TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Paula Etti 221888IVGM

EXPLORING THE USE OF SYNTHETIC DATA IN THE PUBLIC SECTOR: A FRAMEWORK AND CASE STUDY BASED ON THE EXAMPLE OF THE ESTONIAN POLICE AND BORDER GUARD BOARD

Master's Thesis

Supervisor: Mahtab Shahin
MSc

Co-supervisor: Dr Liina Kamm
PhD

Tallinn 2024

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Paula Etti

19.05.2024

Abstract

The expanding utilisation of data-driven services by governments has heightened the need for reliable data access. In law enforcement agencies (LEAs), the intricate nature of the datasets that encompass personal and strategic information, imposes significant challenges on data sharing due to legal constraints, privacy, and security. These factors heavily restrict data access to external entities such as researchers or developers. To address these complexities, synthetic data—artificially generated data that closely mimics real data—has emerged as a promising solution. This thesis examines the potential application of synthetic data in LEA settings to navigate the legal and technical hurdles involved in processing sensitive data. By focusing the capability of synthetic data to preserve known data structures while ensuring privacy, this thesis explores its potential primarily for sharing data outside the organisation, but also takes a look at the internal analytical needs of LEAs. This thesis also reviews different methodologies for creating and analysing synthetic data, and discusses the implications and future directions for their use within LEAs.

Keywords: synthetic data, privacy enhancing technology, law enforcement agency, public sector, data privacy, legal constraints, data sharing, data analysis, synthesis, framework

The thesis is written in English and is 83 pages long, including 11 chapters, 5 figures and 6 tables.

Annotatsioon

Sünteetiliste andmete kasutamise võimalused avalikus sektoris: raamistik ja juhtumiuuring Eesti Politsei- ja Piirivalveameti näitel

Koos andmepõhiste teenuste laialdasema kasutamisega valitsusasutustes on suurenenud vajadus usaldusväärsete andmete järele. Isikuandmeid või strateegilist teavet sisaldavate andmestike keerukas olemus on privaatsuse, turvalisuse ja juriidiliste piirangute tõttu väljakutseks andmete töötlemisel, takistades andmete jagamist teadlaste ja väliste arendajatega. Teadlased peavad paljulubavaks lahenduseks sünteetilisi andmeid – kunstlikult loodud andmestikke, mis jäljendavad pärisandmeid. See magistritöö uurib sünteetiliste andmete võimalikku rakendamist õiguskaitse valdkonnas, pöörates tähelepanu tundlike andmete töötlemisega seotud juriidilistele ja tehnilistele väljakutsetele. Kuna sünteetilistel andmetel on võime säilitada teadaolevaid statistilisi seoseid tunnuste vahel, tagades samal ajal privaatsuse, uurib magistritöö sünteetiliste andmete potentsiaali eelkõige andmete organisatsioonist väljastamisel ja heidab pilgu ka õiguskaitseasutuse sisemistele analüüsivajadustele. Selles magistritöös vaadeldakse ka erinevaid metoodikaid sünteetiliste andmete loomiseks ja analüüsimiseks ning käsitletakse nende kasutamise mõjusid ja tulevikusuundi õiguskaitseasutustes.

Märksõnad: sünteetilised andmed, privaatsuskaitse tehnoloogia, õiguskaitseasutus, avalik sektor, andmete privaatsus, õiguslikud piirangud, andmete jagamine, andmete analüüs, süntees, raamistik

Acknowledgments

This thesis has been prepared as part of the Lessen Data Access and Governance Obstacles (LAGO) project. The project is funded by the European Union grant agreement number 101073951. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

The author of this thesis works at Cybernetica AS as a member of the team responsible for the LAGO project.

I extend my deepest gratitude to the experts participating in the interviews and generously sharing their expertise in the realm of law enforcement and synthetic data. Their insights and guidance have been invaluable in shaping the trajectory of this research. I am indebted to the Police and Border Guard Board for their support and cooperation throughout the duration of this project. I also thank the Estonian Ministry of Interior and the IT and Development Centre at the Estonian Ministry of Interior for enriching the quality of this thesis. I also thank the experts from the Data Protection Inspectorate and Ministry of Economic Affairs and Communications who provided valuable input through the survey.

I would also like to express my appreciation to my colleagues at Cybernetica AS, especially Karl Hannes Veskus, Maria Toomsalu and Hiroki Kaminaga, for their insightful recommendations.

I am deeply grateful to both of my supervisors, Mahtab Shahin and Liina Kamm, for their invaluable guidance and support throughout my thesis writing process.

List of Abbreviations and Terms

AI	Artificial Intelligence
AI Act	Artificial Intellicence Act
AIML	Artificial Intelligence and Machine Learning
BPMN	Business Process Model Notation
DD	Dataset Distillation
DM	Diffusion Model
DP	Differential Privacy
DPO	Data Protection Officer
EIG	Estonian Information Gateway
EU	European Union
GAN	Generative Adversarial Networks
GMM	Gaussian Mixture Model
GDPR	General Data Protection Regulation
HE	Homomorphic Encryption
LEA	Law Enforcement Authority
LLM	Large Language Model
ML	Machine Learning
MPC	Multi-party Computing
NaN	Not-a-Number
NN	Neural Networks
PDPA	Personal Data Protection Act
PET	Privacy-enhancing Technology
PIA	Public Information Act
PPA	Estonian Police and Border Guard Board
SaaS	Software-as-a-Service
SMIT	IT and Development Centre at the Estonian Ministry of the Interior
TEE	Trusted Excecution Environments

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Background and Problem Statement

The European Union (EU) is moving towards ensuring better data availability [1], which would give a boost to the economy throughout the union. The post-GDPR efforts of the EU culminated in the ambitious European Strategy for Data, with the Data Governance Act and Data Act as pivotal legislations aiming to integrate a cross-sectoral architecture for data access [2]. Governments are offering more data-driven services, leading to more data storage [3]. Estonia is currently making a concerted effort to advance public sector innovation by extensively adopting proactive public sector services that utilise artificial intelligence (AI) and machine learning (ML) techniques [4]. Quality data fuels digital economies and societies [5], helping to make better strategic decisions and improve services and products. Therefore, the demand for data access, particularly data gathered using public resources, continues to rise [6].

Law enforcement authorities (LEAs) have vast, complex and sensitive datasets, which contain both personal data and strategic information of the organisation. The sensitivity of the information limits data sharing outside the organisation, such as with researchers or external developers. Data processing can be legally and technically complicated throughout the entire process, especially when dealing with sensitive data, such as personal data or tactical information of a LEA. Therefore, it is necessary to look for suitable solutions that facilitate data processing and provide a sufficient level of privacy and security.

The Estonian digital agenda 2030 [7] underscores the importance of prioritising fundamental human rights, such as privacy protection, and the roadmap for implementing privacy-enhancing technologies (PETs) in Estonia [8] highlights the potential of synthetic data to significantly assist various Estonian public sector institutions by enabling service development and testing. Moreover, current research endeavors explore synthetic data generation as a viable, efficient, and privacy-preserving substitute for real-world data collection [9]. Synthetic data can bridge gaps when real data is scarce or legally complex to process [10, 11, 12], providing an alternative to personal data for complex statistical analyses, as well as for AI and ML (AIML) research [13].

The Estonian internal security development plan 2020-2030 [14] envisages an innovative approach to ensure internal security, utilising smart and innovative solutions. E-government

services necessitate rigorous testing to ensure quality, encompassing performance, correctness, and usability, but data protection regulations limit the use of personal data in such testing scenarios, rendering purely random data inadequate for uncovering all issues [15]. Synthetic data allows for extensive testing and training without risking exposure of real data. Synthetic data not only preserves the structure and characteristics of real data but also speeds up e-government service enhancement and supports data-rich government sectors by allowing more comprehensive analyses [8, 15]. Moreover, synthetic data serves as a crucial privacy-promoting measure, enabling the generation of reliable data that retains essential features of the original, supports controlled information release, helps mitigate biases, and proves useful in diverse applications such as AI training and enhancing cybersecurity [5, 16].

The rapidly increasing volume of data offers substantial opportunities for enhanced decision-making and innovation, yet it also raises significant privacy concerns, particularly in sensitive sectors such as law enforcement. The handling and disclosure of data come with a host of risks, ranging from data quality to privacy and cybersecurity—challenges that are exacerbated when third parties process the data held by LEAs. Nonetheless, the strategic use of synthetic data as a PET in law enforcement is highly promising, as it helps alleviate many of the current challenges in data management.

1.2 Purpose and Importance

The objective of this thesis is to analyse data processing challenges related to LEAs, given the sensitive and regulated nature of the field. Synthetic data has garnered significant attention as a substitute for real data, providing a sophisticated approach to addressing data scarcity and privacy issues, and bolstering data analytics capabilities while maintaining known data structures to ensure privacy during analysis [17, 18, 19]. Considering the constraints in LEAs and the potential of synthetic data, the objective of this thesis is to investigate whether introducing synthetic data in the LEA settings could alleviate complexities arising from legal concerns, privacy, and security.

This thesis examines the current practices of using synthetic data in LEAs, provides an overview of the potential of synthetic data, highlights the solutions that would alleviate current legal, technical, or other obstacles, outlines limitations of synthetic data, and proposes avenues for future research. It contributes to the activities of the police and LEAs in general, finding ways to protect sensitive data (for example, personal data and tactical information) more efficiently. Based on the results of the research work, it is possible to give recommendations or make suggestions for improving the efficiency of certain processes. This thesis also lays the groundwork for further research into the deployment

of PETs like synthetic data in LEAs. This thesis denotes the creation of synthetic data as synthesis.

The main objectives of this master's thesis are:

1. To study the practices of using synthetic data in LEAs;
2. To study different methods of generating synthetic data, highlighting the advantages of the method and opportunities for improvement;
3. Analyse the possibilities of using synthetic data based on the example of the Estonian Police and Border Guard Board (PPA) in two categories: for releasing data to researchers and external developers;
4. Analyse the legal landscape regarding syntehtic data;
5. Provide a practical overview of the process of creating synthetic data.

The topic is relevant because it is not easy to acquire data from LEAs for research work. In certain cases, in addition to submitting a detailed research request to process such data, it may also be necessary to obtain permission from the Data Protection Inspectorate or an ethics committee to process personal data or to assess whether conducting the study is ethical. Sharing data with external development partners is technically and legally complex. In addition, various methods of generating synthetic data, including their advantages and disadvantages, have not yet been studied very thoroughly. Despite its promise, synthetic data generation lacks systematic frameworks for safe deployment, presenting challenges that require tailored solutions [16].

Synthetic data generation is vital in domains with limited data availability and concerns about privacy and confidentiality [10]. Therefore, synthetic data holds great promise in today's digital society as a solution for processing sensitive data, but its properties remain underexplored and the abundance of models combined with a lack of comprehensive literature presents challenges [20]. This thesis will fill a certain gap in the field of scientific research regarding facilitating the work of LEAs.

From the perspective of this master's thesis, the important stakeholders who could benefit from this research are the following:

(a) Representatives of the PPA, primarily the management, who can take the decision to introduce synthetic data in the organisation;
(b) Researchers involved in the study of PETs;
(c) Policymakers who can provide impetus for greater research and adoption of synthetic data at the national level;

(d) Other organisations that have been interested in using synthetic data but have not yet done so, for example due to a lack of sufficient knowledge or skills.

1.3 Previous Work

There is limited prior research on the use of synthetic data in LEAs, for example [21, 22, 23, 24, 25]. Synthetic data has previously been used by Buil-Gil et al. [21] in the geographic crime analysis that examined the accuracy of crime statistics, where synthetic data replicating Manchester's characteristics from the United Kingdom (UK) Census parameters was employed. According to the study, while synthetic data offers a promising avenue to mitigate biases in fine-grained crime analysis, its use requires refinement and validation. The study recommended leveraging census data for workday populations, merging diverse datasets, and innovating techniques to enhance accuracy, ensuring synthetic data becomes a reliable tool for understanding crime patterns. There has been also a study [22] conducted by Pina-Sánchez et al. exploring synthetic crime data to better understand crime patterns and alleviate the gap between police records and 'true' level of crime which found that by forecasting the likehood of victimisation for every artificial resident, it is possible to compute both occurrences of crimes and their recording rates across various spatial scales.

Dokoupil focused on his study [23] on eyewitness identification, a crucial aspect of criminal proceedings, addressing the challenges associated with assembling police photo lineups, which often involve manual or semi-automatic processes. He developed a variation of the StyleGAN2 model, trained on a dataset of missing persons, to create synthetic facial images for use in police lineups. Key achievements included producing images of reasonable quality and ensuring these images were indistinguishable from real human photos, as evidenced by a user study where participants were not significantly better at identifying synthetic images. The model achieved sufficient diversity, although it was limited to the diversity present in the training dataset, and allowed for controlled outputs via an implemented encoder. However, challenges remained in generating rare facial features absent from the training data, with only partial success through fine-tuning attempts, and suggestions for future improvements were noted for further enhancement of the model's capabilities.

Research by Brunton et al. [24] has highlighted gaps in our understanding of measurement error in police-recorded crime data across different spatial scales and area characteristics. The study underscores the limitations of traditional assumptions about uniform police under-counting, and introduces a novel method using a synthetic population for England and Wales, allowing for more accurate predictions of victimisation at various spatial levels and better understanding of how measurement errors correlate with area characteristics.

This approach represents a significant advancement in the computational analysis of crime data, encouraging further innovation in this field.

The use of synthetic data has been effectively employed in creating datasets on trafficking victims to enhance data-driven collaboration within the counter-trafficking community, as highlighted by Edge et al. [25]. The study explored the feasibility of synthetic data for anomaly detection in complex social networks, linking its potential applications to law enforcement and predictive policing. Predictive policing, as defined by Perry et al. [26], involves using analytical techniques to identify likely targets for police intervention, aiming to prevent crime or solve past crimes through statistical predictions.

This thesis provides a comprehensive examination of synthetic data, focusing on its creation, use, and governance within an organistional context. The structure of the thesis is organised into several key sections to cover the diverse aspects of synthetic data. Section 2 details the research methods employed to gather and analyse data, setting the foundation for the investigation. Section 3 offers a detailed definition of synthetic data, outlines its primary characteristics, and discusses the inherent risks associated with its use. Section 4 summarises various techniques and technologies involved in the generation of synthetic data, providing a technical backdrop to the topic. Section 5 explores the legal context surrounding synthetic data, highlighting regulatory challenges and considerations necessary for compliance. Section 6 presents feedback from experts in synthetic data, data protection, and data governance, enriching the thesis with professional perspectives and experiences. Section 7 analyses the specific environment of the PPA, detailing their experiences and use cases with synthetic data. Section 8 discusses the analytical process undertaken within the study, including significant findings and lessons learned from the practical application of synthetic data. Section 9 provides a proposed framework for the implementation of synthetic data within an organisation, aiming to guide practical application. Section 10 outlines strategic recommendations for future actions that can enhance the understanding and use of synthetic data. Section 11 draws final conclusions, encapsulating the research findings and underscoring the potential impacts and future directions for the field of synthetic data.

2. Methodology

2.1 Hypothesis and Research Questions

The hypothesis of this thesis is that the use of synthetic data in a LEA ensures better protection of sensitive data and improves the efficiency of the authority's processes.

To support the hypothesis, three research questions were formulated.

1. How do LEAs currently deploy synthetic data?
2. Which processes within LEAs currently hold the greatest potential for the utilisation of synthetic data?
3. What are the key barriers to adopting synthetic data on a larger scale?

2.2 Research Methods and Design

The primary focus of this thesis was on enhancing the efficiency of law enforcement processes while simultaneously safeguarding data integrity and privacy. Given the importance of fairness and bias within LEAs, the thesis also delved into metrics related to privacy, utility, and fairness. To gain a legal theoretical perspective, this thesis investigated data protection laws and regulations and highlighted the three most pressing issues regarding synthetic data. By integrating these theoretical approaches, the thesis aimed to offer a comprehensive understanding of the role of synthetic data in the realm of law enforcement.

This research was experimental, conducted without predefined models or guidelines because the field of data synthesis is still evolving. In order to carry out this research, the following tasks were performed:

1. secondary data analysis (qualitative document analysis) to define the theoretical and legal framework and to identify possible uses of synthetic data in LEAs;
2. semi-structured interviews (empirical data collection);
3. expert surveys (empirical data collection);
4. statistical analysis and data synthesis with open data (to demonstrate the process of generating synthetic data).

This thesis focused on the use of synthetic data within LEAs, examining key practices,

potential new use cases for synthetic data, and factors that may hinder the implementation of synthetic data. It also explored the obstacles to data dissemination outside the organisation and reviewed scientific and legal literature. Interviews helped to elucidate organisational practices and provided insights into the use of synthetic data within the PPA. The conducted expert surveys provided input on the overall potential of synthetic data but also revealed possible obstacles to its wider adoption.

This thesis also conducted data analysis, including data synthesis, using the UK Police open data as an example. The analysis involved examining both the original and synthetic datasets to understand the structure and characteristics of the synthesised data. Additionally, a framework was proposed to assist public sector organisations in the initial implementation of synthetic data. Recommendations were also provided on how to enhance the broader adoption of synthetic data.

2.3 Interviews

To gain an overview of the data processing operations associated with the organisation, four interviews were conducted with experts of PPA who had diverse backgrounds and experiences. Additionally, an interview was carried out with experts from the Ministry of the Interior and the IT and Development Centre at the Estonian Ministry of the Interior (SMIT) to obtain a more strategic perspective regarding the implementation of syntehtic data. One of the objectives of the interviews was to discover the organisation's current experiences with synthetic data and to investigate where and in which processes the introduction of synthetic data could offer the most potential. Additionally, this thsis aimed to identify the main obstacles to the adoption of synthetic data and research how to overcome the barriers.

To obtain qualitative input, semi-structured interviews with the experts in the field were conducted. Responses from the interviews were the basis for creating Chapter 7 of this thesis. The names and positions of the interviewees were not disclosed in this thesis due to ethical and security concerns; only their expertise in the field was highlighted (see Table 1).

Table 1. *List of interviewees*

Nr	**Field of Specialisation**	**Organisation**
1	Information Systems Development	Police and Border Guard Board
2	Synthetic Data	Police and Border Guard Board
3	Research Work	Police and Border Guard Board

Continues...

Table 1 – *Continues...*

Nr	Field of Specialisation	Organisation
4	Data Protection	Police and Border Guard Board
5	Data Management	Ministry of the Interior
6	Development Expertise	The IT and Development Centre of the Ministry of the Interior

When formulating interview questions, consideration was given to the interviewee's role within the organisation and their previous experience with synthetic data. The interview questions (see Appendix 2) were grouped according to different topics, allowing the specialist in the field to specifically address questions related to their area of expertise or experience. The privacy notice required for conducting interviews is provided in Appendix 3 of the thesis.

2.4 Surveys

In addition to interviews, two surveys were conducted, with the first aimed at synthetic data experts from the private sector and the second at data management and data protection experts from the public sector. While the first survey aimed to gather immediate opinions on the positive and negative attributes and potential of synthetic data, the second survey sought to provide additional insights specifically from a data management and data protection perspective. The survey overview presents detailed insights from five experts on synthetic data, data management and data protection focusing on the potential, challenges, and future prospects of synthetic data. In this master's thesis, the contributors are anonymous, except for Ott Velsberg and Urmo Parm, who have graciously permitted the use of their names as experts. List of respondents to the survey can be seen in Table 2.

Table 2. *List of respondents to the survey*

Nr	Field of Specialisation	Organisation
1	Urmo Parm, Head of Technology	Data Protection Inspectorate
2	Ott Velsberg, Chief Data Officer	Ministry of Economic Affairs and Communications
3	Developer of PETs	Cybernetica AS
4	Synthetic data privacy expert	Cybernetica AS
5	Privacy-preserving machine learning expert	Cybernetica AS

3. Synthetic Data

3.1 Definition and Categories

In 1993, Donald B. Rubin introduced synthetic data [27]. Synthetic data, also known as fake data or artificial data, a replica of original data generated artificially, transforms processes by enabling organisations to share complete datasets, without the sensitive information [28, 16, 29, 30]. It mirrors original data statistically, offers solutions for efficient processing, and allows to pursue goals while reducing the risk of data breaches and re-identification [28, 29].

An early comparison between synthetic and real data [31] demonstrated that statistical inferences from synthetic data closely match those from original data, addressing data availability challenges [32]. Synthetic data can be described as data generated through a process that learns the characteristics of authentic data [33]. In the rapidly advancing realm of AI, the creation and utilisation of synthetic datasets have gained growing importance [34]. When generated appropriately, synthetic data can accurately replicate the statistical patterns of the real data they are derived from while ensuring the anonymity of real individuals [35, 36].

