Master Thesis U.S.E.

MSc International Management

SYNTHETIC DATA

FROM DATA DROUGHT TO SYNTHETIC PLENITUDE:

Economic Implications of Internal Data Sources in Times of Scarcity

Abstract: This thesis explores how international firms utilize synthetic data, and examines its advantages and challenges. The research questions focus on exploring how this technology can solve limitations to data access from a legal and strategic perspective, as well as the impact of internally generated knowledge on managerial choices. The study follows a thematic analysis approach based on semi-structured interviews, following the dynamic capabilities theories and the resource-based view to comprehensively analyze the practical impact of synthetic data. The results suggest that synthetic data offers significant strategic benefits, increasing the flexibility and adaptability of international firms. As few firms are already using this technology, it does also represent a source of sustained competitive advantage, especially allowing to obtain essential or unique knowledge, and safely share it across borders. Additionally, the findings suggest that there are challenges of practical nature, like demanding initial research, that may pose obstacles to a successful adoption of the technology.

JEL-codes: *F23 O33*

Keywords: *Synthetic data - Data Privacy - Data analysis - Thematic analysis*

Martina Dinardo

Student number: 1572121

email: m.dinardo@students.uu.nl

Supervisor

Professor Samuele Murtinu

Second Reader

Professor Vincent Kunst

The copyright of this thesis rests with the author. The author is responsible for its contents and opinions expressed in the thesis. U.S.E. is only responsible for the academic coaching and supervision and cannot be held liable for the content

Table of Contents

Introduction

Context

"*Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway*" declared author and consultant Geoffrey Moore (Moore, 2012). Over the past few years, the importance of data analysis skyrocketed, becoming an essential base of competition for firms on a global scale. While the premise that data is unlimited may hold theoretical validity, collecting data is becoming increasingly challenging due to stricter regulations that stem from a rise in consumer awareness and privacy concerns. Moreover, the competitive advantage offered by big data and AI is compromised by the limited size and generic nature of legally compliant data sets (Nikolenko, 2021). In times of scarcity and at the dawn of AI, could the implementation of synthetic data represent a new renewable resource? The high interest in the field of computer studies has highlighted the great advantages and potential challenges of this promising technology, but what are the concrete business implications in real life setting?

Before attempting to reflect on these questions, it must be established what exactly is synthetic data. It is not real data, but it is generated from real data, through a process known as "synthesis", which yields a dataset that maintains the same statistical properties of the real data, therefore analysis results will be comparable to those obtained with the underlying real data (Emam et al., 2020). The global rise of this new type of data is linked to two main reasons (De Wilde et al., 2024): on one hand there is an increased need of data to train AI, and this availability is impacted by data scarcity and privacy issues of traditional data; on the other, the ease of access to new technologies has made it more approachable to create AI systems, even without extensive engineering knowledge. These megatrends, along with the large demand of data in every field and the high analysis value of synthetically generated data (Emam et al., 2020), have made synthetic data one of the most interesting and promising new technologies.

Limitations to Data Access

Nowadays it is widely acknowledged that data is an essential source of knowledge for firms of any kind and scale. Already in 1984 Samiee (1984) realized: "*managing international information flows in MNCs is as important as managing the company's assets or its production*", and later in 2011 the McKinsey Global Institute's report on big data

(Manyika et al., 2011) highlighted the critical role of data analysis in all aspects of economic activities, stating: "*Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation, and growth simply couldn't take place without data.*"

Since then, data has reshaped the way firms conduct their operations. Embracing the data revolution is now a necessary prerequisite to ensure competitive advantage (Wamba et al., 2015), and with the diffusion of the sensors of the internet of things, data seems to be truly everywhere. In the words of Mr. Radi, analyst at Centogene, "*every mouse click, keyboard button press, swipe or tap is used to shape business decisions. Everything is about data these days. Data is information, and information is power*" (Bathla, 2021). This abundance of data supported the success of firms, especially international ones, allowing them to adapt their products, services and strategies thanks to market-driven analysis.