El Emam et al. divided synthetic data into three categories [29]. The initial category is derived from authentic data, such as authentic datasets that can contain also personal data. The second category is independent of real data—the data is simply invented by generating. The third category represents a blend of these two approaches, meaning that real data and synthetic data are combined.

Synthetic data presents an attractive solution for enabling widespread data access for analysis, allaying privacy and confidentiality concerns [6]. Used judiciously, synthetic data fosters cross-dataset learning while safeguarding privacy and mitigating data incompleteness or bias [16]. Furthermore, synthetic data not only facilitate secure data sharing with minimal privacy apprehensions but also support data augmentation, artificially boosting data volume and enhancing the performance of ML models [36].

3.2 Characteristics

According to the "PET technology ratings" [37], the accompanying table for the roadmap for implementing PETs in Estonia [8] and the concept of PETs [38], the potential of synthetic data was evaluated by experts in four categories as presented in Table 3.

Table 3. *Expert ratings for synthetic data according to [37]*

Aspect	**Rating**
Development complexity—additional effort required based on the technology to get the system up and running	Implementation complexity is low. Setup time is average. There are many configurations and options. Requires moderate skills
Maintenance complexity—additional effort required based on the technology to keep the system running	Implementation complexity is low. Performance and energy consumption are low. Data synthesis can be done through both an on-premises solution and a purchased service
Accuracy—whether the computational result is correct	Inaccurate. Loses attributes or entries and adds noise. The simplest attack against a correctly built system is data unlinkability—by chance, it might be possible to synthesise data about a real existing person. It does not offer auditability, does not protect the integrity of processing, and agreed-upon objectives. The guarantee against re-identification is statistical
Technology maturity—how easy it is to procure	Maturity is average
Technology availability	Integration project is required. Availability of products or frameworks is average, with few commercial deployments

Table 3 illustrates the maturity of synthetic data. It is clear that synthetic data requires research and pilot projects to fully understand its potential. The literature has delved into various aspects of synthetic data, including diversity, privacy preservation, and utility. These topics are further explored in the following sections.

3.2.1 Privacy

Synthetic data serves as a replacement for real data that cannot be publicly shared due to privacy constraints [17]. Privacy or disclosure control of the synthetic dataset means what extent can the synthetic dataset provide privacy protection without revealing the identity or confidential information of individuals within the dataset [39, 40]. Robust privacy assurances facilitate effective safeguarding of personal data and sensitive information within an organisation, while model accuracy is pivotal for both result analysis and avoidance of bias.

Synthetic data can maintain the privacy of individuals or sensitive information since it is generated rather than collected from real-world sources. Model accuracy and privacy preservation are primary focuses for researchers as they assess whether synthetic data is both informative and suitable for sharing [17]. However, there is a risk of revealing original data properties, and inadvertent disclosure of private or proprietary information can occur [41]. Concurrently, concerns regarding the possible exposure of identities and sensitive information compel data collectors to restrict data access [6].

Therefore, the scientific analysis of synthetic data is crucial to identify an appropriate model for generating adequate data. Not every use case requires the same level of privacy in the dataset. An inevitable aspect of synthetic data generation involves an intrinsic balance between privacy and utility [18, 42]. In some instances, a lower level of privacy is acceptable. Data of lower quality can be constructed to mirror the framework of the source data but does not retain any of the connections and might include unrealistic covariate values [18]. While this data does not carry the risk of disclosing information, its utility is limited to grasping the data format [18]. However, when dealing with LEA's data, privacy assurances become one of the central challenges. High-quality synthetic data inherently carries a greater risk of information disclosure [18]. Hence, it is crucial to find an appropriate method for generating synthetic data and use suitable PETs.

Privacy considerations are paramount in data processing, therefore it is recommended to utilise additional protective measures in data synthesis. For instance, it has been found that employing differential privacy (DP), that offers formal mathematical privacy protection, in data synthesis enhances security and contributes to ensuring privacy [43]. A study proposes a general method for generating differentially private synthetic data, involving three steps: (1) selecting low-dimensional marginals, (2) measuring them with noise addition, and (3) generating synthetic data preserving the measured marginals [44].

However, DP also poses challenges, particularly in defending against membership and

reconstruction inference attacks, and it may not fully protect specific values like categorical IDs or names [43]. Where needed, it is advised to employ privacy-enhancing methods like multi-party computing (MPC), homomorphic encryption (HE), and trusted execution environment (TEE) alongside DP to bolster security assurances [10]. In addition to MPC, HE and TEE, it is recommended to further explore the potential of technology called Byzantine adversaries, that focuses on addressing potential issues related to data corruption, communication failure, or malicious attacks that deviate from the learning model [10].

3.2.2 Bias Mitigation and Fairness

In the case of LEAs, addressing fairness and bias is crucial, with bias mitigation being a key aspect of data processing. Bias refers to a consistent error in decision-making processes leading to unfair outcomes [45]. Through careful design and generation techniques, synthetic data can mitigate biases present in real data, promoting more fair and accurate analyses.

Bias can manifest in various forms, such as within the datasets themselves, the methods and tools used to generate synthetic data, or through the interpretations made by individuals. This multifaceted presence of bias poses significant risks, including the potential alteration of individuals' self-perception and their perceptions of others, which could change their opportunities and interactions, thereby necessitating rigorous measures to ensure fairness [45]. To address these concerns, it is crucial to use diverse datasets as bias can emerge from data collection, algorithmic design, and human interpretation, leading to potentially unjust outcomes [45]. Moreover, fairness assessments in synthetic data focus on whether equity is preserved across different groups or demographics from the original data, ensuring that biases are either carried over or mitigated. Despite advances in synthetic data generation, which allow for the creation of realistic and sensitive simulations, biases may persist, emphasising the need for continual evaluation and improvement of these technologies [13].

In real-world scenarios, data is often observed without clear knowledge of its generation mechanism [46]. Also, factors such as under-representation can result in structurally missing data or inaccurate correlations and distributions, which will be reflected in the synthetic data generated from biased ground truth datasets [13]. Hence, dataset diversity is closely associated with the issue of bias. Also, despite efforts to classify bias in data and algorithms, the connection to fairness in ML-based decision-making systems remains limited, and bias mitigation efforts often overlook specific biases in the data [46].

To avert adverse outcomes, it is crucial to understand how and where bias enters the entire modeling pipeline and explore potential mitigation strategies [46]. Correcting biases in ground truth datasets is crucial to prevent errors in correlations and distributions from being mirrored in biased datasets [13]. Incorporating methods for rare event detection and correction can be crucial for a synthetic dataset service [13]. Research also suggests that assumptions regarding the data generation process play a pivotal role in shaping the interpretation of bias within the relevant use case [46].

3.2.3 Interpretability

The lack of a widely agreed-upon definition of interpretability frequently yields divergent viewpoints, however, the concept is associated with the following terms: clarity, understandability, simplicity, and readability [47]. Understandability is also influenced by psychological factors, relying not just on the phenomenon and its explanation but also on the individual receiving the explanation [48]. In general, interpretability measures how easily a user can comprehend a model's reasoning [49], and assesses the ease with which insights and conclusions can be drawn from the synthetic data.

Interpreting data requires knowledge and experience, and it is also important to know which method of synthetic data generation was used and the nature of the data synthesis tool. Studies indicate that technologies play a crucial role in improving data interpretability and facilitating the creation of an open government [47]. Indicators are also relevant. A study [50] shows that the precision, recall, and authenticity metrics prove valuable, offering insights into both distribution accuracy and individual sample quality.

Interpretability can also be understood through the metrics of fidelity and utility. This approach reveals that an interpretable model is not just about transparency, but also about how faithfully it represents the underlying data (fidelity) and how useful it is in practical applications (utility).

3.2.4 Fidelity

Fidelity in synthetic data measures how closely it mimics the original data's statistical properties and patterns, encapsulating its degree of similarity and realism compared to the original dataset. Levels of fidelity (low, mid, and high) and types (physical, psychological, and conceptual) are associated with fidelity in simulations, and evidence suggests that all levels can be beneficial when used appropriately [51]. Although high fidelity enhances utility, it may compromise privacy by accurately representing the original data

too closely [41].

Fidelity can be measured using metrics like mean absolute error, root mean square error, and correlation coefficient. Synthetic data can be evaluated also through a statistical resemblance score, which compares its features to those of the original dataset, and a marginal distribution likeness score, assessing it based on each feature's marginal probability distribution [17]. However, in certain scenarios, such as testing responses to unexpected events, realism may not be essential [41]. Ultimately, realistic, high-quality synthetic data supports the advancement of data analysis methodologies by allowing researchers to test and refine techniques in contexts similar to those originally intended, fostering deeper insights and promoting open scientific dialogue [5, 18]. Additionally, making data accessible alongside published studies promotes open scientific dialogue [18].

Ebert-Uphoff and Deng conducted a research [52] wherein they utilised synthetic data to improve result comprehension by deciphering the configuration of a graphical model using observed spatio-temporal data, thereby pinpointing the interrelations within the observed physical system. Thus, synthetic data can be also valuable for interpreting real-life phenomena if it is interpreted correctly.

3.2.5 Utility

Utility means the usefulness of the synthetic data for specific analytical or operational tasks, and the dataset's capacity to preserve the statistical structure identical to that of the original sample from which it originated [39, 53]. The utility of a synthetic dataset lies in how accurately it mirrors real data [29], and how effectively the synthetic data can be used for specific analytical tasks or applications. Some industries are increasingly using synthetic data to enhance data utility and privacy, recognising its value as an innovative solution [54].

Establishing theories and metrics for evaluating the utility of synthetic data in maintaining the desired characteristics of the original data is still an ongoing area of research. Utility can be measured with model performance metrics, decision impact analysis and statistical tests. Several studies have examined the utility and reliability of ML algorithms trained on synthetic data, yielding promising results [32]. Synthetic data has demonstrated effectiveness in improving ML models and testing processes [41]. The utility of the disseminated data greatly hinges on models' capacity to encapsulate crucial relationships present in the initial data [55]. Types of synthetic data and the corresponding utility metrics according to El Emam et al. [29] can be seen in Table 4.

Table 4. *Type of synthetic data vs utility metrics*

Type of synthetic data	**Utility**
Derived from authentic nonpublic datasets	May exhibit considerable magnitude
Derived from authentic public datasets	May be substantial, albeit constrained by the de-identification or aggregation common in public data
Derived from an existing model of a process, which can also be simulated in an engine	Will be contingent upon the faithfulness of the pre-existing generative model
Informed by analyst expertise	Hinges on the analyst's familiarity with the domain and the intricacy of the phenomenon
Derived from generic assumptions not tailored to the phenomenon	Expected to be minimal

Striking a balance between privacy and utility is crucial, especially in law enforcement where accurate representation of data is essential. Whether a high utility score is needed or not depends on the specific use case and the goals of data processing. In certain scenarios, possessing high utility holds significant importance [29]. Conversely, in other instances, moderate or even low utility may suffice [29], especially in some simpler data processing cases like performance testing. However, balancing privacy and utility is akin to navigating through a delicate dance of complexity. The level of the mentioned metrics must be assessed on a case-by-case basis, depending on the purposes of data processing.

3.3 Advantages

Diversity. Creating artificial datasets is a cutting-edge method for distributing data [55]. Synthetic data can offer a broader range of scenarios and instances compared to real data, enabling more comprehensive testing and analysis. Sampling methods that incrementally increase diversity along one dimension at a time overlook the opportunity to create intersectional datasets [56]. Additionally, introducing a degree of randomness in the synthetic data generation process can enhance data variability, although careful control is necessary to ensure realism [30].

Scalability. Synthetic data has seen widespread application in various real-world scenarios, notably in addressing privacy concerns in datasets [17]. For instance, synthetic data

can replace sensitive real data, allowing for public testing of ML models while preserving privacy [17]. This is primarily because synthetic data can be generated in large quantities, facilitating the scalability of experiments and analyses without additional data collection efforts.

As the demand for training data grows, programmers, data analysts and data scientists are increasingly turning to synthetic data, which can be scaled to meet the training requirements of ML models. Findings indicate that the primary factor impacting the robustness of downstream models is the quantity of data, with other factors playing a lesser role [57]. Synthetic data is also used in oversampling unbalanced datasets, where synthetic data is generated to balance the ratio of positive and negative samples during model training [17].

Issues related to scalability have been studied in several research works, such as [17, 58, 59], and it has been found that synthetic data can help alleviate scalability concerns arising from insufficient, incomplete or sensitive raw data. Bösche et al. [58] recommended to use a hybrid approach to address the tension between authentic and scalable datasets.

3.4 Riskis Related to Synthetic Data

Model Performance Limitations. Synthetic data is critical in AI applications, yet it poses substantial challenges and risks due to potential deviations from real-world data, as highlighted by Hao et al. [34]. These deviations include disparities in features and class distributions, leading to biased predictions and reduced model fidelity. Additionally, errors, biases, or incomplete information in synthetic datasets can impair accuracy and generalisation across diverse conditions, with a lack of real world inconsistencies and detailed nuances further limiting model effectiveness in realistic environments [34].

Further complicating the landscape, Lu et al. [60] discuss how biases in datasets significantly affect dataset distillation (DD), underscoring the necessity for tailored bias mitigation strategies. They propose a mathematical definition of biased DD, with detailed explorations into its implications deferred to future studies. This ongoing research aims to extend experiments to larger datasets, more complex models, and advanced DD techniques, potentially improving understanding and handling of biases in synthetic data [60].

Ethical and Social Concerns. Synthetic data is pivotal in AI applications but introduces ethical, social, and technical challenges due to potential biases and discrepancies from real world data [34]. Concerns also extend to the potential misuse of synthetic data to create fictional scenarios that could misinform or lead to adverse societal impacts [34].

The ethical landscape is further complicated by the use of synthetic data in sensitive areas such as law enforcement and healthcare, where biases can significantly impact fairness and equity [61, 62].

Fairness issues are also prevalent in dataset biases and the fairness feedback loops in ML model development, affecting how these models perform and the societal policies they influence [39, 63]. Various methods and strategies have been proposed to mitigate these biases in synthetic datasets, including algorithmic reparation interventions which aim to rectify biases embedded in data ecosystems [63]. In more technical aspects, biases in data often stem from covariate imbalances during the data collection or analysis phases, impacting the representation of certain groups or features in datasets [64]. Approaches to address these include ensemble-based learning, cost-sensitive learning, and single-class learning which target the learning stage of analyses to manage covariate imbalance [64].

The creation of synthetic face data for AI applications raises significant concerns about bias in facial recognition systems, which could lead to unfair decisions based on demographic or non-demographic attributes [65]. Researchers have suggested employing frameworks like quality-diversity generative sampling to uniformly sample data across diverse groups despite biases inherent in the data source [66]. Moreover, integrating domain-specific expertise can enhance the authenticity and realism of synthetic data, making it more representative of real world contexts [34]. DD is also gaining attention as a method to create condensed, yet representative synthetic datasets, although existing approaches often neglect the impact of inherent biases [60].

Inherent biases in AIML model training remain a persistent challenge, stemming from various factors such as demographics and physical attributes, and techniques like over-sampling or applying DP may inadvertently exacerbate these issues [32]. Furthermore, as cybersecurity increasingly integrates AI, human traits and biases in AI could amplify challenges in this field [67].

Security concerns. Cyberattacks have escalated in complexity and frequency, increasingly being used by states and groups to destabilise societies and integrate into their foreign policy strategies, as highlighted by the recent surge in aggression from Russia towards Ukraine. This has prompted Estonia to prioritise comprehensive national defence, secure digital solutions, and societal resilience during crises, especially given the country's reliance on information systems for critical infrastructure [68].

ENISA's 2023 Threat Landscape report [69] categorises data-related threats into breaches,

leaks, and manipulation. Data breaches involve deliberate cyberattacks to access and disclose sensitive data. Data leaks usually result from misconfigurations or human errors leading to unintentional data exposure. Data manipulation, a rising concern with the advent of AIML, aims to corrupt reliable data to sabotage AIML accuracy and skew reality perception, including tactics like data poisoning and information manipulation.

The introduction of synthetic data presents new adversarial risks and security concerns. Synthetic data, while beneficial for training AI models where real data availability is constrained, may also be exploited for malicious purposes, such as destabilising AI models during adversarial attacks. This vulnerability underscores the need for robust mechanisms to detect and mitigate threats posed by synthetic data manipulation [34, 70].

In cybersecurity, deep learning techniques are employed to identify malicious patterns, yet the challenge of acquiring large, privacy-compliant training datasets remains. Generative Adversarial Networks (GANs) have been highlighted as a solution to generate synthetic attack data, enhancing the training of cybersecurity models [39]. Moreover, the potential manipulation of voice assistants using synthetic voices necessitates the integration of advanced spoofing and deepfake detection technologies [67]. While synthetic data offers considerable advantages for enhancing cybersecurity model robustness, its use also introduces significant risks that must be carefully managed through a balanced approach that incorporates both natural and synthetic data to ensure the reliability and security of cybersecurity systems.

In terms of data generation, synthetic datasets can be produced internally or sourced from specialised providers. While in-house processing keeps data within the organisation, mitigating certain risks, challenges in expertise and suitable tool availability persist. When engaging third-party services, security of data synthesis and transmission must be critically evaluated. Techniques such as TEEs have been found effective in securely outsourcing data synthesis to untrusted servers, thus protecting both original and synthetic data [71].

Legal Compliance Challenges. Using synthetic data in LEAs can encounter significant regulatory hurdles, as highlighted by the complexities associated with data protection laws and ethical considerations. The legality of employing synthetic data often pivots on several crucial factors, including the source of the original data, its contents (whether it includes sensitive or personal data), the purpose and method of its processing, necessary consents, and the impact of the synthesis process.

Additionally, regulatory bodies require that data processing models be transparent and

interpretable, which can be challenging for the synthetic data generation process [34]. Furthermore, when generating synthetic data, compliance with EU AI regulations may be necessary as the synthesiser or data processing software may classify as AI system, thereby imposing additional requirements on these technologies. Legal nuances are discussed further in Chapter 5.

4. Data Synthesis

4.1 Overview

Synthetic data generation utilises various methodologies, from statistical and mathematical models to more complex AIML technologies, to create data that mirrors real-world data characteristics while ensuring privacy and diversity [5, 11, 72]. These methods are designed to generate data that cannot be distinguished from the original, serving functions especially where privacy constraints are significant and use of real data is limited [19, 33]. The generation process not only involves fitting models to real datasets to capture their statistical properties but also employs advanced techniques like deep learning to address the limitations of real data, such as imbalances and discrimination [34, 73].

As the need for inclusive data-sharing grows, synthetic data has become crucial in AI and ML realms, notably in model training and validation [19, 29]. This trend is supported by a proliferation of synthetic data generators, particularly those leveraging AIML techniques, although their utility remains largely exploratory at this stage [32, 35, 73]. Notably, DP has been incorporated to enhance disclosure control in synthetic data generation, reinforcing privacy without compromising data utility [32, 74].

The broad applicability of synthetic data is expected to make it a primary resource for AI training by 2030, as predicted by industry forecasts, which anticipate synthetic data comprising a significant portion of AIML training data [32, 75]. This shift emphasises the importance of enhancing generative models using diverse datasets to improve the accuracy and effectiveness of ML models trained on synthetic data [76, 77].

4.2 Statistical Models and Rule-Based Generation

In rule-based generation, data is generated based on predefined rules, constraints, and statistical distributions. These rules can be derived from domain knowledge or observed patterns in real data. Statistical models for generating synthetic datasets often rely on analysing the distribution, relationships, and characteristics of real data, aiming to produce synthetic data with similar properties by simulating these statistical features [34]. Rule-based generation allows for fine-grained control over the characteristics and relationships of generated data, and ensures that generated data conforms to specific constraints or regulations. At the same time, rule-based generation may struggle to capture complex

and non-linear relationships present in real data. In addition, extensive domain expertise is required to define accurate rules. Employing rule-based methods to generate a dataset minimises the risk of re-identifying individuals from the original data while preserving the overall structure and internal distributions [18].

When real data is unavailable, analysts can generate synthetic data by understanding and utilising the expected distribution of the dataset. This involves creating randomised samples from various distributions, mimicking the original data's distribution characteristics through probability density functions for continuous data and probability mass functions for discrete data. Techniques like linear interpolation and kernel density estimation enhance the generation process by creating new data points and estimating complex, multimodal distributions, respectively, thereby effectively capturing the intricate nature of the data environment [34].

4.3 Artificial Intellicence and Machine Learning Models

The use of ML in law enforcement for predicting crime patterns and resource allocation is growing due to its potential in data analysis and model construction [78]. Synthetic datasets are becoming crucial in balancing privacy and data utility in ML, with innovations in data synthesis allowing for the preservation of statistical properties without direct reliance on real data, significantly enhancing privacy [73, 79, 80, 81].

In data generation, traditional model-based methods and advanced techniques like GANs play key roles. Model-based generation mimics real data distributions effectively, especially in linear relationships [34], while GANs, leveraging adversarial learning, produce high-quality synthetic samples and maintain statistical characteristics with minimal privacy risks [32, 36, 43]. Moreover, GANs, particularly suitable for object-centric images, still face challenges like non-convergence and mode collapse, and require careful hyperparameter tuning [43, 82].

Large language models (LLMs) and diffusion models (DMs) are also emerging as powerful tools for data synthesis. LLMs like GPT-3.5 excel in generating synthetic data for narrow domains, while DMs capture dynamic information accurately, enhancing ML training across diverse domains [34, 82]. Despite their potential, these models face challenges in ensuring data privacy and addressing biases inherent in training data, which can impact the effectiveness and fairness of the generated datasets [83, 84]. The integration of sophisticated generative models is crucial for advancing synthetic data capabilities, which helps in addressing privacy concerns, enhancing data utility, and maintaining fairness in ML applications [34].

4.4 Deployment Methods

Synthetic data deployment method refers to the manner and location where the data are synthesised. These methods can vary depending on the specific needs, such as available resources, and constraints, such as data sensitivity and regulatory considerations, of the organisation or certain project. Discrepancies in security provisions between data at-rest environments and compute environments pose architectural challenges for organisations, demanding tailored solutions across diverse security contexts [43].

It is possible to use on premises or cloud based solutions, but it is also possible to combine these for a hybrid solution. There are three main deployment models:

1. On-premises solutions, where a data synthesiser can be installed within the organisation.
2. Cloud-based solutions, where a data synthesiser is deployed on the cloud provided by a third party (vendor).
3. Hybrid solutions that are the combination of the two previous solutions.

While the first method allows the organisation to have complete control over the synthesis, for the second and third methods, the organisation's control decreases depending on the specific solution. Deploying software within enterprise networks is a multifaceted endeavor, influenced by organisational requirements, structural intricacies, and legacy system limitations, while the emergence of cloud transformation introduces a spectrum of deployment alternatives such as on-premises, hybrid, public, and multi-cloud architectures [43].

On-Premises Solutions. Organisations may deploy synthetic data generation and management systems within their own infrastructure. This approach provides full control over data handling and security, granting complete authority over both the data and technological infrastructure, free from reliance on third parties or vendor restrictions [85]. Such an on-premises solution provides the organisation with the greatest control over data processing, but requires dedicated hardware and IT resources.

Cloud-Based Solutions. Cloud computing services are revolutionising business and governmental operations at an accelerating pace [86], offering flexibility and scalability. Web analytics tools enable organisations to gather insights into user interactions on their websites, aiding in web usage optimisation [87].

Despite public sector organisations favoring cloud-based services like software-as-a-service (SaaS) solutions [88], challenges emerge from security risks and limited customisation capabilities, impeding widespread adoption of the SaaS model [86]. Most SaaS products, accessed through a web browser or similar interface and hosted in the cloud, encounter various challenges spanning legal issues [88], data privacy aspects [87], and security concerns [86]. Utilising a cloud solution raises concerns regarding lawful data processing, organisational obligations to safeguard digital assets, and potential vendor lock-in, thus necessitating the presence of a contingency plan at all times [87, 88].

Research [86] has examined these challenges using game-theoretical models to study the interactions between cloud service providers, on-premises software vendors, and consumers with diverse usage patterns and security needs. The findings reveal complex effects on consumer decisions, vendor strategies, and societal welfare, highlighting the intricate relationship between security, customisation, and cloud service adoption [86]. Illustrating with an Austrian SME as an example, a study demonstrates the practical application of secure cloud-based storage services, emphasising the importance of minimising the system's attack surface and fortifying essential software components against network-based threats from a security standpoint [89]. Therefore, when utilising cloud services, it is crucial to assess and mitigate associated risks, necessitating both a comprehensive risk assessment and, particularly for personal data processing, a data protection impact assessment, integral steps in the adoption of cloud services.

Hybrid Solutions. Organisations may opt for a combination of on-premises and cloud-based solutions, depending on factors like data sensitivity, computational requirements, and regulatory compliance. Establishing a hybrid platform for synthesis entails intricate considerations in system design and architecture, encompassing security, scalability, and distributed state management, while managing data distributed across multiple at-rest locations presents formidable hurdles such as segmentation complexities, data discovery challenges, and stringent security requirements [43].

Hybrid solutions can arise in situations where a data synthesis model is trained in a cloud environment on non-sensitive data, such as synthetic data derived from original datasets. Subsequently, the trained model is transferred to an on-premises solution where it can be further trained and fine-tuned with original data. Hybrid data synthesis tools, which combine cloud and on-premises resources, offer the flexibility and scalability of cloud computing along with the control and security of on-premises infrastructure, making them a potentially ideal solution for organisations seeking a balanced approach to data management. However, ensuring secure access to data for model training requires

adherence to industry best practices, alongside addressing challenges such as training at source, temporary privilege escalation, data leakage prevention, and encryption of intermediate data [43].

4.5 Synthetic Data Generation Steps

4.5.1 Data Collection and Data Preparation

The first step involves data collection and preparation, which can be further divided into two substeps: data cleaning, and data analysis [72]. Research underscores the pivotal role of data preprocessing in enhancing data quality and minimising bias, revealing a historical lack of investment in this area among non-technology companies [43]. This meticulous process requires significant attention to detail to ensure subsequent smooth operations, aiming to refine raw, often incomplete and inconsistent data into a clean, comprehensible dataset suitable for training with generator models [72].

Data collection begins with goal setting—the organisation must figure out what the purpose of data processing is or what the desired result is. Thereafter, it is possible to assess which subsets of data are needed to fulfill these goal. At its core, synthetic data diverges from real data while possessing equivalent statistical characteristics [29]. However, there are also methods for generating synthetic data that do not use real data. Section 3.1 discussed how synthetic data can be created using real data, without any real data, or by combining real data with randomly generated data.

Next, suitable data needs to be acquired. It is important whether the data is already available or needs to be acquired from external sources. Data collection entails gathering and measuring featured variables. This is crucial for ensuring accuracy, enabling informed decision-making, and identifying patterns or trends within the dataset [72]. Ensuring high-quality and diverse training data involves collecting up-to-date data from multiple sources, validating and cleaning it [43].

Data cleaning, a critical preprocessing step, involves standardising and normalising formats, addressing missing values and outliers, and sometimes labelling and annotating data [43]. This process is vital for addressing missing values, noise, outliers, and inconsistent data, while data analysis and extraction entail reviewing and refining the dataset, necessitating additional attention to handle edge cases and outliers for ensuring the robustness of the synthetic data generated [43, 72].

The next substep of data preparation is analysis, since understanding the characteristics, structure, and distribution of the original data is crucial. This substep involves data exploration, visualisation, and analysis to identify patterns and dependencies [72]. Data quality requirements encompass consistency, accuracy, integrity, timeliness, interpretability, and believability, highlighting the importance of both pre- and post-processing in ensuring that a generative model can effectively handle realistic datasets [64, 72].

The level of dataset bias largely depends on the preprocessing of the data. While data-driven models excel at solving real-world problems, acquiring relevant data can be challenging, and the need for more diverse data often leads to the creation of data samples based on existing metadata—a common practice that can inadvertently introduce bias into the dataset [90]. Navigating the varied landscape of bias identification and mitigation presents challenges in determining the most suitable approach for specific contexts, lacking a universally applicable solution [91].

Current approaches for mitigating data bias can be broadly categorised into three groups: rebalancing, algorithmic, and post-processing approaches [64]. According to Juwara et al. [64], rebalancing aims to create datasets that better reflect the true population, starting with biased data and striving to produce balanced datasets that approximate the underlying population data, with matching and stratification typically applied before statistical modelling to establish balance and facilitate appropriate comparisons among baseline covariates. Propensity score adjustment, on the other hand, is employed during the analysis stage by directly incorporating the scores as weights into the regression model, addressing bias in a more nuanced manner [64].

4.5.2 Model Selection and Training

Model selection starts with choosing a suitable approach or generative model based on the data characteristics, such as dimensionality, complexity, and distributional properties. The complexity and nature of the use case, along with the associated task (e.g., capturing statistics, preserving query answers, classification), as well as the chosen evaluation criteria, may restrict the model choice [43]. As previously noted, no universal synthesis method exists, and this presents challenges in different data domains.

Different domains like single table, time-series, sequential, and multi-table may require specific knowledge to be encoded into the model to achieve higher utility and scalability [43]. The suitable model also needs to be trained, and this requires clean data. Addressing the effective deployment of AI approaches like ML and deep learning algorithms in real-world applications, data emerges as a crucial component, yet privacy concerns and limited public

accessibility of many datasets pose challenges in training ML models, which heavily rely on extensive training data [17].

The training step may involve parameter tuning and optimisation to enhance model performance. Given the uncertainty surrounding how a model would perform beforehand, it is often necessary to train and compare different models to determine the most suitable one for a particular use case [43]. Synthetic data complements real data during model training by addressing gaps in coverage, with models trained using augmented data occasionally outperforming those trained solely on real data [41].

4.5.3 Data Generation

Following the training phase, models enable inference, empowering users to generate synthetic data based on specified configurations, including sample quantity, value range, and distribution, with the objective of aligning the data distribution in real data with that of the generated data [72]. Employing ML models to capture the structure and statistical distribution of the original data (A) facilitates the generation of a synthetic dataset (A') from it, with the preservation of the statistical properties of A in A' enabling data analysis to use A' in analyses with similar results as if they were using A [30].

Over the past few years, extensive research has been conducted on synthetic data generation and evaluation, primarily dividing into two main streams [17]. According to Ling et al. [17], one stream focuses on the advancement of novel synthetic data generation algorithms [74, 92], and the other stream involves the application of established generation methods to diverse datasets across various domains, with subsequent evaluation using different metrics [93].

When analysts engage with synthetic datasets, they should yield analysis outcomes akin to those from real data [29]. Leveraging domain-specific knowledge from disciplines such as computer graphics, physics, and cognitive science can enrich the realism of synthetic data [34]. A comprehensive grasp of physical laws and cognitive processes enables more accurate generation of diverse scenarios, thereby bridging the gap between synthetic data and real-world contexts [34]. Synthetic data serves to either substitute gathered data by retaining or emulating its characteristics, or to complement collected data, enhancing its comprehensiveness or bolstering privacy safeguards [5].

4.5.4 Evaluation

Evaluation means assessing the quality and utility of the synthetic data generated. This step involves measuring various metrics, such as similarity to the original data, privacy [17], and utility for specific tasks [43]. Evaluation is a vital yet challenging aspect of synthetic data generation as synthetic data approximates real data but may not capture all outliers, which can be crucial. The quality of synthetic data is closely tied to the input data and the generation model used. Biases in the source data can transfer to the synthetic data, underscoring the need for rigorous validation. Evaluation assesses quality and compares model performance using predefined metrics [72].

Evaluating the real-world applicability of synthetic data involves several critical metrics [17]. Measuring the desired properties of synthetic data, such as utility, fidelity, diversity, authenticity, and fairness, is essential for building trust [43]. Privacy is paramount in ensuring that identities from the original dataset cannot be discerned in the synthetic data, and simultaneously, similarity assessment is crucial as the synthetic data must accurately capture the information present in the original data [17]. However, these properties are difficult to quantify, and defining them poses challenges [43].

Evaluation of the generated data can be conducted to assess quality and compare model performance using specific metrics [72]. Auditing the model is essential to ensure the privacy of synthetic data [43]. Offering controlled generation, synthetic data allows precise manipulation of its properties and contents [41]. For instance, in testing fraud detection algorithms, known fraudulent patterns can be injected for evaluation [41]. Rigorous manual checks are essential to ensure accuracy before integrating synthetic data into ML models.

During the training phase, preprocessed data acts as input, aiming to harmonise various features and characteristics—including data types, value ranges, patterns, and distributions—with those of the output schema [72]. Research has demonstrated that models trained on a hybrid dataset, comprising both synthetic and real data, and tested on real data tend to outperform those trained solely on real data [19]. In the realm of deep learning, pre-training has gained widespread adoption to boost model performance, especially in scenarios where training data for a specific task is limited [57].

Integrating synthetic data into enterprise data systems presents various challenges, including versioning complexities, compatibility issues, and efficient allocation of computational resources [43]. This step may also involve adjusting model parameters, incorporating expert feedback, and improving the data generation pipeline to enhance the quality and utility of synthetic data. Also, efficient data versioning, synchronisation, metadata management,

model portability, and resource allocation management are critical for seamless synthetic data integration [43].

5. Legal Implications of Synthetic Data

5.1 Pseudonymity and Anonymity

In the realm of data synthesis, the requirement for real world data is a common prerequisite. Given that these datasets often contain personal information, the generation of synthetic data must adhere to strict data protection rules. In the EU, the processing of personal data is primarily governed by the Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation; GDPR) [94]. This legislation sets the framework for ensuring that personal information is handled safely and securely.

While GDPR lays down general rules to protect natural persons in relation to the processing of personal data, Directive (EU) 2016/680 of the European Parliament and of the Council [95] lays down the specific rules regarding the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security, respecting the specific nature of those activities. Directive (EU) 2016/680 Recital 11 clarifies that if such a body or entity processes personal data for purposes other than for the purposes of the Directive (EU) 2016/680, GDPR applies. Since this thesis primarily discusses the sharing of data from LEAs with external developers or researchers, which relates more to activities and objectives other than those mentioned in Directive 2016/680, this section will mainly focus on the requirements of the GDPR and the associated challenges. In addition to the aforementioned EU legal acts, data processing in Estonia is also regulated nationally by the Personal Data Protection Act (PDPA) [96] and the Public Information Act (PIA) [97], along with several sector-specific special laws.

While synthetic data addresses challenges related to data access and privacy, it also introduces legal complexities under data protection laws. Data synthesis involves various methods and tools, leading to significant variations in the legal implications of the data. A central legal issue is classifying synthetic data as either anonymous or pseudonymous, each defined distinctly under the GDPR. This distinction is crucial because the processing of personal data versus anonymised data carries different legal ramifications and potential risks to individuals' rights and freedoms [98]. Unfortunately, the criteria for determining whether synthetic data qualifies as personal or anonymous data under GDPR requires case-by-case analysis. This classification depends on multiple factors, including the synthetic

data generation method, the synthesiser model, and the original training data. As the field of synthetic data is still evolving, each dataset derived from personal data requires individual assessment to ascertain if it meets GDPR's criteria for anonymity.

Determining whether data is pseudonymous or anonymous involves assessing probabilities of re-identification and inference risks, and as trust in anonymisation wanes, defining the desired model within data protection law becomes increasingly crucial [28]. Therefore, anonymisation processes must render data subjects unidentifiable, meeting high standards set by European data protection legislation [1]. The ongoing debate requires careful consideration of balancing data utility and protection within legal frameworks [28].

GDPR Article 4(1) stipulates that the term 'personal data' *means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.* When generating synthetic data based on personal data, adherence to the GDPR requirements is essential until the data reaches at least an anonymised state.

The GDPR defines pseudonymous data as personal data that cannot be attributed to a specific data subject without additional information. GDPR Article 4(5) states that pseudonymisation is the processing of personal data in such a manner that it can no longer be attributed to a specific data subject without the use of additional information, provided that this 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. Furthermore, GDPR clarifies in Recital 26 that the principles of data protection should extend to any data concerning a known or identifiable individual. Properly generated synthetic data often avoids direct attribution to individuals and traditional identifiability tests, but it may still be considered pseudonymous if it retains key properties or patterns that closely mirror the original dataset [30].

According to GDPR Recital 26, data protection principles do not apply to anonymous information, which refers to data not related to an identifiable individual or to data anonymised to the extent that individual identification is no longer possible [30]. Anonymisation processes, such as grouping variables, omitting variables, or adding noise, often lead to significant data loss, which can negatively impact their use, particularly in scientific research [54]. Anonymisation is often critiqued for reducing data utility and providing sub-optimal privacy outcomes, leading to the rise of DP as a superior alternative that promises

near-ideal privacy, though at the cost of utility, and in the domain of data synthesis, ML-enabled synthetic data, which approximates but does not exactly replicate the original data, can be further enhanced by integrating it with DP to optimally balance privacy and utility [99].

The decision between using anonymised or synthetic data depends on the specific needs and trade-offs considered by researchers [54]. Advocates argue that properly generated synthetic data without direct mappings to individuals qualifies as anonymous, providing robust data protection, but critics contend that it may still enable one-to-one relationships or facilitate sensitive information inference, challenging its anonymity status [30]. It is essential to assess whether the synthetic data has achieved anonymity or if it is merely pseudonymised data. While synthetic data offers potential advantages for preserving the statistical integrity of original data and enhancing privacy, it necessitates careful examination to ensure it meets the stringent anonymity criteria set by GDPR. The ongoing debate on whether synthetic data can be considered truly anonymous or if it remains a form of pseudonymised data highlights the need for a thorough legal and technical analysis to ensure compliance with data protection laws [30]. This evolving landscape indicates a growing need for comprehensive data protection strategies that incorporate both traditional and innovative PETs to navigate the complexities of modern data use.

5.2 Legal Basis for Synthesis

From a data protection perspective, in addition to the question of whether synthetic data qualifies as anonymous or pseudonymous, there is another challenge—the legal basis for synthesising real-life data that contains personal information. GDPR Article 5 establishes the principles relating to the processing of personal data. According to GDPR Article 5(1)(a), personal data must be processed lawfully, fairly, and in a transparent manner in relation to the data subject. GDPR Article 6(1) stipulates that personal data processing is lawful only if it satisfies at least one of the following conditions: (a) the data subject has consented to the processing for one or more specific purposes; (b) the processing is necessary for contract performance or to take steps at the request of the data subject before entering a contract; (c) it is necessary for compliance with a legal obligation; (d) it is required to protect the vital interests of the data subject or another person; (e) it is necessary for performing a task in the public interest or exercising official authority; and (f) it serves the legitimate interests of the controller or a third party, provided these interests do not override the data subject's fundamental rights and freedoms. Point (f) does not apply to processing carried out by public authorities in the performance of their official duties.

According to GDPR Article 6(2), member states may enact specific provisions to refine the application of GDPR rules, particularly concerning compliance with points (c) and (e), by defining more precise processing requirements and other measures to ensure lawful and fair processing. GDPR Article 6(3) stipulates that the legal basis for processing under points (c) and (e) must be established through Union law or the law of the member state to which the controller is subject. Therefore, if PPA wishes to perform data analysis necessary for fulfilling its lawful tasks, the legal basis must derive from EU or Estonian legislation. For example, if a LEA intends to perform data synthesis based on personal data to use synthetic data for purposes such as developing or testing an information system, such a legal basis should be stipulated in the legislation regulating the activities of the LEA. In their operations, PPA must adhere to the sector-specific legislation, for example, Police and Border Guard Act, Law Enforcement Act, Code of Misdemeanour Procedure, Peanal Code, and internal organisational rules. PPA's long-term development objectives are outlined in the internal security development plan (plan for 2020-2030 [14]).

Among the data-related legislation, the most significant national laws in Estonia are PDPA and PIA. PDPA sets out standards for the implementation of GDPR and for the transposition of Directive (EU) 2016/680. PDPA specifies requirements for national processing of personal data. The purpose of PIA is to ensure public access to information intended for public use, enabling society to monitor the performance of public duties in line with democratic and open society principles. The referenced legislation also regulates matters related to national databases and open data.

Currently, the use of synthetic data is not regulated at the national level. Therefore, data synthesis requires case-by-case interpretation of existing norms, which creates an administrative burden. Given that the development and testing of information systems are relevant to the entire public sector in Estonia, not just LEAs, it would be prudent to regulate the use of synthetic data in legislation such as PDPA or PIA. This would enable a broader range of public sector institutions to leverage the potential of synthetic data in the development of technological processes.