As often occurs when new technologies evolve and spread, data collection and its use have been largely unregulated until fairly recently, therefore making it easy to access and exploit. However, legislators worldwide are catching up with the rapid pace of data, and with these increasing restrictions come increasing costs and burdens for firms that need to acquire external data to operate optimally. Data scarcity has proved to be a valid concern: there is already an unmet demand for data (OECD 2019) and it is destined to worsen as limitations will become more restrictive. Data access will represent a complex challenge to overcome without addressing the issues of traditional data, such as privacy concerns, intellectual property rights and ethical use (Richards et al., 2014). National and International regulations are starting to pull the reins on unrestricted data collection, and an increasing number of firms have begun to recognize the necessity of finding an alternative for real data that can comply with the emerging requirements.

The literature fully agrees on the importance of having a good access to external knowledge in the context of international businesses, since "*knowledge is the preeminent resource of firms that wish to become and/or remain globally competitive*" (Kedia et al., 2012). Acquisitions, Alliances, joint ventures, or creating subsidiaries are among the primary ways in which a firm can access international knowledge (Fernhaber et al., 2009), however privacy issues and regulations on data protection still pose glaring obstacles to the flow of information across countries (Voss, 2020). When access to external knowledge is hindered, international firms could suffer an increase to the liabilities of the foreigners (Zaheer & Mosakowski, 1997), caused by impossibility to obtain inaccessible or heavily protected

necessary data. In this context, synthetic data could represent a key tool for international firms, thanks to the fact that it allows the internal generation of knowledge and the ability to create, from sensitive real data, compliant datasets that can be safely shared, ensuring a smooth cross-border transfer of knowledge.

Moreover, Synthetic data has the potential to not only address real data's issues, but it also offers tools for innovation in the field of market analysis, complementing and enhancing the analytical power of data. Different studies have tested to what extent synthetic data compare to real data, and recent findings confirm that they do offer results equally as valid (Lucini, 2021). Thus, synthetic data represents a unique opportunity to access quality data that is individually tailored to the needs of specific firms. In light of these findings, Synthetic data has become a compelling option to address the increasingly rigorous regulations and the consequent data scarcity.

International firms seem to be the most interested in the adoption of this new technology for a variety of reasons. Smaller firms do not usually need to collect and process massive volumes of data, and domestic firms usually have less problem accessing data from their own country's consumers. Moreover, International firms on one hand have a larger pool of resources to aid the implementation of an innovative technology, compared to smaller domestic firms; on the other, they are the ones affected by limitations on cross-border exchanges of data, due to privacy regulations and the necessity to comply with both their home country and host countries regulations, the latest being a particularly gray and complex legal area. Additionally, international firms require a deep understanding of both their domestic and international customer base needs, making synthetic data high customization capabilities a key feature for their global success.

Information management is an essential aspect for all modern firms, but even more so for international firms that have a considerable volume of data to transfer between headquarters and subsidiaries worldwide, and this data needs to be protected during the exchange process. Studies have suggested that firms invest abroad to seek external knowledge and reduce R&D costs (Chung & Yeaple, 2008), and also that Advanced and Emergent MNEs rely on international R&D activities to gain knowledge (Awate et al., 2015): Advanced MNEs focus on competence exploitation and competence creation, while Emerging MNEs aim to reach the level of industry leaders. However, when such activities are not possible due to restriction to data access, synthetic data allows firms to generate valuable data internally.

This research does not only aim to participate in this literature discussion on the relevance of access to knowledge, but also to contribute practically to the managerial implications of relying on an internally generated source of knowledge, such as synthetic data.

However, no innovation comes free of potential downsides. Despite the existence of evident benefits associated with this new technology, a cautious approach is fundamental to ensure that the potential challenges, risks, limits and the proposed solutions are weighed adequately, such as bias amplification and low interpretability (Giuffrè & Shung, 2023). Taking all factors into consideration, it appears clear that there is a plethora of information about the potential of this technology and its risks.