PDPA Section 6 establishes requirements for the processing of personal data for scientific research needs. According to PDPA Section 6(1), personal data may be processed without the consent of the data subject for scientific research, particularly in a pseudonymised form or one that provides equivalent protection, and must be replaced by pseudonymised data or data in a similarly protective format before its transmission for processing. However, the DPDA does not define what qualifies as research within the meaning of the said law. In principle, the interpretation of the PDPA Section 6(1) would allow for the issuance of synthetic data for research purposes, as synthetic data is at least in pseudonymised

form if not anonymous. For this, it is necessary for the researcher to have an appropriate legal basis for processing the data. Generally, appropriate legal bases can be derived from GDPR Article 6(1)(a), which involves the consent of the data subject, or 6(1)(f), which pertains to the legitimate interest of the researcher. However, in order to make a decision on releasing data for research purposes, the mere presence of an appropriate legal basis for the researcher is not sufficient—PPA must also evaluate other principles relating to the processing of personal data as outlined in GDPR Article 5(1), for example, purpose limitation, data minimisation, storage limitation, and confidentiality. Therefore, PPA must assess a wide range of nuances depending on the purposes of data processing and ensure that both personal data and the organisation's strategic information are effectively protected.

5.3 Data Synthesiser as an Artificial Intelligence System

In data synthesis, it is crucial to consider the method used for synthesising the data. When data is processed using technology that qualifies as an AI system, the data processor must adhere not only to data protection requirements but also to the regulations governing AI set forth by the EU such as the Artificial Intellicence Act (AI Act) [100, 101] and AI Liability Directive [102]. According to AI Act Article 1(2)(a), the AI Act establishes harmonised rules for the marketing, implementation, and use of artificial intelligence systems (AI systems).

The AI Act establishes specific requirements for LEAs, but also granting them notable exceptions that include a waiver from conformity assessments and the permission to begin real-world testing of high-risk AI systems without prior authorisation [101]. When assessing whether the synthesiser may be subject to the requirements of the AI Act, it is necessary to evaluate whether it falls under the AI system definition provided by the AI Act. According to the AI Act Article 3(1), *'AI system' is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*. It is also crucial to assess whether a LEA qualifies as an AI system provider or deployer, as the specific requirements for AI operators depend on their role in the AI system lifecycle.

The AI Act adopts a risk-based approach where operators of lower-risk AI systems face lighter requirements, while those involving high-risk AI systems must adhere to stricter compliance standards. Annex III [101] of the AI Act outlines high-risk AI systems used within law enforcement. The high-risk categories include (a) AI systems that are intended

for use by law enforcement or EU institutions to assess the risk of individuals becoming victims of criminal offenses; (b) AI systems that are used by law enforcement or EU bodies as tools akin to polygraphs; (d) AI systems designed to evaluate the reliability of evidence during the investigation or prosecution of crimes; (e) AI systems used by law enforcement to assess the risk of individuals committing or recommitting crimes, not solely based on profiling; and (f) AI systems for profiling individuals, as specified in Article 3(4) of Directive (EU) 2016/680, during the detection, investigation, or prosecution of crimes. These provisions emphasise a balance between technological advancement and legal as well as ethical considerations.

Research conducted by Strmečki and Pejaković-Đipić [103] examined the different methods for data analysis used by LEAs and explored the relationship between privacy and security perceptions related to personal data collection via conventional and AI methods. The initial hypothesis, which posited a causal link between these perceptions, was not supported as no significant correlation was found, and recommendations for future research included distinguishing situations by their security relevance. Additionally, it was observed that individuals are likely influenced more by the type of data rather than the collection method when providing personal data to law enforcement.

5.4 Open Data

According to PIA § 3^1(1), open data refers to the secondary use of public information that is not restricted by law or by procedures established by law. Open data is gaining importance in fostering innovation and driving economic growth, particularly in the realm of developing AI applications, which is a strategic objective for Estonia's digital advancement [104].

Through the Estonian information gateway (EIG), a website allowing access to public information (PIA § 32^1(1)), everyone can access both unrestricted public sector data and licensed data shared by private and third-sector entities [105]. Reuse of the data is permitted based on licenses determined by data providers, allowing for both commercial and non-commercial purposes [105].

This master's thesis addresses open data because they are anonymised data that facilitate easier analysis compared to raw, unprocessed original data. The COVID-19 crisis underscored the value of open data as citizens increasingly demanded accessible and comprehensible information, such as statistics on virus spread [104]. Recognising the significance of open data, the Ministry of Economic Affairs and Communications has made it a priority in its efforts to lead Estonia's digital state development, focusing on

enhancing the quality of open data and promoting systematic data release by public sector entities [104].

While open data enables many opportunities, it also comes with various challenges. The primary issue with open data revolves around two aspects. Firstly, to create open data, the original datasets undergo extensive cleansing, resulting in the loss of significant information. When attempting to analyse different datasets together, there is a lack of connections or no links at all between them, meaning that features enabling correlation have been removed. Therefore, this does not allow open data to be successfully utilised in more complex research. Additionally, the use of open data is not particularly feasible in the development of AIML technology. Secondly, considering the current security situation, we should indeed carefully assess the amount and type of information we make freely accessible to everyone.

As the EU moves towards greater data transparency, it is crucial to consider whether all disclosure obligations and requirements are reasonable, especially in the context of data economy and security. In a prior study [106], correlations have been identified between the societal implications of the digital economy and its impacts on cybersecurity. These interconnections suggest that addressing cyber threats should not solely focus on technical aspects, but also on social factors [106].

In recent decades, the nature of information technology security incidents has evolved, moving from sporadic attacks on information systems to deliberate, focused, and intricate cyber threats capable of targeting individuals, institutions, or even entire nations [107]. The current state of war in Europe adds even greater pressure to security efforts, necessitating heightened vigilance and proactive measures to ensure safety and protection [108]. Given the wartime situation in Europe, it has become even more critical to carefully determine who should have access to specific datasets and what values these datasets should contain.

6. Expert Surveys

6.1 Public Sector Experts

This subsection presents the results of an expert survey, with the questions detailed in Appendix 4. Both public sector experts consented to be named in this thesis. The Ministry of Economic Affairs and Communications is one of Estonia's primary government institutions that aims to utilise data for the state's benefit while preserving privacy. The Data Protection Inspectorate aims to ensure that people's privacy is protected during data analysis. Thus, the results of the two surveys provide insights from both mentioned perspectives.

Assesment of Synthetic Data. According to Ott Velsberg, an expert at the Ministry of Economic Affairs and Communications, synthetic data plays a crucial role in anonymising original, sensitive data such as delicate personal information or competitively sensitive data. Additionally, the expert finds that it is necessary to analyse how data integrity is ensured, including the irreversible process—meaning that it would no longer be possible to derive the original data from synthetic data. According to the expert, synthetic data represents a significant future direction for the Ministry of Economic Affairs and Communications, which has previously conducted projects in this area.

In the opinion of Urmo Parm, an expert at the Data Protection Inspectorate, synthetic data has become an indispensable part of today's data protection landscape, essential for various needs such as testing information systems and training AI. Regarding the use of personalised data, there must always be a legal basis, which can often be difficult to establish, especially in the public sector where any processing of personal data must have a legal basis derived from the law.

Potential of Synthetic Data. Ott Velsberg sees the potential of synthetic data in the provision of research and also other purposes like statistical analysis where statistical similarity is ensured, and data creation. The expert highlighted the considerable potential of synthetic data in facilitating research and other related applications. He noted that synthetic data is particularly valuable in statistical analysis due to its ability to ensure statistical similarity with real-world data. Additionally, synthetic data plays a pivotal

role in data creation, enabling researchers to generate robust datasets when actual data is unavailable or confidentiality constraints apply.

Urmo Parm sees the greatest potential for synthetic data in the development of AI and related technologies, which require vast amounts of training data to achieve accurate and unbiased outputs. While algorithms for mass use are often trained on public data, specific AI solutions for organisations can only be trained using data generated through their own operational processes. He added that another area of potential is market-wide analyses and studies, which currently always utilise original raw data. The process is complex, particularly when personal data is involved, often requiring permissions to be obtained. The use of synthetic data offers statistical weights similar to the original data, and simpler analyses could be performed with them.

Shortcomings of Synthetic Data. Ott Velsberg found that the main shortcomings of synthetic data are the statistical properties and the privacy risk. He considers under statistical properties whether synthetic data has sufficiently similar statistical characteristics; this also needs to be validated. The expert believes that one of the biggest challenges in the broader adoption of synthetic data is the general lack of knowledge, for example, regarding what is sufficient, how to validate data, which methodologies should be used for creating synthetic data, whether there is signal in synthetic data. On the other hand, problems such as overfitting and related issues also arise. In addition, he finds that concerns with privacy and anonymisation are linked, for instance, whether it is ensured that a person or party cannot be identified or if the synthetic data are merely pseudonymised. The expert sees that the legal classification of synthetic data is one of the significant challenges in the wider adoption of synthetic data. Another aspect is the generally low awareness and expertise on the topic of synthetic data.

Urmo Parm clarifies that there are also bottlenecks in the use of synthetic data, primarily in the data creation process. The goal is to retain the statistical properties of the original data in the synthetic data, necessitating the processing of original data. For personal data, such processing must have a legal basis. The private sector may use legitimate interest, assuming the interests of the organisation and the individual have been properly balanced beforehand. In the public sector, it may be prudent to create a legal mandate at the legislative level to allow certain datasets to have legal synthetic clones.

Obstacles to the Adoption of Synthetic Data. According to Ott Velsberg, the biggest obstacles to the adoption of synthetic data are the protection of original data, ensuring the

irreversibility and integrity of anonymised data, and assessing the interaction of synthetic data.

Urmo Pram reiterated that today there is no suitable legal basis for creating synthetic data from original data.

Previous Experiences with Synthetic Data. According to Ott Velsberg, the ministry has primarily encountered synthetic data in the creation of statistical and language datasets. A synthetic dataset has also been created for the development of the synthetic twin of the traffic registry.

According to Urmo Parm, the Data Protection Inspectorate do not have experience with the use of synthetic data, however, they have participated in various discussions regarding synthetic data. He knows that as a result of the study on privacy technologies, the government was advised to consider the concept of a digital twin, but whether there has been any progress with this idea is currently unknown.

6.2 Private Sector Experts

This subsection aims to provide an overview of the current state of synthetic data from the perspective of experts deeply knowledgeable in the field. It is organised around the viewpoints of three experts regarding various aspects of synthetic data, with the survey questions detailed in Appendix 4.

Potential of Synthetic Data. The experts discuss the various benefits and applications of synthetic data across different industries. Expert 1 sees synthetic data as pivotal in generating test data and enhancing access to datasets in restricted fields. The most immediately useful use-case for synthetic data in the expert's opinion could be generating large amounts of synthetic test data. A second major use case is providing access for researchers and the general public to data sets they normally would not have access to. Expert 1 argues, that this can significantly increase the amount of science done in areas where data access has been the limiting factor, such as health care, law enforcement, and finance. Additionally, synthetic data can help comply with privacy regulations and decrease the risk of data leakage by reducing the attack surface through limiting the number of systems that have access to the original sensitive data. Expert 1 notes that synthetic data can also aid in achieving compliance with privacy regulations.

Expert 2, however, expresses concerns about the effectiveness of synthetic data in preserving privacy when used for analysis, suggesting that the balance between utility and privacy might not always be achievable. Expert 2 recommends using DP to safeguard synthesised datasets as there have been successful inference attacks on synthesised datasets that do not use DP. In the expert's opinion DP should also be used when using synthetic data to publish open data.

Expert 3 highlights the transformative potential in industries with strict privacy needs, envisioning a market for synthetic data that supports data-poor startups. The impact is likely to be high in industries where data sharing has traditionally been particularly difficult from a privacy protection perspective, such as the healthcare, financial and defense industries. Expert 3 also mention the novel use of synthetic characters like synthetic actors in Japanese advertising to avoid scandals and reduce costs, pointing to the broader implications for industries reliant on public perception.

Complexities Regarding Synthetic Data. When addressing the complexities of creating synthetic data that accurately mirrors real-world scenarios, Expert 1 outlines the inherent trade-off between data privacy versus the similarity of synthetic data to real data. A synthetic dataset that maximises privacy will necessarily be more generic and less similar to the real data, when in counterpoint compared to a synthetic dataset that attempts to emulate the real data perfectly will necessarily leak something about the original dataset. Expert 1 therefore argues that maximising privacy dilutes the real world applicability of the data, necessitating a careful balance tailored to specific use cases.

Expert 2 raises legal and ethical issues, particularly the risk of inadequately synthesising a real individuals' data, which could hinder compliance with data protection laws. Training a model for data synthesis is classified as data processing and may also constitute secondary data usage; thus, it is uncertain if organisations can legally use real data for this purpose in practice. Expert 2 adds that another open question is that of accidentally creating data of a real person: Expert 2 found that another important issue to investigate in relation to synthetic data is the possibility of accidentally synthesising real data about a person who actually exists (i) whose data in the original dataset; or (ii) whose data belongs to the original dataset not in the training set that was used for creating the synthesis model; (iii) someone whose data does not belong to the original dataset (i.e., does not have a speeding ticket, a criminal record or a specific disease).

Expert 3 discusses technical challenges like catastrophic forgetting, where updating models with new data without access to the original dataset can lead to significant information

loss, affecting the model's accuracy and reliability. For example, the assumption is that the model MA is initially trained using the dataset DA. If there is a need to update MA using a new dataset DB, but one can no longer access DA for various reasons, then updating MA only with DB might result in the loss of unique statistical properties from DA. This loss is referred to as catastrophic forgetting.

Different technologies and methodologies. The creation of high-quality synthetic datasets is influenced by the choice of technology and methodology, as explained by the experts. Expert 1 emphasises that the quality of synthetic data heavily depends on the quality of the training data, indicating that poor input cannot produce high-quality synthetic outputs. In the expert's opinion, creating synthetic data with high utility starts with collecting high-quality training data. High-quality outputs cannot be derived from low-quality inputs. Additionally, most synthesis models need large volumes of training data, making the acquisition of such data a crucial challenge in data synthesis. However, as similar issues affect other machine learning systems, data collection methods are rapidly advancing.

Expert 2 notes that the choice of technology for data synthesis should be purpose-driven, suited to the specific needs of the data synthesis task, whether for testing or analysis. This also allows for the use of less accurate methods in scenarios such as generating test data or in environments with limited computational resources.

Expert 3 points to the use of advanced models like GANs and methodologies such as federated learning, which enhance data privacy while collaborating across datasets. The quality of a model hinges on the quantity and diversity of the data used. Federated learning enhances privacy in ML by enabling data owners to collaborate without sharing their data, and privacy can be further improved—albeit at the expense of utility—by combining secret computation and differential privacy.

Synthetic data as a booster to ML models and algorithms. In terms of enhancing ML models, Expert 1 compares the inclusion of synthetic data in training datasets to sophisticated noise addition, which can help prevent overfitting. Expert 2 supports this view by acknowledging the value of synthetic data in diversifying training datasets.

Expert 3 extends this idea to data augmentation, where synthetic data is used to create additional data points when existing data is insufficient, thereby enhancing the robustness of machine learning models. The process of generating new data similar to existing data

when there is a shortage is known as up-sampling or data augmentation, which is useful for purposes such as validation and comparison. In the opinion of the expert, synthetic data is effectively utilised in these scenarios and, although not an ideal example, can also serve to simulate attacks on systems.

Future of Synthetic Data. Regarding the future of synthetic data, Expert 1 is optimistic about the integration of synthetic data into mainstream applications, drawing parallels to the acceptance of large language models like ChatGPT. Expert 1 believes that the privacy-preserving nature of synthetic data will boost its acceptance. Similarly the rising importance of privacy and the protecting of private data gives a perfect background to introduce synthetic data generation into existing systems, as one of the main selling points of data synthesis is that synthetic data preserves the privacy of the input data.

Expert 2, however, cautions against over-reliance on synthetic data, similar to past over-confidence in data anonymisation techniques. In combination with DP, the expert finds that synthetic data has potential for creating test data or data for education. However, according to the expert, it is still too early to draw conclusions about whether synthetic data will also be suitable for research work.

Expert 3 suggests a paradigm shift towards storing only the models used to generate data, reducing the risks associated with data breaches and minimising storage costs. Explaining in more detail, the idea is to save only the models that generate the original-like data and delete the raw data, rather than keep them intact. This allows companies to reduce the amount of data stored and also reduces the risk of direct leakage of customer data.

Ethical Considerations. The ethical implications of synthetic data, particularly concerning biases and fairness, are complex. Expert 1 references initiatives aimed at creating safe and trustworthy AI but acknowledges challenges in ensuring that synthetic data does not perpetuate existing biases or infringe on copyright laws. According to the expert, there are numerous initiatives working towards safe and trustworthy AI (e.g. ECSTAI, NIST's AI Safety Institute, IBM's research team). However, there are also many ethics questions raised in the entertainment field both in the way the AI and specifically data synthesis models are trained, and how their output is used. A notable issue on the training side is that models are trained on copyrighted data and can generate (arguably) copyrighted output. According to the expert, the same issue can arise with respect to privacy, as when synthesis models are trained on personal data, the output may also leak personal data. Using the synthetic data can also come with ethical problems. As an example, the expert pointed

out that last year the Writers Guild of America was on strike partly to protest the use of generative AI in media, as they feared their jobs would be made redundant and they will be replaced by data synthesis models.

Expert 2 warns that synthetic data is likely to inherit and possibly amplify the biases present in the original datasets. Expert 3 discusses technical measures to maintain diversity and quality in synthetic data outputs, which can help mitigate bias. When addressing the so-called mode collapse—the issue of producing a limited variety of outputs—the inception score and Frechet inception score are widely utilised. Additionally, according to Expert 3, multidimensional scaling has also potential by focusing on measuring pairwise distances. To prevent mode collapse, a variety of GAN methods are being explored.

Misconceptions and Myths. Addressing common misconceptions surrounding synthetic data, the experts offer clarifications and real-world insights. Expert 1 corrects the widely held belief that synthetic data is inherently anonymous, pointing out that because it mimics real-world data, there is a significant chance it might inadvertently leak information about the original dataset. Expert 1 highlights the complexity of ensuring that synthetic data does not include personal data accidentally, which remains a challenging area in privacy engineering. Accordin to the expert, combating such problems is hard, but they can often be mitigated by using standard practices in privacy engineering.

Expert 2 challenges the assumption that synthetic data must be derived from real data, emphasising that synthetic data can also be created based on random or rule-based generation methods. Expert 3 dispels the notion that synthetic data can enhance the original data's value, stressing that while synthetic data maintains statistical properties as a whole, individual data points often do not exceed the quality of the original dataset.

The Best Field for Synthesis. In discussing industries where synthetic data has shown promise, Expert 1 mentions its significant impact in creative fields such as art, movies, and music, where generative models are used to create new content from text prompts or to replicate voices of deceased actors. Expert 1 reflect on the potential of these technologies to change content creation fundamentally.

Expert 2 highlights practical applications in health studies in Canada, where synthetic data facilitates feasibility studies with easier-to-obtain confidentiality agreements compared to real data. Expert 2 also notes the use of synthetic data in hackathons and machine learning competitions. Expert 3 provides examples from healthcare and autonomous

vehicle development.

Advice for Adopting Synthetic Data. The experts share strategic advice for organisations considering incorporating synthetic data into their processes. Expert 1 advises a cautious approach, emphasising the need for thorough analysis and consultation with experts to avoid potential pitfalls in this relatively new field.

Expert 2 advocates for a clear understanding of the intended use of synthetic data, suggesting that organisations should prioritise privacy and clearly define the characteristics of the synthetic dataset to ensure it serves its intended purpose without compromising data privacy. According to Expert 2, although original data typically offers the best utility, there are compelling reasons to use synthetic data, such as privacy concerns. It is crucial that synthetic data sufficiently differ from the original to ensure privacy while maintaining usability for analysis.

Expert 3 recommends clarifying the objectives, goals, and quality requirements of synthetic data projects upfront and stresses the importance of regular updates and evaluations to keep the models relevant and effective.

Area Requiring Investigation. Finally, the experts identify key research areas that warrant further exploration to advance the field of synthetic data. Expert 1 calls for more research into membership inference attacks, which could reveal whether an individual's data was used in a dataset, potentially undermining privacy protections. Expert 1 also suggest examining the impact of legislative developments, like the EU AI Act, on synthetic data usage.

Expert 2 focuses on developing methods to measure utility and privacy more effectively and ensuring that synthetic data does not accidentally recreate real individuals' identities. Expert 3 highlights the need for ongoing research into attacks and countermeasures related to synthetic data to enhance security and utility in practical applications.

7. Utilisation of Synthetic Data in Law Enforcement

7.1 Data Analysis and Related Challenges

The research on synthetic data and its application in law enforcement underscores the importance of handling sensitive information with utmost confidentiality, a principle governed by various legal acts such as the Public Information Act and the Code of Criminal Procedure, among others. The Ministry of the Interior supervises a substantial number of databases, primarily managed by the PPA, which underscore the necessity of a robust legal justification for data processing.