Research Question

Currently, the literature on synthetic data has its roots primarily in computer science, since identifying and testing the full scope of this technology is still an ongoing effort. Nevertheless, proof of exponential interests in synthetic data adoption for more than research purposes is already evident. Global synthetic data generation market was valued at USD 194.5 millions in 2022 and the forecast expects that it will reach USD 3.4 billions by 2031 (Straits Research report, 2023).

Many firms in different industries are starting to follow the path of tech giants in the use of synthetic data to deal with limitations to data access. Therefore, a systematic understanding of the use of this data and the role of internally generated knowledge could provide valuable insight into its practical impact for international firms,

Consequently, the core research question will focus on exploring beyond the applications that the literature suggests, understanding **how international firms use synthetic data to address data scarcity from a legal and strategic perspective, and how internally generated knowledge can impact the managerial choices of international firms.**

Answering these questions will provide relevant insight on new tools for firms that rely on data and are concerned with the tightening of national and international regulations, the impact on access to knowledge, as well as the possibility of obtaining new data-based strategic resources.

Structure of the paper

Drawing from both fields' literature, this work aims to offer economic insight to build a bridge between computer science and management. This study will attempt to offer a systematic overview of: the actual use of synthetic data employed by international firms; the relevant technicalities related to their generation and adoption; the legal and strategic value of synthetic data; the mechanism to integrate them with real data; the techniques to address the main challenges and limitation; and the prospected future developments.

The following sections of this paper are structured to systematically delve into the multifaceted nature of synthetic data, its practical implementation in international firms, and its broader economic impact on knowledge access. First, the literature review will present a comprehensive overview of the present state of the technology, its main characteristics, its benefits and its challenges and limitations. Secondly, the theoretical framework will lay the necessary groundwork for the subsequent research; then, the empirical methodology will indicate the rationale behind the choice of a qualitative path with an interview-based approach. The results and interpretation section will showcase all the findings to then discuss their practical implications and how they distance themself or agree with the relevant literature.

Through this structured and thorough analysis, this work aims to provide comprehensive insight on synthetic data impact on international firms operations, decisions and knowledge access, as well as suggesting directions for more extensive and further-reaching research.

Literature review

To determine whether the literature effectively captures all the potential applications of synthetic data that real firms could make use of, it is essential to better understand what exactly is the set of benefits and challenges associated with this new technology. The following sections will highlight the current key discussions surrounding synthetic data applications and implications, with the aim of contextualizing the successive discussion on the results.

Synthetic data synthesis

The outcomes of synthetic data can be a variety of output, from text and tabular data to images and videos (De Wilde et al., 2024). Sallier (2020) outlines the objective of generating synthetic data, specifying that, given the original dataset and its outputs, there are two main intents behind the synthesis: the first it to obtain new data that ensures confidentiality of the original data; the second is making certain that the results from the synthetic analysis correspond to those obtainable from the real data, with the highest degree of proximity. However, there is not a unitary definition for synthetic data, since there are various degrees of synthesis leading to different types of data:

Dummy files: "*A nonexistent file which is treated by a computer program as if it were receiving its output data*" (McGraw-Hill Dictionary of Scientific & Technical Terms, 2003). Dummy files are usually used in testing stages and they do not retain any analytical value in terms of inferential validity, therefore there is no disclosure risk (United Nations Economic Commission for Europe, 2022).

Fully synthetic files: the values of all attributes are synthetic (Hundepool et al., 2012), all of the variables undergo the synthesis process in an effort to obtain adherence with the original data set and ensure low disclosure risk (United Nations Economic Commission for Europe, 2022).

Partially synthetic files: only a group of variables are synthesized, such as the most sensitive (Hundepool et al., 2012).

Hybrid data: presents both original and synthetic data (Hundepool et al., 2012).

The core distinction between these types of data resides in the level of correspondence with the original data set. The choice of the right data needs to take into account the trade-off between anonymization and statistical relevance (Raghunathan, (2021).