Experts within the PPA continue to enhance their data analysis skills through regular training sessions, partnerships with technical bodies like IT and Development Centre at the Estonian Ministry of the Interior (SMIT), and continuous professional development. They leverage advanced tools like WebFOCUS, Tableau, programming language R, and Excel to navigate complex data analysis tasks, adapting their approaches to meet the specific demands of their roles. The use of relational databases and SQL queries is common, yet tools are chosen based on the specific requirements of each department.

A significant focus of recent initiatives has been the development of technological solutions like predictive models for deployment of police forces and tools for free text analysis. These innovations are designed to improve operational efficiency and data quality. The predictive model, for instance, uses historical data to anticipate police calls, enhancing PPA's responsiveness to incidents within specific geographic areas.

However, the advancement in data analysis capabilities often outpaces the development of corresponding legal frameworks, leading to potential challenges in data protection and privacy. Experts point out the difficulty in maintaining security without hindering data utility, highlighting the importance of sophisticated strategies and technologies to prevent breaches and misuse.

Another key challenge lies in effectively communicating analysis results to non-technical stakeholders, transforming complex data into comprehensible narratives that can influence decision-making processes. This often requires a deep understanding of business processes and data generation, areas where gaps in knowledge can lead to miscommunication and ineffective solutions.

Resource constraints also pose a problem, particularly in terms of budget allocations for necessary human resources, which are crucial for managing extensive data analysis tasks. Moreover, the architectural integration of data systems and the safe communication between databases remain complex issues that require meticulous attention to both technical and legal standards.

The potential of synthetic data is recognised as a solution to some of these challenges, particularly in protecting individual identities while allowing for meaningful data analysis. However, the use of pseudonymised data, while helpful, still carries the risk of re-identification, underscoring the need for careful consideration of how synthetic data is generated and used.

While synthetic data offers promising solutions for secure and efficient data handling in sensitive sectors like law enforcement, it must be managed within a carefully structured legal and technical framework to ensure both effectiveness and compliance with privacy standards. The ongoing challenges of legal justification, resource allocation, and system integration underline the complexity of data management in a rapidly evolving digital landscape.

7.2 The Use of Synthetic Data in LEAs

The research on the use of synthetic data within LEAs has revealed only a few use cases. Synthetic data has been employed in two distinct cases in PPA to enhance data management and analysis despite limited experience in its creation.

In the first instance, synthetic data was utilised to substantially increase the volume of data available for statistical analysis in a rare situation where traditional data extraction methods were not applicable. The expert described this use case as unusual, employing a method akin to bootstrapping to simulate a process model and gather adequate statistical metrics. This approach was necessary due to the insufficient historical data which hindered obtaining reliable statistics.

In another scenario, synthetic data played a crucial role in a procurement process by creating a sample dataset derived from general statistics. This was further synthesised using a specially developed algorithm, a simple number mixer, to ensure the new data adhered to the same distribution patterns. The dataset was then made public, a decision driven by the need to share data while also obfuscating specific details to address security concerns. This method of data synthesis provided a balance between transparency and privacy.

Moreover, experts have engaged in international collaboration projects such as LAGO[1], STARLIGHT[2], and METICOS[3] to further explore the potential of synthetic data. These projects often face challenges due to the complex nature of analysing sensitive data like that of the PPA. Alternatives such as using open data in place of original datasets have been considered, particularly in initiatives like METICOS.

The legal framework surrounding synthetic data remains underdeveloped and does not specifically address its use, creating a grey area that also impacts the private sector. This lack of a clear legal structure underscores the need for further research and development of tools capable of generating and handling synthetic data effectively.

7.3 Potential of Synthetic Data

The potential of synthetic data has been underscored in various sectors, highlighting its capability to transform practices across development, testing, research, education, and the creation of digital twins. Recent insights from interviews and research suggest that synthetic data holds promise in reducing the duplication of efforts in data creation and enhancing privacy. For instance, synthetic data is viewed as a PET under GDPR guidelines, offering a method to handle data that mitigates privacy breach risks. A particular highlight was an expert's comment that synthetic data could safeguard individual privacy since the data used does not directly correspond to personal data, yet maintains underlying patterns.

However, experts also acknowledged some challenges in the use of synthetic data, particularly its application in scenarios where data content is critical, such as in specific research settings. The need for a certified synthesiser was emphasised to ensure legal compliance and reliability in its application. The potential of the synthesis process could be significantly increased if standardisation of its use were partially achieved.

Furthermore, the discussions revealed practical applications of synthetic data in educational contexts, such as in law enforcement training, where synthetic versions of organisational data systems are employed for training without risking data integrity. The use of synthetic data in creating digital twins also promises more efficient data processing and outcome accuracy compared to traditional statistical methods. This would allow for the analysis of various aspects by combining information from multiple databases.

Looking forward, the strategic adoption of synthetic data is seen as a key component in

[1] https://lago-europe.eu
[2] https://www.starlight-h2020.eu/about
[3] https://meticos-project.eu/about/

advancing internal security measures, especially with the Ministry of Interior's aim to integrate smart technologies and AI by 2030. The potential for synthetic data to facilitate the creation of secure, efficient, and innovative data handling processes in cloud-based environments is particularly promising, indicating a shift towards more dynamic and less resource-intensive methodologies in system development and beyond.

7.4 Obstacles of Synthetic Data and its Adoption

The adoption of synthetic data faces numerous obstacles stemming from a combination of technical, legal, and organisational factors. Data collection is generally targeted for specific uses, and current regulations do not typically encompass the secondary use of data for synthesis. This lack of explicit permission complicates the issuance of synthetic data for research purposes. Moreover, existing forecasts and analyses within organisations are based solely on real data, as endorsed by internal regulations, which do not extend to the creation and use of synthetic data.

The legal environment significantly influences the use of synthetic data. Under the GDPR, synthetic data is acknowledged as a PET, designed to minimise the impact of processing personal data. Despite this, the implementation of synthetic data is curtailed by the need to validate the reliability of data synthesisers and to address ethical considerations, particularly those concerning AI. Such factors highlight a prevailing conservatism within law enforcement and other legal frameworks that adopt a risk-based approach.

Experts suggest that synthetic data could theoretically function as anonymous data exempt from personal data protection laws, which would simplify many processes if true anonymity were guaranteed. However, this potential is undermined by technical challenges, such as ensuring synthesisers do not inadvertently disclose original data, and legal ambiguities concerning the status of synthetic data. There is no definitive guideline to determine when data derived from personal data cease to be personal themselves, leaving a grey area in legal interpretation.

Furthermore, even encrypted or synthesised data can inadvertently reveal information about the original data set. This is seen in cases where synthetic data must avoid replicating exact records from the original dataset, potentially allowing individuals to deduce which data were omitted based on expected probabilistic patterns. Such complexities indicate that synthetic data can still carry traces of the original information, thus complicating claims of complete anonymisation.

In practical terms, the broader adoption of synthetic data within organisations is hindered by

a lack of knowledge, skills, and appropriate tools. While on-premise solutions are favoured for their security, cloud-based solutions are not entirely dismissed but are constrained to specific, secure servers to avoid data breaches. The path to wider acceptance of synthetic data also involves collaborative research and potential regulatory adjustments that might ease these barriers.

While synthetic data holds significant potential for various applications by offering a form of data that ostensibly circumvents privacy concerns, its practical use is heavily restricted by a myriad of legal, technical, and cultural challenges. These issues must be addressed to facilitate broader adoption, particularly in sectors like law enforcement and public administration that are traditionally less inclined towards innovative data practices.

7.5 Development and Testing of Information System

In the realm of information system development and testing, one of the primary hurdles identified involves the management and sharing of data, particularly when dealing with external entities like developers. Research and interviews highlight significant time consumption in generating test data, with experts pointing out the inefficiency of creating non-reusable data combinations during system testing. A notable suggestion from these discussions is the potential development of a universal tool that could generate test data resembling real-life scenarios, thereby reducing the need to manually craft data sets for various test cases.

Experts also emphasised the challenges faced when testing systems with sensitive data, such as biometrics. Legal and ethical concerns frequently arise, especially around the usage and processing of personal data. The necessity to obtain and manage consent from individuals whose data is used poses a continuous ethical and legal challenge, complicating the testing processes. This issue is prevalent not only in one jurisdiction but is a common challenge globally, underscoring the need for innovative solutions to navigate these complexities without breaching legal and ethical boundaries.

Moreover, the interviews underscored the need for legal reforms to better support the testing and development of information systems. Changes to existing laws like the Police and Border Guard Act and the PDPA were suggested to create a more robust legal framework that could facilitate more effective testing while adhering to privacy and security standards. The discussion also touched on the broader implications for both the public and private sectors, where the handling of biometric and other sensitive data remains a contentious issue, often entangled in political and societal debates.

The development and testing of information systems are mired in technical, legal, and ethical challenges. There is a clear recognition of the need for regular review and adaptation of laws and testing principles to keep pace with technological advancements and societal expectations. The call for higher-level legal assurances and the development of tools to simplify testing processes reflect a concerted effort to address these multifaceted issues effectively.

7.6 Sharing Data with External Developers

In the context of sharing data with external developers, collaboration is also undertaken with the SMIT which primarily oversees the process, serving both as an internal development partner and a controller for external engagements. SMIT ensures that data sharing is conducted securely and aligns with the organisation's data protection and confidentiality standards. Occasionally, external providers are enlisted to supplement development tasks, where strict guidelines on data handling are maintained.

The expert elaborated that data preprocessing involves certain protective measures such as anonymisation, which is preferred over pseudonymisation within the organisation. Anonymisation effectively removes identifiable fields, ensuring individuals cannot be recognised, which is a critical step before data are shared with developers. This approach is crucial to maintain confidentiality and adhere to security protocols that govern what developers can and cannot do with the data.

Further challenges emerge in the reuse or secondary processing of shared data, which is strictly prohibited unless for the predefined purposes. This policy aims to guarantee responsible data handling and adherence to usage terms agreed upon initially. The expert highlighted sector-specific challenges, such as the necessity for developers to work onsite due to certain project requirements, which often leads to increased costs and logistical hurdles.

To mitigate some of these challenges, the utilisation of synthetic data was discussed as a potential solution. If synthetic data could be treated with a lower level of security, this would facilitate easier and more flexible usage in software testing and other development phases. Currently, even synthetic data requires rigorous security measures, akin to those used for personal data.

The discussion also covered internal practices at SMIT concerning the outsourcing of developments. The code produced by external developers undergoes a thorough review process to ensure quality. Tools like SONAR are employed to aid in this process, contributing

to higher standards in code quality across various projects. SMIT aims to further enhance this aspect by developing and disseminating tools that assist in code quality checks across different teams.

On an organisational level, SMIT coordinates extensive development activities, aiming for a coherent strategy across all projects to avoid overlap and maximise resource efficiency. With over 400 employees and multiple projects running concurrently, it is crucial to maintain a unified strategy across the entire administrative area, underscoring the importance of strategic planning and advanced tool utilisation in improving operational efficiency and preventing redundancy.

7.7 Sharing Data with Researchers

Currently, research using data from the PPA mainly focuses on internal security. However, the subjects fluctuate based on current societal relevance. Topics such as border control, police operations, crime prevention, and investigation have been prominent, as have issues related to domestic violence, migration crises, security threats, and cybercrime. These shifts reflect broader societal concerns at the time.

Sharing data for research is highly dependent on the specifics of each request and the nature of the information involved. Typically, it is essential to approach each case individually, considering the type of data requested, whether it is merely statistical, intended for internal use, or classified as confidential by the state.

For research, it is recommended to use open data whenever possible. However, the detail in open data often falls short of research needs. Experts suggest that data sharing principles are contingent on the specific nature of each request, considering the type of data involved, its intended use, and its confidentiality status.

Expert opinions highlight a key question: what exactly qualifies as research? This affects data sharing decisions, especially when data is intended for internal use only. In such cases, organisations might limit access to complete documents, offering instead redacted versions to protect sensitive information.

Another approach involves conducting interviews to gather necessary data while avoiding sensitive details. This method, though not ideal for rigorous research, helps control the dissemination of information. When it comes to protective measures, the The Estonian Academy of Security Sciences plays a vital role in training and research, implementing strict protocols to safeguard the confidentiality and security of data.

The volume of research requests the PPA handles annually highlights the importance of data in academic studies, with the main themes reflecting current societal issues like crime prevention, domestic violence, and cybercrime. Despite the eagerness to support student research, logistical and regulatory hurdles often complicate data access. These include the challenges of handling sensitive or big data and adhering to privacy regulations.

The discourse around synthetic data raises concerns about its validity for scientific research. Experts worry that synthesising data might alter essential relationships or introduce biases, which could mislead research findings, particularly in fields reliant on accurate data patterns like ML.

The legal framework also presents significant challenges. For instance, the PIA limits internal use of certain research to a maximum of ten years, after which data becomes public. However, the work methods and strategic information of the PPA do not become outdated as quickly. This situation poses risks, especially as the volume of digital information grows, necessitating updates to legislation that better accommodates modern data challenges.

While there is a robust system in place for managing data requests and ensuring data security, the evolving needs of research and the complexities of legal and regulatory frameworks continue to pose significant challenges. Adjustments, such as clearer definitions of research and potential changes in data protection laws like GDPR or PDPA, could help mitigate these issues and enhance the effectiveness of data use in research.

Figure 1 illustrates the process reflecting the scenario where a researcher wishes to conduct research in the field of law enforcement. The figure was created using PLEAK [109], an advanced analysis tool designed for conducting privacy audits on existing systems and developing new systems with privacy considerations. PLEAK enables the modelling of business processes using the business process model notation (BPMN).

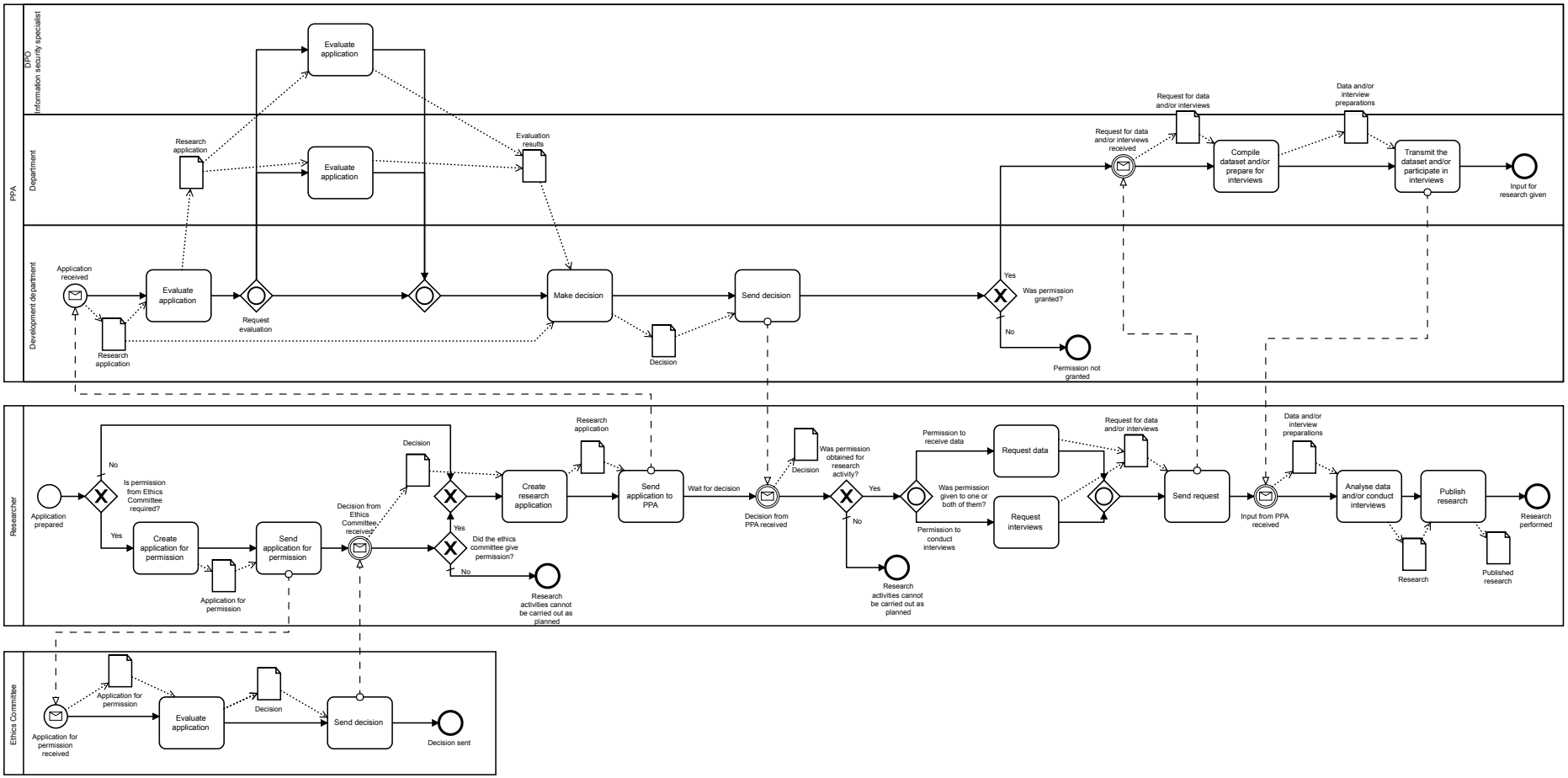


Figure 1. A BPMN diagram of the process of requesting data access for research from PPA

Figure 1 illustrates the process of requesting data access for research from PPA. According to PPA [110], research at the PPA is conducted in line with police priorities and supports studies that enhance police development, and all research must adhere to ethical and data protection requirements. Researchers must submit detailed and signed applications to the PPA, including necessary documents like questionnaires or interview questions [110, 111]. When processing special category personal data (see GDPR Article 9(1)), it is also necessary to obtain permission from an ethics committee or the Data Protection Inspectorate [112]. The referenced permission must be applied for before submitting a research proposal to PPA. PPA's decision is communicated within one month, and completed research must be submitted in PDF format within a month of completion [110]. Therefore, a researcher must account for the time PPA takes to review the research application, particularly if it is also necessary to obtain permission from an ethics committee or the Data Protection Inspectorate for processing special category personal data.

Once the research application has been submitted to the PPA, the development department begins to process it, involving the relevant internal department, such as the department whose area of responsibility relates to the researcher's topic. The evaluation of the application also involves a data protection specialist (DPO) and, if necessary, an information security specialist. Once all parties have provided their assessments, a final decision is made. The decision informs the researcher whether they can proceed with the research as proposed or not. If the research can continue, the researcher will then communicate directly with the relevant internal department, which will provide the necessary data for the research, or with the interviewees.

7.8 Open Data

PPA maintains a cautious approach towards open data, exercising discretion over the release of information, except where legally restricted. A collaborative decision-making process involving various stakeholders, including the dataset owner and open data manager, ensures that the released data is both sensible and safe, focusing heavily on personal data protection. Historical interests from journalists or researchers often influence the content of datasets. The efforts of specialists and analysts ensure the timely and structured publication of data, which is then listed on the Estonian open data portal with complete metadata and provided in a machine-readable format to maximise usability.

PPA's methodical data management reflects a strong commitment to transparency balanced with the need to protect sensitive information. This approach fosters public trust and enhances the usability of open data for various stakeholders. Publishing open data is a burdensome task that requires significant effort to maintain secure, frequently updated

datasets. Once the initial setup for open data publication is complete, however, updates become less resource-intensive, facilitating more efficient data accessibility.

Feedback on the use of published data is considered, though formal evaluation is not performed regularly. Requests for dataset modifications are accommodated if feasible, and any errors found are promptly corrected. PPA receives a handful of inquiries annually about the open data it publishes, highlighting a moderate level of engagement from the public.

There is an ongoing challenge to enhance public awareness and capability in data analytics, as many lack the necessary skills to utilise the available open data effectively. The use of synthetic data for open data publication is seen as riskier compared to other possible use cases due to the potential for broader misuse. Therefore, the disclosure rules are strict, focusing on safeguarding against potential misuse and ensuring data cannot be used to identify individuals.

8. Practical Data Synthesis. A Case Study

8.1 Data Pre-processing and Synthesis

Researchers are trying to understand the nature of synthetic data and its potential. For these reasons, one of the objectives of this thesis was to conduct a data analysis and explore cybercrime data. Subsequently, the goal was to synthesise data based on the model trained on real cybercrime data and perform the exact same analysis on the synthesised data as was done on the original data. This would have allowed for a comparison of the analyses conducted on the original and synthetic data to see what the results would have been. This could have been used to assess the potential and usability of synthetic data and to provide recommendations, also for future research.

For this purpose, real data regarding cybercrime were requested from the PPA, which, unfortunately, they were unable to provide. Therefore, open data have been used as the basis for the data analysis. However, it is important to note that the open data are cleaned, anonymised and aggregated, and therefore, using open data to train the synthesis model will not yield the same results as using real data for training the synthesis model. Despite this, the process of generating synthetic data has been carried out to introduce the data synthesis process to a broader audience.

The process started with data collection and analysis. The raw data (initial dataset) was taken from the open data portal of the Police of the UK Police, and is titled "101 call handling" [113, 114]. Open data published by PPA and Ministry of Justice of the Republic of Estonia were also considered, but open data from the UK Police offered more processing opportunities due to its additional fields and a (slightly) larger amount of data points. The dataset was available under the the Open Government Licence v3.0 and in a CSV (comma-separated values) format, facilitating ease of download and subsequent analysis. The analysed metrics were as follows:

- Count—number of records (entries) analysed;
- Mean—the average value;
- Std—standard deviation, which measures the amount of variation or dispersion in the data;
- Min—minimum value;
- 25%—25th percentile (a quarter of the values lie below this number);

- 50%—middle value of the dataset (median);
- 75%—75th percentile (three quarters of the values lie below this number);
- Max—maximum value.

For the purpose of this thesis, the service prototype for data synthesis [71] (synthesiser prototype) developed in Cybernetica AS was used. The synthesiser prototype that uses a Gaussian Mixture Model utilises the Sharemind HI platform, which in turn is based on Intel's Software Guard Extensions trusted execution environment (TEE) technology, for ensuring security requirements [71]. Sharemind HI enhances data security by allowing owners to encrypt data on-site and upload it for processing without decryption, ensuring that only authorised users can access or query the data, while logging all activities for accountability [115]. TEEs enable privacy-enhancing data synthesis [71] that is crucial in law enforcement settings. Cybernetica AS is currently developing a new synthesis tool that can be used internally within an organisation as an on-premises solution which better meets the expectations of clients who operate in a sensitive sector and do not wish to transmit data outside the organisation.

The synthesiser prototype is engineered to meet the requirements of use cases by safeguarding data even from the service provider via Sharemind HI [71]. Process of data synthesis with the synthesiser prototype unfolds in three stages:

1. The data owner encrypts their CSV file and uploads it to the Sharemind HI server;
2. Server enclaves, that contain code and data for performing security-sensive computations, decrypt the file, develop and train a synthesiser model, and generate synthetic data from this model;
3. The data owner retrieves and decrypts the synthetic data using the Sharemind HI client [71].

Before proceeding with data synthesis, it was essential to comprehend the structure, features, and distributions of the initial dataset. Therefore, the initial dataset was pre-processed. The pre-processing involved rather minimalistic changes using Python code. The free text fields were removed and missing values were supplemented with not-a-number (NaN) labels. NaNs are commonly used to represent missing or undefined data in data analysis and serve as a prevalent choice in modelling, primarily functioning as a stand-in for indeterminate values [116].

Additionally, the structure of the data in the initial dataset was modified, to reduce the number of columns and increase the number of rows, which is better suited for the used synthesis tool. Separate columns for the year and month were added, and the data in a

	Force	Year	Month	Total calls	Answer time	Abandonment rate
0	South Yorkshire Police	2014	8	38077	41.6	7.100
1	West Yorkshire Police	2014	9	92439	277.0	31.740
2	Norfolk Constabulary	2014	11	24653	11.0	2.670
3	Nottinghamshire Police	2014	5	34906	19.0	2.800
4	South Yorkshire Police	2015	3	22513	49.0	0.089
...	...	...	...	...	...	...
474	Cleveland Police	2014	7	24194	10.0	2.100
475	North Wales Police	2014	5	28215	10.0	2.800
476	Gwent Police	2014	12	14335	44.0	7.000
477	Wiltshire Police	2014	12	24721	2.0	0.300
478	South Wales Police	2014	6	41721	16.0	2.000

Figure 2. Dataset with 479 rows and 6 columns

single category were collected into a common column by month, so that a separate row was created for each month for each division. For example, if the initial dataset had columns "Answer time 2023-12" and "Answer time 2023-11", and data for one division were on the same row, then during pre-processing, columns "Year", "Month", and "Answer time" were created, resulting in two rows for the same division with data respectively 2023, 12, xxx and 2023, 11, yyy). After pre-processing, the clean dataset was created as shown in the Figure 2.

The cleaned dataset in CSV format could then be uploaded into the synthesiser prototype. The ML model was trained using the cleaned dataset. Python (Jupyter notebook [117]) was required solely for pre-processing and for data analysis and comparison afterwards. The model training and synthesis were carried out in the synthesis prototype. The cleaned dataset is also referred to as original dataset from now on. The generated synthetic dataset is shown in Figure 3.

The CSV file with the results could be downloaded from the synthesiser prototype. Subsequently, the synthesised CSV and the cleaned CSV were compared in Jupyter Notebook.

	Force	Year	Month	Total calls	Answer time	Abandonment rate
0	Metropolitan Police Service	2014	4	77473	217.719735	26.216924
1	Metropolitan Police Service	2014	7	54154	183.334332	25.676414
2	Leicestershire Police	2014	10	130557	335.332544	28.104608
3	West Yorkshire Police	2014	9	100493	257.963098	27.932000
4	West Yorkshire Police	2014	10	112151	262.208179	27.228849
...	...	...	...	...	...	...
9995	Hampshire Constabulary	2014	6	37299	75.899286	5.591867
9996	City of London Police	2014	5	49983	55.669462	10.167321
9997	City of London Police	2014	6	52693	44.707526	7.455309
9998	Derbyshire Constabulary	2014	7	22001	37.357100	8.687639
9999	City of London Police	2014	6	47998	144.737721	4.555152

Figure 3. Synthesised dataset with 10000 rows and 6 columns

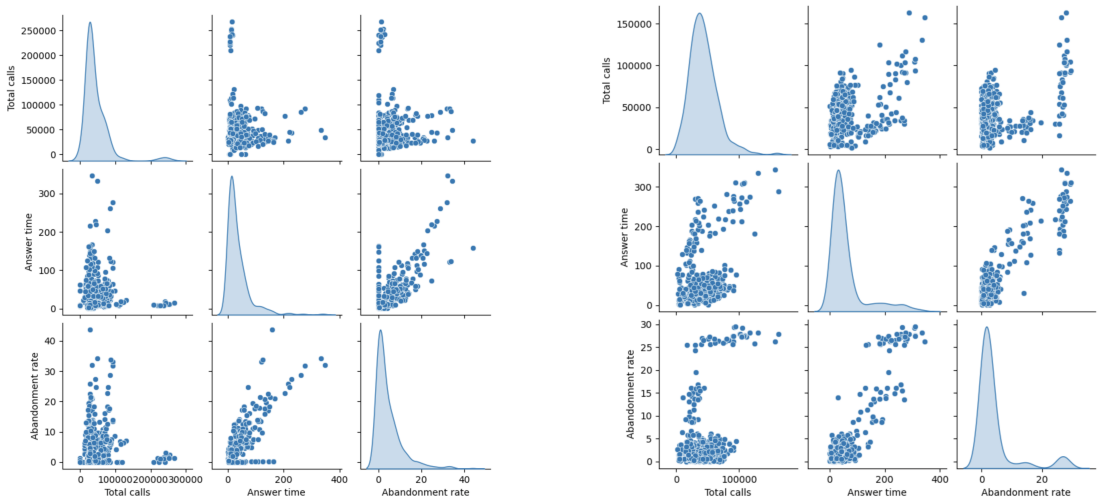
8.2 Data Analysis and Evaluation

8.2.1 Data Analysis

A comparison between the original and synthetic datasets was concluded using Python. Pairplots were used for visualisation, because they give an overview of the connection between each pair of variables in a dataset. In the analysis, only numerical values were evaluated, that is, total calls, answer time and abandonment rate were compared. Total calls represents the number of calls received that month in the given Force. Answer time refers to the average time taken to answer calls, measured in seconds. Abandonment rate refers to the percentage of calls that are abandoned by the caller before being answered.

In this use case, interpretability has been analysed through the metrics of utility and fidelity, which provide a comprehensive understanding of the synthetic data. Similar to privacy analysis, conducting a bias analysis is challenging because open data are already cleansed and anonymised. Without access to the original dataset, relevant conclusions cannot be drawn regarding the privacy and biases assessment of the synthetic dataset.

Figure 4 vividly demonstrates that the distributions of the synthesised dataset (Subfigure (a)) are similar to those of the original dataset (Subfigure(b)). Although the pairplots of the datasets look quite similar visually, they have also some visual differences. To evaluate the characteristics of synthetic data, several metrics like R-squared scores (R^2), fidelity and utility were analysed. Figure 4 shows the independent variables—total calls, answer time



(a) original (b) synthesised

Figure 4. Pairplots of the original data (Subfigure (a)) and the synthesised data (Subfigure (b)).

and abandonment rate—which were the basis for calculating the R^2 score.

R-squared scores. R^2 is a statistical measure that shows how much variance in the dependent variable can be predicted from the independent variables in a regression model, assessing the model's suitability for data processing. The score values range from 0 to 1, where an R^2 value 0 means that the model predicts 0% of the relationship between the dependent and independent variables. For the R^2 value 1, the model predicts the values in question 100% correctly. A negative R^2 score suggests that the regression model performs poorly, failing to explain any meaningful variation in the dependent variable.

R^2 scores were computed for the trained machine learning model. For each column (both original and synthetic datasets), a linear regression model was trained, and the R^2 scores for these models were calculated. The purpose was to compare how much better or worse the model trained with synthetic data performed compared to the model trained with real data.

The primary distinctions between the two datasets can be discerned through the comparison of their R^2 scores:

- Original dataset R^2 scores: [-0.00385, 0.62744, 0.72030].
- Synthetic dataset R^2 scores: [0.11343, 0.47730, 0.47262].

R^2 scores were rounded to five decimal places for better readability. In the context of

the original dataset, the R^2 scores vary, with one negative value and two positive values. Positive R^2 scores indicate a moderate to strong fit of the regression model to the data, with the model explaining a substantial proportion of the variance in the dependent variable. A negative R^2 score -0.00385 indicates that the linear regression model did not learn sufficiently to analyse the first column of the original dataset. This value is very close to zero, indicating that the model explains virtually none of the variability of the response data around its mean, essentially performing no better than a model that simply predicts the mean of the dependent variable every time. R^2 score 0.62744 suggests that approximately 62.74% of the variance in the dependent variable is predictable from the independent variables. R^2 score 0.72030 indicates that about 72.03% of the variance in the dependent variable can be predicted, which is a reasonably good fit. Metrics of the original and synthetic dataset can be seen in Table 5.

Table 5. *Metrics of the original and synthetic dataset*

Metric	**Total calls**		**Answer time**		**Abandonment rate**	
	Original	**Synthetic**	**Original**	**Synthetic**	**Original**	**Synthetic**
count	479	10000	479	10000	479	10000
mean	44144.68	50297.38	35.57	66.26	4.56	6.44
std	38248.44	34924.69	44.13	49.96	6.28	4.26
min	275	404	1	1.03	0	0.004
25%	24714	31049	9.67	28.72	0.16	3.31
50%	30784	44302.5	19	54.60	2.20	6.12
75%	54561.5	59768	44.50	92.63	6.45	8.71
max	267933	266919	347	344.26	43.80	30.28

In contrast, the models trained on the synthetic dataset exhibit R^2 scores that are consistently positive but generally lower than those of the original dataset. Although positive, these R^2 scores suggest a comparatively weaker fit of the regression model to the synthetic data, implying that the synthetic dataset may not fully replicate the underlying patterns and relationships present in the original dataset. R^2 score 0.11343 shows that only about 11.34% of the variance is explained by the model, which is quite low. R^2 scores 0.47730 and 0.47263 indicate that nearly 47.73% and 47.26% of the variance, respectively, can be explained by the model. These are moderate scores, indicating a fair prediction but significantly less effective than the best models trained on the original dataset, even though the model had significantly more data to train on.

Examining these statistics provides insights into the inherent disparities between the two datasets. Notably, the synthetic dataset exhibits larger mean values and greater variability,

as evidenced by the higher standard deviations across all features. Moreover, the range of values, as indicated by the minimum and maximum values, differs notably between the two datasets. These discrepancies in statistical characteristics underscore the main differences between the original and synthetic datasets, highlighting potential variations in underlying patterns and distributions.

When comparing the R^2 scores of models trained on original data versus synthetic data the models trained on the original data generally perform better, particularly in the best-case scenario (72.03% vs. 47.73% variance explained). The lower performance on the synthetic data could be due to several factors including (i) loss of nuanced data characteristics during the synthetic data generation process; or (ii) possible overfitting of models on the original data which do not generalise as well on slightly different or less complex synthetic data. Inherent differences in data distributions and relationships between variables exist in the synthetic data compared to the original data. Absolute differences between the original and synthetic datasets can be seen in Table 6. These differences have been explained in more detail in the following paragraphs.

Absolute difference refers to the absolute value of the difference between corresponding elements from two datasets: original and synthetic. It quantifies the discrepancy without considering the direction of the difference. Absolute difference between original and synthetic datasets suggests a comparison between the real data and the synthetic counterpart to analyse accuracy or deviation. Comparing the number of records (count) might be aimed at verifying if the synthetic dataset accurately mirrors the size of the original dataset, which is essential for validating the utility of synthetic data in simulations or modeling.

Table 6. *Absolute difference between the original and synthetic datasets*

Metric	**Total calls**	**Answer time**	**Abandonment rate**
mean	6152.70410	30.69411	1.87983
std	3323.751285	5.82508	2.01718
min	129	0.03232	0.00360
25%	6335	19.05233	3.14850
50%	13518.5	35.59864	3.92322
75%	5206.5	48.12624	2.26234
max	1014	2.73835	13.51920

Mean difference, standard deviation difference, and range difference. Mean difference refers to the absolute difference between the means of the original and synthetic

datasets. Standard deviation difference indicates that the absolute difference between the standard deviations of the original and synthetic datasets. Range difference means that the absolute difference between the ranges (max – min) of the original and synthetic datasets.

These deviations indicate differences in the central tendency and variability of the data between the original and synthetic datasets. The mean differences suggest that the synthetic dataset generally estimates higher values for answer time and abandonment rate but slightly underestimates for total calls. The standard deviation differences show a higher consistency in values for the synthetic dataset, particularly noticeable in the total calls and abandonment rate. The range differences, although relatively small for total calls and answer time, are more significant for abandonment rate, indicating a notable discrepancy in the spread between minimum and maximum values in this category. This could suggest that the synthetic dataset may not fully capture extreme cases or outliers as effectively as the original dataset in some categories, particularly abandonment rate.

Fidelity. The outcome of the analysis of the datasets in question reveals several key insights about the fidelity of the synthetic data compared to the original data. Fidelity, in this context, refers to how closely the synthetic data replicates the key characteristics and distributions of the original data.

The average absolute difference of 6152.7 in total calls between the original and synthetic data is quite high, indicating that the synthetic data may not replicate the total call volume accurately, and therefore showing low fidelity in this aspect.

The high standard deviation (3323.75) further indicates that the discrepancy of the call statistics is not consistent but varies significantly across the dataset, suggesting that the method used to generate synthetic calls needs to be improved or needs more refinement to better capture the variability in call volumes.

The mean difference of 30.694 seconds points to a moderate level of discrepancy regarding answer times. This suggests that while the synthetic data does not perfectly replicate answer times, the degree of error might be acceptable depending on the context of the analysis. If precise timing is crucial, this would be a concern. When we are analysing the mean difference regarding the abandonment rate, a relatively lower standard deviation (5.825 seconds) in answer time differences than in total calls implies that the synthetic data manages to maintain a somewhat consistent error margin in this metric, indicating moderate fidelity.

An average difference of 1.879% in abandonment rates may significantly affect analyses depending on the abandonment rate thresholds critical to the operational context. This suggests that the synthetic data may not reliably replicate abandonment behaviors. The maximum difference observed (13.519%) is particularly concerning as it indicates that in some instances, the synthetic data greatly misrepresents the abandonment rate, which could lead to incorrect conclusions or decisions based on the synthetic data.

These metrics collectively suggest that the synthetic data does not fully capture the characteristics of original data with high fidelity. There are significant differences in the metrics, which could impact the usefulness of the synthetic data for tasks that require high accuracy and replication of original data behaviors.

Utility. Utility evaluation indicates the extent to which the synthetic dataset deviates from the original one. Similarly to before, three key variables—total calls, answer time and abandonment rate—are considered.

The mean value of 6152.7 suggests a significant discrepancy in the total number of calls recorded between the original and synthetic datasets. A standard deviation of 3323.75 also suggests high variability in this discrepancy across different data points, indicating that the synthetic dataset may not consistently replicate the original dataset's call volumes.

The answer time difference about 30.7 seconds indicates that the synthetic data values typically deviate from the original data by nearly half a minute on this metric, and this could be significant depending on the context. The standard deviation is relatively smaller (5.825 seconds), which suggests that the measure is more consistent than that of total calls.

The mean difference in abandonment rate is 1.88%, which could represent a substantial variation depending on the industry standards and thresholds for acceptable abandonment rates. The maximum difference reaching up to 13.52% highlights extreme cases where the synthetic data greatly misrepresents the abandonment rates of the original data, possibly affecting the utility of the synthetic dataset for simulations or models where abandonment rate is a critical factor.

Utility results may raise concerns about the reliability of the synthetic dataset in accurately mimicking the characteristics of original data. Significant differences in variables like total calls and abandonment rates suggest that the synthetic dataset may not be suitable for applications where exact or near-exact replication of these variables is necessary.

The utility of the synthetic dataset can vary based on the specific application. For example, if the purpose is to test systems under varied conditions rather than to replicate precise behavior, then some level of discrepancy might be acceptable. If using the synthetic data for training predictive models or conducting simulations, adjustments or calibrations might be necessary to account for the observed discrepancies, ensuring the models are robust and can generalise well when applied to real world data.

8.2.2 Evaluation

The adequacy of the synthetic data should be evaluated in the context of its intended use. There are significant differences in measured variables, which could impact the usefulness of the synthetic data for tasks that require high accuracy and replication of original data behaviors. If the use case can tolerate some level of error, the current synthetic dataset might still be viable. However, for precise analytical tasks, these discrepancies might limit the utility of synthetic data.

In addition, the observed discrepancies highlight areas where data synthesis could be improved. By refining the data synthesis algorithms or perhaps incorporating more complex modeling techniques, the fidelity of the synthetic dataset might be enhanced. These insights emphasise the need for thorough validation and possibly recalibration of the synthetic data generation process to enhance its accuracy and applicability.

The metrics show that the synthetic data does not convey the desired relationships very well. At the same time, the goal was also not to overtrain the ML model. Additionally, the synthesiser prototype uses Sharemind HI service that performs data synthesis side-channel securely, which may also affect the model's accuracy. As the main focus of the used synthesis tool was to be easy to use and generally applicable to a wide variety of use-cases, the available model parameters are limited. This also limits the potential utility of the generated synthetic data as the model can not be fine-tuned to the use-case and dataset.

It is highly likely that the usefulness of the synthetic data is affected by the fact that the synthesis was performed on open data, i.e., data that had already been cleaned, which itself lacked the necessary relationships. In addition to the aforementioned, the outcome is also influenced by the fact that open data has already been anonymised in such a way that outliers are removed from the dataset, and further training smooths this distribution even more, eliminating the next set of outliers. However, the most significant factor affecting the analysis results was that the initial data was insufficient for training the model. Less than 500 rows constitute a very small dataset, and in order to generate better synthetic data, at least ten times more data would be required.

If an organisation lacks knowledge, experience, or skills in data synthesis, it is possible to use existing tools or purchase the synthesis process as a service. The method primarily depends on the purpose of data processing, the desired outcomes, and organisational or regulatory requirements. One influencing factor is also the scarcity of training data. It would certainly be helpful to have more training data and the opportunity to experiment with various models and parameters in order to find the most suitable model for the given use case.

9. Framework for Implementing Data Synthesis

To implement synthetic data in an organisation, a structured and methodical approach is necessary to maximise effectiveness and ensure compliance with legal standards. Therefore, a detailed six-step framework for integrating synthetic data is proposed (see Figure 5). In addition to the four primary steps for generating synthetic data proposed in Section 4.5, two additional steps are presented: the preparation phase and the production phase, which create a comprehensive framework. This framework ensures that synthetic data is implemented thoughtfully and responsibly, maximising benefits while minimising risks across the organisation.

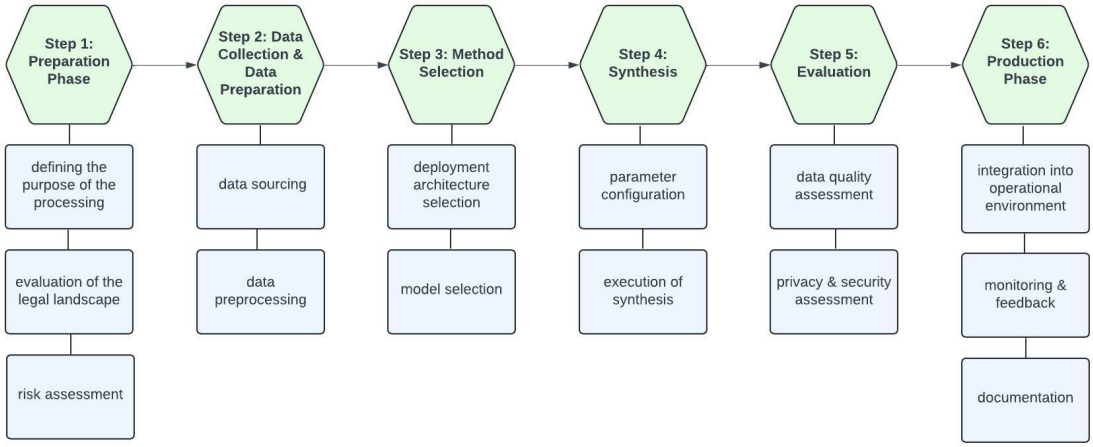


Figure 5. Framework for implementing data synthesis in an organisation

Step 1: Preparation Phase. The preparation phase involves 3 substeps: (1) defining the purpose of the processing, (2) evaluation of the legal landscape, and (3) risk assessment.

It is essential to identify specific use cases where synthetic data is required, such as enhancing data privacy, improving data security, expanding datasets for ML, or simulating testing scenarios. Clear objectives should be established to clarify how synthetic data can be leveraged to fulfill these requirements. This ensures that the production of synthetic data aligns with broader organisational goals.

A thorough understanding of the regulatory environment governing data handling practices in applicable jurisdiction must be achieved. This includes regulations related especially to data processing, data governance, and AI systems (see Chapter 5). It is crucial to ensure that all necessary consents, licenses, or agreements are in place for data processing.

A detailed risk assessment should be conducted to identify and evaluate the specific risks associated with using synthetic data. Both technical risks, such as data breaches or system failures, and non-technical risks, such as reputational damage or operational disruptions, must be considered. This approach enables the organisation to anticipate potential challenges and implement mitigation strategies effectively. See more in Section 3.4.

Step 2: Data Collection and Data Preparation. Step 2 is divided into two substeps: (1) data sourcing and (2) data preprocessing, including cleaning and analysis.

The initial task is to identify and collect the data based on which synthetic data will be generated. It is imperative to ensure that these datasets are obtained legally and ethically, securing proper consent or agreement where required. Thereafter it is necessary to evaluate data sources for their quality and relevance to the intended synthetic data applications. This evaluation helps in choosing datasets that are not only rich in information but also pertinent to the specific needs of the synthetic data use cases.

The collected data must be cleansed to remove inaccuracies, duplicates, and irrelevant entries. This cleaning process improves the overall quality of the synthetic data by eliminating noise and errors that could distort the generation process. Preprocessing techniques might be needed to standardise and normalise the data. This substep is crucial in preparing the data for effective synthetic generation, ensuring consistency across different data points and enhancing the analytical utility of synthetic datasets. See more in Subsection 4.5.1.

Step 3: Method Selection. The third step includes two substeps: (1) selection of the deployment architecture and (2) model selection, training and validation.

Deployment options include using on-premises or cloud-based solutions, or a combination of both to create a hybrid solution. The choice is closely tied to Step 1, involving an assessment of the framework, risks, and objectives regarding the use of syntetic data and concrete use case. See more in Section 4.4.

There is a need to select AIML method that is most suited in the specified use case. The model must be chosen based on its ability to handle the specific requirements of the data's complexity and sensitivity. Models and algorithms chosen must be well-suited for synthetic data generation. Options include GANs, models incorporating DP, or simpler data masking techniques. The choice of model should reflect the desired balance between data fidelity

and privacy. It is also essential to balance the complexity of the model with computational efficiency, considering the available technological resources and the scope of the use case. This balance ensures that the model can be operated sustainably within the organisation's infrastructure. See more in the sections 4.2 and 4.3.

When the AIML method is selected, the next substep is to train the model on the preprocessed data. It is crucial to ensure that the model captures the essential statistical properties of the original data while effectively mitigating the risk of re-identification. Training the model is crucial to adapt accurately to the nuances of the input data without retaining or reconstructing any personally identifiable information. This training phase is pivotal for generating high-quality synthetic data that maintains the utility of real datasets without compromising privacy. After training, validation of the model using separate validation datasets is necessary to assess the model's effectiveness and accuracy before proceeding with full-scale synthetic data generation. Validation helps identify any potential overfitting or underfitting and ensures the model's reliability in generating useful synthetic data. See more in Subsection 4.5.2.

Step 4: Synthesis. The fourth step consists two substeps: (1) parameter configuration and (2) execution of synthetic data generation.

There is a need to configure the parameters that will dictate the specifics of the synthetic data generation process. Parameters should be set to control aspects like the volume of data generated, the level of similarity to the original data, and specific privacy settings. Adjusting these parameters ensures the synthetic data meets the required privacy standards while still being useful for the intended applications. This involves a delicate balance between data utility and privacy protection. See more in Subsection 4.5.3.

With the model trained and parameters set, the synthetic data generation process can be initiated. This step must be closely monitored to ensure that the data generated meets the predefined criteria (defined in Step 1 or 2) and maintains high quality and relevance. Addressing anomalies or deviations from expected outputs and making adjustments to the model or parameters is necessary to align with the goals of the use case.

Step 5: Evaluation. The fifth step is divided into two substeps: (1) data quality assessment and (2) privacy and security assessment.

After generation, there is a need to rigorously evaluate the synthetic data to ensure it accu-

rately replicates the essential statistical features of the original data without compromising privacy. Validation of fidelity, utility and privacy of the synthetic data ensures it is fit for purpose and complies with regulatory requirements. Assessing how well the synthetic data integrates with existing systems and processes is also necessary. Usability testing should include scenarios where the synthetic data is used in place of real data to see if it performs adequately in real-world applications. See more in Section 3.2 and Subsection 4.5.4.

Step 6: Production Phase. The last step is incrporating syntetic data into real world settings and includes three substeps: (1) integration into operational environment, (2) monitoring and feedback, and (3) documentation.

There is a need to integrate the synthetic data into operational environments, replacing or supplementing real data where appropriate. This includes ensuring that all systems, processes, and stakeholders are prepared for and capable of utilising the synthetic data effectively.

Depending on the use case, there might be a need to establish a continuous monitoring system to evaluate the ongoing effectiveness and safety of the synthetic data use. This system should include mechanisms for collecting and analysing feedback from all relevant parties. It is possible to adapt the synthetic data generation methods based on feedback and evolving needs. Records of the process are necessary to improve the synthesis or helpful for ensuring compliance.

10. Recommendations for the Use of Synthetic Data in the Public Sector

Regulatory Framework Development. The current absence of national regulation concerning the use of synthetic data presents an ongoing challenge that necessitates a case-by-case interpretation of existing norms, adding to administrative burdens. To address this, it is recommended that synthetic data usage be specifically regulated within existing legislation such as the Personal Data Protection Act or the Public Information Act. Implementing such regulation would not only alleviate these burdens but also support the broader deployment of synthetic data across various public sector operations in Estonia. Furthermore, regulating synthetic data would significantly propel the development of a digital twin for the e-state, enhancing data analytics and supporting more data-driven decision-making across Estonia.

Comprehensive Analysis of Synthetic Data. A thorough analysis of synthetic data is recommended to better understand its implications and benefits. This analysis should compare real data with its synthetic counterparts across multiple sectors to scientifically assess the potential of synthetic data in specific fields. Such comprehensive research would provide valuable insights that could help regulators make informed decisions regarding legislative changes, thereby facilitating the wider use of synthetic data in the public sector.

Centralised Data Processing by Statistics Estonia. Given the complexities and the developmental stage of tools for generating synthetic data, a solution could be implemented where access to the data is provided through a secure research workstation at Statistics Estonia [112]. The secure research workstation could be adapted for use with various public sector organisations' data, ensuring compliance with data protection regulations, including the deletion of data after research activities conclude. This would assure organisations that data processing adheres to strict data protection standards.

Wider Adoption of Synthetic Data inside the PPA. It is essential to initiate comprehensive training and workshops that enhance employees' understanding and capabilities regarding synthetic data and data analytics. Implementing pilot projects can serve as a practical test bed for synthetic data applications, providing valuable insights and building

confidence within the PPA. Additionally, establishing partnerships with entities experienced in synthetic data use can offer further expertise and learning opportunities. Lastly, it is crucial to develop metrics for evaluating the impact of the use of synthetic data on organisational performance. These recommendations can be applicable to other organisations as well.

11. Conclusion

This thesis investigated the utilisation of synthetic data within the Police and Border Guard Board (PPA), focusing on its potential to streamline processes and bolster the protection of sensitive information. The findings suggest that synthetic data can significantly simplify internal procedures and support the development efforts of law enforcement authorities (LEAs). It enables the sharing of data with external developers for system development or testing without risking exposure of sensitive information, and is favorably viewed for its role in system testing and data distribution to external partners. Synthetic data also allows LEAs to conduct training and simulations for different situations and scenarios without using personal data.

However, the adoption of synthetic data is not without challenges. Primary hurdles include navigating legal constraints, ensuring data protection, and the absence of scientifically validated tools for data synthesis. The possibility of misusing synthetic data, particularly in its publication as open data, raises concerns, necessitating that such data be fully anonymised to avoid privacy breaches. However, there are reservations regarding the ability of synthetic data to accurately represent crucial data relationships, which poses a challenge when precise metrics and relationships are essential.

Various methods for generating synthetic data exist, from on-premise to cloud-based solutions, with the choice largely depending on the goals of data generation and associated legal restrictions. However, the legal framework governing synthetic data is still underdeveloped, creating uncertainties, especially within the public sector. This scenario underscores the necessity for continuous legal review and adaptation to keep pace with technological advancements and changing data usage practices.

Within the PPA, synthetic data has been used in a limited capacity. To facilitate the wider adoption of synthetic data within the PPA it is necessary to enhance understanding of both synthetic data and data analytics in general. The development of a universal tool that could generate test data simulating real scenarios is suggested, which would reduce the manual effort needed for creating test datasets and streamline the testing process.

Researchers and journalists alike are interested in analysing data of PPA. Open data provides a degree of relief—once the dataset is published, data requesters can be directed to the open data. However, the attitude towards open data within the PPA is cautious, with

stringent controls in place to ensure data protection. The process of making data open is complex and challenging, requiring a fine balance between transparency and security.

Data analysis indicated that open data is not the best option for creating synthetic data, particularly when the goal is to convey inter-data relationships. Additionally, the anonymised dataset used was too small for successfully training the synthesiser model. There were also complexities stemming from the synthesis prototype. While data synthesis provides insights into the process of generating synthetic data, researching synthetic data requires a sufficient amount of original data from which future recommendations for the potential of synthetic data can be derived.

There is a recognised gap in the legal framework that does not fully address the specific uses of synthetic data, highlighting the need for further research and tool development for effective generation and management of synthetic data. Experts have pointed out the importance of a certified synthesiser to ensure legal compliance and reliability in applications where precise data content is critical. The potential for standardisation in the synthesis process could greatly enhance its effectiveness.

Despite the potential of synthetic data, the practical application of synthetic data within sectors like law enforcement and public administration is limited by an array of legal and technical hurdles. These challenges need to be addressed through comprehensive research, development of effective tools, and necessary legal adjustments to encourage wider adoption and ensure successful implementation in sensitive sectors. Regular revisions of laws and testing protocols are essential to align with technological advances and societal norms, with calls for robust legal assurances and streamlined testing processes reflecting a concerted effort to tackle these multifaceted issues effectively. Interviews revealed readiness among several stakeholders to participate in research activities that would deepen understanding of the potential and legal characteristics of synthetic data, possibly forming the basis for legislative changes to facilitate its broader and easier adoption as a privacy-enhancing technology.

The hypothesis of this thesis was partially confirmed. Both interviews and scientific literature have recognised the potential of synthetic data in the development and testing of information systems. Interviews with experts from the PPA revealed that synthetic data also holds potential for sharing with external development partners. All experts identified specific workflows where synthetic data could enhance the efficiency of their processes. For example, synthetic data is a valuable tool for law enforcement research and analysis, allowing for the evaluation of various scenarios without the risk of exposing real data. It is also useful for testing and developing software or ML models, ensuring the system's

reliability, efficiency, and security. Additionally, synthetic data can be used for training and simulations, enabling to prepare for different situations and scenarios without relying on personal data.

However, experts were somewhat skeptical regarding the use of synthetic data in more complex research, which requires further scientific analysis to provide accurate assessments. While recognising the potential of synthetic data, the interviewees also noted significant obstacles including legal barriers, the current maturity level of synthetic data which necessitates further research to fully assess its potential, and the need for qualified synthesis tools.

Estonia can enhance its capabilities in the field of data analytics, support the digital transformation of the public sector, and ensure that data processing for research and other purposes is conducted securely and efficiently by adopting the recommendations provided in Chapter 10.

References

[1] European Commission. *European strategy for data: Data Governance Act becomes applicable. Press release.* "[Accessed: 20-12-2023]". Sept. 2023. URL: https://digital-strategy.ec.europa.eu-/en/news/european-strategy-data-data-governance-act-becomes-applicable.

[2] Tommaso Fia. "Fairness in Market Instrumental Data Governance". In: *Available at SSRN 4805796* (2024).

[3] Liina Kamm et al. "Blueprints for Deploying Privacy Enhancing Technologies in E-Government". In: *IFIP International Summer School on Privacy and Identity Management*. Springer, 2023, pp. 3–19.

[4] Mihkel Solvak and Ave Lauringson. "A Case Study of the Public Sector Digital Ecosystem in Estonia". In: *Computer* 57.5 (2024), pp. 44–49.

[5] Michal Gal and Orla Lynskey. "Synthetic Data: Legal Implications of the Data-Generation Revolution". In: *SSRN Electronic Journal* (Apr. 2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4414385. URL: http://dx.doi.org/10.2139/ssrn.4414385.

[6] Trivellore E Raghunathan. "Synthetic data". In: *Annual review of statistics and its application* 8 (2021), pp. 129–140.

[7] Ministry of Economic Affairs and Communications. *Estonian Digital Agenda 2030. Tech. rep.* Accessed: Apr. 29, 2024. [Online]. 2021. URL: https://www.mkm.ee/en/e-state-and-connectivity/digital-agenda-2030.

[8] Dan Bogdanov et al. *Privaatsuskaitse tehnoloogiate Eestis rakendamise teekaart. Aruanne. Versioon 1.1. ID D-16-215*. Mar. 2023. URL: https://www.mkm.ee/digiriik-ja-uhenduvus/analuusid-ja-uuringud?view_instance=0&current_page=1.

[9] Indu Joshi et al. "Synthetic data in human analysis: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[10] Ali Reza Ghavamipour et al. "Federated Synthetic Data Generation with Stronger Security Guarantees". In: *Proceedings of the 28th ACM Symposium on Access Control Models and Technologies*. 2023, pp. 31–42.

[11] Preeti Patel. "Synthetic data". In: *Business Information Review* (2024).

[12] Yingzhou Lu et al. "Machine learning for synthetic data generation: a review". In: *arXiv preprint arXiv:2302.04062* (2023).

[13] Barbara Draghi et al. “Bayesboost: Identifying and handling bias using synthetic data generators”. In: *Third International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR. 2021, pp. 49–62.

[14] Siseministeerium. *Siseturvalisuse arengukava 2020–2030*. URL: https://www.siseministeerium.ee/stak2030.

[15] Liina Kamm et al. “Blueprints for Deploying Privacy Enhancing Technologies in E-Government”. In: *Privacy and Identity Management. Sharing in a Digital World*. Ed. by Felix Bieker et al. Cham: Springer Nature Switzerland, 2024, pp. 3–19. ISBN: 978-3-031-57978-3.

[16] James Jordon et al. “Synthetic Data–what, why and how?” In: *arXiv preprint arXiv:2205.03257* (2022).

[17] Xiao Ling et al. “Trading Off Scalability, Privacy, and Performance in Data Synthesis”. In: *IEEE Access* 12 (2024), pp. 26642–26654.

[18] Aiden Smith, Paul C Lambert, and Mark J Rutherford. “Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility”. In: *BMC Medical Research Methodology* 22.1 (2022), p. 176.

[19] Vishnupriya Buggineni, Cheng Chen, and Jaime Camelio. “Enhancing manufacturing operations with synthetic data: a systematic framework for data generation, accuracy, and utility”. In: *Frontiers in Manufacturing Technology* 4 (Feb. 2024). ISSN: 2813-0359. DOI: 10.3389/fmtec.2024.1320166. URL: http://dx.doi.org/10.3389/fmtec.2024.1320166.

[20] André Bauer et al. *Comprehensive Exploration of Synthetic Data Generation: A Survey*. 2024. arXiv: 2401.02524 [cs.LG].

[21] David Buil-Gil, Angelo Moretti, and Samuel H. Langton. “The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis”. In: *Journal of Experimental Criminology* 18.3 (Sept. 2022), pp. 515–541. ISSN: 1572-8315. DOI: 10.1007/s11292-021-09457-y. URL: https://link.springer.com/content/pdf/10.1007/s11292-021-09457-y.pdf.

[22] Jose Pina-Sánchez et al. “Exploring the impact of measurement error in police recorded crime rates through sensitivity analysis”. In: *Crime Science* 12.1 (July 2023). ISSN: 2193-7680. DOI: 10.1186/s40163-023-00192-5. URL: http://dx.doi.org/10.1186/s40163-023-00192-5.

[23] Patrik Dokoupil. “Generating synthetic data for an assembly of police lineups”. In: (2021). URL: https://dspace.cuni.cz/bitstream/handle/20.500.11956/186895/150059703.pdf?sequence=1&isAllowed=y.

[24] Ian Brunton-Smith et al. “Using synthetic crime data to understand patterns of police under-counting at the local level”. In: *CrimRxiv* (2023).

[25] Darren Edge et al. “Design of a Privacy-Preserving Data Platform for Collaboration Against Human Trafficking”. In: (May 2020). eprint: 2005.05688. URL: https://arxiv.org/pdf/2005.05688.pdf.

[26] Walter L. Perry et al. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Santa Monica, CA: RAND Corporation, 2013. DOI: 10.7249/RR233.

[27] Donald B Rubin. “Statistical disclosure limitation”. In: *Journal of official Statistics* 9.2 (1993), pp. 461–468.

[28] César Augusto Fontanillo López and Abdullah Elbi. *On the legal nature of synthetic data*. "[Accessed: 20-12-2023]". Nov. 2022. URL: https://lirias.kuleuven.be/3976326&lang=en.

[29] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O’Reilly Media, 2020.

[30] César Augusto Fontanillo López and Abdullah Elbi. *On synthetic data: a brief introduction for data protection law dummies*. Sept. 2022.

[31] Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. “Multiple imputation for statistical disclosure limitation”. In: *Journal of official statistics* 19.1 (2003), p. 1.

[32] A. Kiran and S. Saravana Kumar. “A Methodology and an Empirical Analysis to Determine the Most Suitable Synthetic Data Generator”. In: *IEEE Access* 12 (2024), pp. 12209–12228. DOI: 10.1109/ACCESS.2024.3354277.

[33] Samuel A Assefa et al. “Generating synthetic data in finance: opportunities, challenges and pitfalls”. In: *Proceedings of the First ACM International Conference on AI in Finance*. 2020, pp. 1–8.

[34] Shuang Hao et al. *Synthetic Data in AI: Challenges, Applications, and Ethical Implications*. 2024. arXiv: 2401.01629 [cs.LG].

[35] Richard J Chen et al. “Synthetic data in machine learning for medicine and healthcare”. In: *Nature Biomedical Engineering* 5.6 (2021), pp. 493–497.

[36] Chao Yan et al. “Generating Synthetic Electronic Health Record Data Using Generative Adversarial Networks: Tutorial”. In: *JMIR AI* 3 (Apr. 2024), e52615. ISSN: 2817-1705. DOI: 10.2196/52615. URL: https://doi.org/10.2196/52615.

[37] Cybernetica AS. *PET tehnoloogiate hinded*. Accessed: Apr. 27, 2024. [Online]. Mar. 2023. URL: https://www.mkm.ee/digiriik-ja-uhenduvus/analuusid-ja-uuringud?view_instance=0&current_page=1.

[38] Dan Bogdanov et al. *Privaatsuskaitse tehnoloogiate kontseptsioon. Aruanne. Versioon 1.1. ID D-16-175*. Mar. 2023. URL: https://www.mkm.ee/digiriik-ja-uhenduvus/analuusid-ja-uuringud?view_instance=0&current_page=1.

[39] Garima Agrawal, Amardeep Kaur, and Sowmya Myneni. "A Review of Generative Models in Generating Synthetic Attack Data for Cybersecurity". In: *Electronics* 13.2 (2024), p. 322.

[40] Jerome P Reiter. "Estimating risks of identification disclosure in microdata". In: *Journal of the American Statistical Association* 100.472 (2005), pp. 1103–1112.

[41] Tucker Balch et al. *Six Levels of Privacy: A Framework for Financial Synthetic Data*. 2024. arXiv: 2403.14724 [cs.CR].

[42] Paul Calcraft et al. *Accelerating Public Policy Research with Synthetic Data. A report from the Behavioural Insights Team*. Dec. 2021. URL: https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf.

[43] Lauren Arthur et al. "On the Challenges of Deploying Privacy-Preserving Synthetic Data in the Enterprise". In: *arXiv preprint arXiv:2307.04208* (2023).

[44] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. "Winning the NIST Contest: A scalable and general approach to differentially private synthetic data". In: *arXiv preprint arXiv:2108.04978* (2021).

[45] Emilio Ferrara. "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies". In: *Sci* 6.1 (2023), p. 3.

[46] Joachim Baumann et al. "Bias on demand: a modelling framework that generates synthetic data with bias". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 1002–1013.

[47] Raissa Barcellos, Flavia Bernardini, and José Viterbo. "Towards defining data interpretability in open data portals: Challenges and research opportunities". In: *Information systems* 106 (2022), p. 101961.

[48] Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. "What is interpretability?" In: *Philosophy & Technology* 34.4 (2021), pp. 833–862.

[49] Mu-Tien Kuo, Chih-Chung Hsueh, and Richard Tzong-Han Tsai. “Automated Assessment of Fidelity and Interpretability: An Evaluation Framework for Large Language Models’ Explanations (Student Abstract)”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 21. 2024, pp. 23554–23555.

[50] Ahmed Alaa et al. “How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 290–306.

[51] Jeanne M. Carey and Kelly Rossler. *The How When Why of High Fidelity Simulation*. StatPearls Publishing, Treasure Island (FL), 2023. URL: http://europepmc.org/books/NBK559313.

[52] Imme Ebert-Uphoff and Yi Deng. “Causal discovery in the geosciences—Using synthetic data to learn how to interpret results”. In: *Computers & geosciences* 99 (2017), pp. 50–60.

[53] Alan F Karr et al. “A framework for evaluating the utility of data altered to protect confidentiality”. In: *The American Statistician* 60.3 (2006), pp. 224–232.

[54] Stefanie James et al. “Synthetic data use: exploring use cases to optimise data utility”. In: *Discover Artificial Intelligence* 1.1 (2021), p. 15.

[55] Jörg Drechsler. “Using support vector machines for generating synthetic datasets”. In: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2010, Corfu, Greece, September 22-24, 2010. Proceedings*. Springer. 2010, pp. 148–161.

[56] Allen Chang et al. “Quality-Diversity Generative Sampling for Learning with Synthetic Data”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 18. 2024, pp. 19805–19812.

[57] Vivek Ramanujan et al. “On the connection between pre-training data diversity and fine-tuning robustness”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[58] Konrad Bösche et al. “Scalable generation of synthetic GPS traces with real-life data characteristics”. In: *Technology Conference on Performance Evaluation and Benchmarking*. Springer. 2012, pp. 140–155.

[59] Mehrnaz Sabet, Praveen Palanisamy, and Sakshi Mishra. “Scalable modular synthetic data generation for advancing aerial autonomy”. In: *Robotics and Autonomous Systems* 166 (2023), p. 104464.

[60] Yao Lu et al. “Exploring the Impact of Dataset Bias on Dataset Distillation”. In: *arXiv preprint arXiv:2403.16028* (2024).

[61] Thomas Rolland and Alberto Abad. “Improved Children’s Automatic Speech Recognition Combining Adapters and Synthetic Data Augmentation”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 12757–12761. DOI: 10.1109/ICASSP48485.2024.10446889.

[62] Karan Bhanot et al. “The problem of fairness in synthetic healthcare data”. In: *Entropy* 23.9 (2021), p. 1165.

[63] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. *Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias*. 2024. arXiv: 2403.07857 [cs.LG].

[64] Lamin Juwara, Alaa El-Hussuna, and Khaled El Emam. *An evaluation of synthetic data augmentation for mitigating covariate bias in health data*. 2024. DOI: https://doi.org/10.1016/j.patter.2024.100946.

[65] Marco Huber et al. “Bias and Diversity in Synthetic-Based Face Recognition”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2024, pp. 6215–6226. URL: https://openaccess.thecvf.com/content/WACV2024/html/Huber_Bias_and_Diversity_in_Synthetic-Based_Face_Recognition_WACV_2024_paper.html.

[66] Allen Chang et al. “Quality-Diversity Generative Sampling for Learning with Synthetic Data”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.18 (Mar. 2024), pp. 19805–19812. DOI: 10.1609/aaai.v38i18.29955. URL: https://ojs.aaai.org/index.php/AAAI/article/view/29955.

[67] Domna Bilika et al. “Hello me, meet the real me: Voice synthesis attacks on voice assistants”. In: *Computers & Security* 137 (2024), p. 103617.

[68] Riigi Infosüsteemi Amet. *Küberturvalisuse aastaraamat 2024*. Feb. 2024. URL: https://www.ria.ee/kuberturvalisus/kuberruumi-analuus-ja-ennetus/olukord-kuberruumis?view_instance=0&current_page=1#aastaraamatud.

[69] The European Union Agency for Cybersecurity. *ENISA Threat Landscape 2023. July 2022 to June 2023.* Oct. 2023. DOI: 10.2824/782573. URL: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023.

[70] Georgi Ganev and Emiliano De Cristofaro. “On the Inadequacy of Similarity-based Privacy Metrics: Reconstruction Attacks against" Truly Anonymous Synthetic Data””. In: *arXiv preprint arXiv:2312.05114* (2023).

[71] Karl Hannes Veskus. “Privacy-Preserving Data Synthesis Using Trusted Execution Environments”. Available at https://cyber.ee/uploads/Veskus_M_Sc_computerscience_2022_fe493c7761.pdf. Master’s Thesis. University of Tartu, 2022.

[72] Markus Endres, Asha Mannarapotta Venugopal, and Tung Son Tran. “Synthetic data generation: a comparative study”. In: *Proceedings of the 26th International Database Engineered Applications Symposium*. 2022, pp. 94–102.

[73] Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. “A multi-dimensional evaluation of synthetic data generators”. In: *IEEE Access* 10 (2022), pp. 11147–11158.

[74] Haoyue Ping, Julia Stoyanovich, and Bill Howe. “Datasynthesizer: Privacy-preserving synthetic datasets”. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 2017, pp. 1–5.

[75] Alexander Linden. *Is Synthetic Data the Future of AI?* [Accessed: 7-04-2024]. June 2022. URL: https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai.

[76] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. “Utility and privacy assessments of synthetic data for regression tasks”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 5763–5772.

[77] Dilawar Shah et al. “Enhancing the Quality and Authenticity of Synthetic Mammogram Images for Improved Breast Cancer Detection”. In: *IEEE Access* (2024).

[78] Kalmer Keerup et al. “Privacy-Preserving Analytics, Processing and Data Management”. In: *Big Data in Bioeconomy: Results from the European DataBio Project*. Ed. by Caj Södergård et al. Cham: Springer International Publishing, 2021, pp. 157–168. ISBN: 978-3-030-71069-9. DOI: 10.1007/978-3-030-71069-9_12. URL: https://doi.org/10.1007/978-3-030-71069-9_12.

[79] Chris Liu et al. “Synthetic Data Generation Without Real Data: Uncovering Insights in Malware Detection”. In: *Advances in Information and Communication*. Ed. by Kohei Arai. Cham: Springer Nature Switzerland, 2024, pp. 235–255. ISBN: 978-3-031-53963-3.

[80] Fida K Dankar and Mahmoud Ibrahim. “Fake it till you make it: Guidelines for effective synthetic data generation”. In: *Applied Sciences* 11.5 (2021), p. 2158.

[81] Joshua Snoke et al. “General and specific utility measures for synthetic data”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 181.3 (2018), pp. 663–688.

[82] Quang Nguyen et al. “Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[83] Zhuoyan Li et al. “Synthetic data generation with large language models for text classification: Potential and limitations”. In: *arXiv preprint arXiv:2310.07849* (2023).

[84] Ruixiang Tang et al. “Does synthetic data generation of llms help clinical text mining?” In: *arXiv preprint arXiv:2303.04360* (2023).

[85] Carolina Fortuna et al. “On-premise artificial intelligence as a service for small and medium size setups”. In: *Advances in Engineering and Information Science Toward Smart City and Beyond*. Springer, 2023, pp. 53–73.

[86] Zan Zhang, Guofang Nan, and Yong Tan. “Cloud services vs. on-premises software: Competition under security risk and product customization”. In: *Information Systems Research* 31.3 (2020), pp. 848–864.

[87] Jonas Gamalielsson et al. “Towards open government through open source software for web analytics: The case of Matomo”. In: *JeDEM - eJournal of eDemocracy and Open Government* 13.2 (Dec. 2021), pp. 133–153. DOI: 10.29379/jedem.v13i2.650. URL: https://jedem.org/index.php/jedem/article/view/650.

[88] Björn Lundell et al. “Data Processing and Maintenance in Different Jurisdictions When Using a SaaS Solution in a Public Sector Organisation”. In: *JeDEM - eJournal of eDemocracy and Open Government* 14.2 (Dec. 2022), pp. 214–234. DOI: 10.29379/jedem.v14i2.749. URL: https://jedem.org/index.php/jedem/article/view/749.

[89] Bernd Gastermann et al. “Secure implementation of an on-premises cloud storage service for small and medium-sized enterprises”. In: *Procedia Engineering* 100 (2015), pp. 574–583.

[90] Shubham Gujar et al. “Genethos: A synthetic data generation system with bias detection and mitigation”. In: *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*. IEEE. 2022, pp. 1–6.

[91] Tiago P Pagano et al. “Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods”. In: *Big data and cognitive computing* 7.1 (2023), p. 15.

[92] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. “The synthetic data vault”. In: *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE. 2016, pp. 399–410.

[93] Khaled El Emam et al. “Utility metrics for evaluating synthetic health data generation methods: validation study”. In: *JMIR medical informatics* 10.4 (2022), e35734.

[94] European Union. “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)”. In: *Official Journal L119* 59 (May 4, 2016), pp. 1–88.

[95] European Union. “Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA”. In: *Official Journal L 119* 59 (May 4, 2016), pp. 89–131.

[96] *Personal Data Protection Act*. URL: https://www.riigiteataja.ee/en/eli/ee/507112023002/consolide/current.

[97] *Public Information Act*. URL: https://www.riigiteataja.ee/en/eli/ee/503052023003/consolide/current.

[98] Giuseppe D’Acquisto. “Synthetic data and data protection laws”. In: *Journal of Data Protection & Privacy* 6.3 (2024), pp. 227–239.

[99] Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. “Privacy and synthetic datasets”. In: *Stan. Tech. L. Rev.* 22 (2019), p. 1.

[100] Europen Commission. “Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts”. In: (Apr. 2021). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%5C%3A52021PC0206.

[101] EU Presidency. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Analysis of the final compromise text with a view to agreement. Brussels, 26 January 2024. Interinstitutional File: 2021/0106(COD). No. Cion doc.: 8115/21*. Jan. 2024. URL: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[102] European Commission. *Proposal for a directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*. Sept. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%5C%3A52022PC0496.

[103] Simona Strmečki and Silvija Pejaković-Đipić. "Data protection, privacy and security in the context of artificial intelligence and conventional methods for law enforcement". In: *Digitalization and Green Transformation of the EU*. Faculty of Law, Josip Juraj Strossmayer University of Osijek, 2023. DOI: 10.25234/eclic/27462. URL: http://dx.doi.org/10.25234/eclic/27462.

[104] Majandus- ja Kommunikatsiooniministeerium. *MKM tutvustab avaandmete foorumil järgmise kahe aasta tegevuskava*. Accessed: May 2, 2024. [Online]. Oct. 2020. URL: https://www.mkm.ee/uudised/avaandmetest-saavad-kasu-nii-riik-kui-ettevotjad.

[105] Eesti avaandmed. *Andmestikud*. URL: https://avaandmed.eesti.ee/datasets?emsId=12&emsId=13.

[106] Elena G Popkova and Kantoro Gulzat. "Contradiction of the digital economy: public well-being vs. cyber threats". In: *Digital Economy: Complexity and Variety vs. Rationality 9*. Springer. 2020, pp. 112–124.

[107] Mario Spremić and Alen Šimunic. "Cyber security challenges in digital economy". In: *Proceedings of the World Congress on Engineering*. Vol. 1. International Association of Engineers Hong Kong, China. 2018, pp. 341–346.

[108] Kaitsepolitsei. *Aastaraamat 2023-2024*. Apr. 2024. URL: https://kapo.ee/sites/default/files/content_page_attachments/Aastaraamat%5C%202023-2024.pdf.

[109] *PLEAK. Privacy Leakage Analysis Tools*. Accessed: Apr. 18, 2024. [Online]. URL: https://pleak2.cyber.ee/home.

[110] Police and Border Guard Board. *Conducting Research at the PPA*. Accessed: May 12, 2024. [Online]. URL: https://www.politsei.ee/et/uurimistoeoede-laebiviimine-ppas.

[111] Police and Border Guard Board. *Application for Conducting Research at PPA*. Accessed: May 12, 2024. [Online]. URL: https://www.politsei.ee/files/Anal%C3%BC%C3%BCs%5C%20ja%5C%20statistika/ukktaotlus.pdf?e184d386db.

[112] Statistics Estonia. *Use of confidential data for scientific purposes*. Accessed: May 12, 2024. [Online]. URL: https://www.stat.ee/et/avasta-statistikat/kusi-statistikat/konfidentsiaalsete-andmete-kasutamine-teaduslikul-eesmargil.

[113] UK Police. *Open data*. URL: https://data.police.uk/data/open-data/.

[114] The National Archives. *Open Government License for public sector information*. URL: https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

[115] Cybernetica AS. *Sharemind HI. Product overview*. Accessed: May 7, 2024. [Online]. URL: https://sharemind.cyber.ee/sharemind-hi/.

[116] Federico Cismondi et al. "Missing data in medical databases: Impute, delete or classify?" In: *Artificial intelligence in medicine* 58.1 (2013), pp. 63–72.

[117] Jupyter. *Jupyter Notebook: The Classic Notebook Interface*. Accessed: May 8, 2024. [Online]. URL: https://jupyter.org.

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I, Paula Etti,

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Exploring the use of synthetic data in the public sector: a framework and case study based on the example of the Estonian Police and Border Guard Board", supervised by Mahtab Shahin and Dr Liina Kamm,
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

19.05.2024

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Interview Questions

A General questions:

1. What is your role within the PPA? What specific topics do you engage with on a daily basis?
2. How do you stay up to date about of new developments in the field of data analysis, and how do you apply new knowledge in your work?
3. Which data analysis tools or software do you use in your work?
4. In your opinion, what is the most significant or paramount challenge in data analysis today?

B Questions on organisational structure and internal procedures

1. What are the main internal procedures for data processing at PPA in the following categories: data sharing for research, information systems development and publication as open data?
2. What are the main safeguards to ensure the confidentiality and security of data?
3. What problems have you encountered? What are the biggest obstacles? How have you overcome them?

C Questions regarding data transmission and practices for research

1. How many datasets does PPA issue for research each year?
2. Who are the main persons to whom the datasets are issued?
3. What are the main research areas or research topics that are investigated based on data of the PPA?
4. What problems or obstacles are encountered in issuing data for research?
5. Are there procedural or regulatory obstacles?
6. What are the main cases where it is not possible to issue data for research? How many such cases are there when the release of data is refused?
7. How do you think these problems or obstacles could be overcome?

D Questions about transfer of data to external developers for the development or testing of information systems

1. How and on the basis of which principles is data shared with external developers for the development or testing of information systems?
2. What are the processes for data pre-processing, for example, cleaning and protection (e.g., pseudonymisation or anonymisation) before sharing it with developers?
3. How is the security of law enforcement agency data ensured when providing data to developers?
4. What measures are in place to prevent or control data reuse or secondary processing after sharing data with developers?
5. What are the obstacles to sharing data with external developers?
6. How do you think these problems or obstacles could be overcome?

E Questions regarding open data

1. How is data selected and prepared for publication as open data?
2. What are the main rules and internal procedures for disclosing law enforcement data?
3. How is the security and privacy of PPA's data ensured when it is disclosed?
4. How is the use and feedback of published data evaluated? Has the analysis of open data made it possible to make PPA's processes more efficient or has it created new knowledge?
5. Have there been incidents with open data?

F Questions regarding the use of synthetic data

1. Does PPA use or has it ever used synthetic data in its processes?
2. What do you see as the main benefits and challenges of using synthetic data in law enforcement?
3. In your opinion, how does the legal environment affect or prevent the use of synthetic data in PPA?
4. What are the future plans or perspectives for the use of synthetic data in PPA?

G Additional Questions

1. Are there any other significant topics that have not been covered above?

Appendix 3 – Privacy Notice

PRIVACY NOTICE

Paula Etti is conducting interviews as part of her master's thesis titled "Exploring the use of synthetic data in the public sector: a framework and case study based on the example of the Estonian Police and Border Guard Board" (hereinafter "thesis").

By participating in the interview, you agree to the terms of processing personal data as stated in this privacy notice. Please confirm your consent to participate in the interview by digitally signing this privacy notice.

(1) Processed Personal Data

During the research activity, I will process your personal data (first and last name, contact information, job position) in accordance with applicable data protection norms.

I will conduct a semi-structured interview. During the interview, I will take notes and provide you with a summary of the interview before using it in my thesis for your review. You may make corrections or additions as needed.

I will use the results of the interviews in pseudonymised form. The interview questions will be published as an appendix to the thesis.

(2) Legal Basis for Processing Personal Data

The processing of personal data is carried out under GDPR Article 6(1)(a) (consent of the data subject). I will only use personal data for the purpose of writing my thesis. Participation in the interview and any data disclosure are voluntary.

(3) Retention of Personal Data

I will retain summaries of the interviews until the defense of my thesis, and delete them no later than June 20, 2024. Signed privacy notices will be deleted by the end of 2024.

(4) Data Subject Rights

Consent for the processing of personal data may be withdrawn at any time by informing me via email at [email address].

Contacts

For any questions or issues regarding data protection, you can contact Paula Etti as the data controller responsible for your personal data via email at [email address], or you can contact the Data Protection Inspectorate by phone at 627 4135 or via email at info@aki.ee.

Appendix 4 – Survey Questions

Survey questions for the synthetic data experts

1. What do you see as the potential of synthetic data in various industries, particularly in addressing data privacy concerns and facilitating data-driven decision-making?
2. Could you elaborate on the complexities involved in generating synthetic data that accurately reflects real-world scenarios? What are some of the main challenges you have encountered in this process?
3. From your perspective, how do different technologies and methodologies contribute to the creation of high-quality synthetic data sets?
4. In what ways can synthetic data be utilised to enhance machine learning models and algorithms? Are there any specific techniques or approaches you find particularly effective?
5. How do you envision the future of synthetic data generation and usage? Are there any emerging trends or advancements that you believe will significantly impact its adoption and application?
6. Can you discuss the ethical considerations associated with the use of synthetic data, especially concerning biases and fairness? What measures can be taken to address these concerns?
7. From your experience, what are some common misconceptions or myths surrounding synthetic data, and how would you address them?
8. Are there any specific industries or domains where synthetic data has shown particularly promising results or applications? Could you provide some examples or case studies?
9. What advice would you give to organisations looking to incorporate synthetic data into their data strategies or development processes?
10. Lastly, what areas of research or development do you believe warrant further exploration in the field of synthetic data?

Survey questions for the data governance and data protection experts

1. What is your assessment as an expert in the field of data protection regarding synthetic data?
2. Where or in what processes do you see the greatest potential for synthetic data?
3. What are the limitations of synthetic data?

4. What are the biggest obstacles from the perspective of public sector organisations in implementing synthetic data?
5. Does your organisation have any experience with synthetic data? If yes, what kind?
6. Would you like to add anything else in addition to the above?