This classification will be useful to better follow the firms' choice on the types of data they are adopting.

Benefits of collecting and sharing more data

Thanks to their inherent characteristics, synthetic data can help to safely collect large amounts of data, however it is important to understand why having more data available could be beneficial in the first place. Efficient data analysis could lead to a wide range of

improvements in research for collective welfare. Ilieva & Mcphearson (2018) propose advancements in urban sustainability thanks to the employment of social media data, which offer unique insight of real time urban dynamics. Moreover, if users' privacy is unequivocally guaranteed, they would be encouraged to donate their data, which would be a key asset for investigations that are complex without updated information, such as the illegal wildlife trade (Di Minin et al., 2018; Di Minin et al., 2019). Social media data could also be used to cross-validate data collected through traditional channels like surveys (Di Minin et al., 2015). Additionally, more data would contribute to the noble pursuit of cultural conservation (Toivonen et al., 2019).

The spread of synthetic data implementation would ensure that initiatives of this nature can evolve without concerns for the source of the information. This would also foster the safe sharing of data; in the words of Fernando Lucini, global lead data science & ML engineer of Accenture: "*data assumes even greater power and importance when it's shared*" (Lucini, 2021). Lucini highlights examples of achievements made possible by data sharing, thanks to synthesis processes, such as the collaboration between the american National Institutes of Health and the synthetic data engine of Syntegra, for the release of a synthetic version of an entire dataset of 2.7 million COVID patients, for global research purposes (Mendelevitch et al., 2021; Syntegra, n.d.).

Furthermore, synthetic data would not only increase the amount of data collectable, but could also enhance the quality of analysis. For instance, Lucini also reports advancements in the American justice system, like in the case of MostlyAI's research. The AI company, that focus on synthetic data, claims that they solved the limitations of the infamous COMPASS, the software used in courts to predict recidivism: Dressel & Farid (2018) proved that the software was not more accurate than people with no criminal expertise, while MostlyAI sustains that in their new and adjusted data the gap between the African Americans and Causasians that received high risk scores is reduced from the original 24% to almost nothing (1%), with "*minimal compromises to predictive accuracy*" (Tiwald et al., 2020).

It is clear that synthetic data is a promising technology that warrants deeper research to assess its true potential, but the literature does not specify whether these advantages related to safely collecting and sharing more data are feasible in real international business settings.

Compliance to privacy regulations through anonymization

The main issues linked to real data could be summarized in the concept of scarcity, and the primary obstacles are data protection regulations. Firms who wish to access, collect and use external data from consumers, now need to comply with specific legal requirements, and international businesses need to juggle between domestic and foreign countries' regulations.

The major contribution of adopting synthetic data is represented by its inherent anonymization, which allows it to bypass the legal complications linked to real data. This presents an opportunity to shift the focus from adapting traditional data in order to meet new regulations and to fit the firm's needs, towards investing in an entirely new category of personalized data. The concept of simulated datasets that could solve privacy concerns was first explored in 1993 (Rubin, 1993). Nowadays, thanks to modern advancements in the anonymization embedded in the synthesis process, many privacy issues can be effectively contained (Ruiz et al., 2018). This can be especially relevant for international firms since regulations are increasingly limiting cross-country flowing of data, which is even more problematic for the emergence of digital MNEs (Coche et al., 2024) that heavily rely on data.

The two most important regulations are currently the CCPA (California Consumer Privacy Act, 2018) and the GDPR (General Data Protection Regulation, 2018). As soon as GDPR was enforced, the number of companies that provide online services used by websites (“technology vendors”) decreased by 15% (Garrett et al., 2023). This was observed particularly in the case of European visitors. Several studies have then confirmed that the changes imposed by these regulations were not limited to the country of origin of the rules (Garrett et al., 2023), suggesting that they have a global effect, therefore impacting international firms particularly.

Empirical evidence shows that these firms face the biggest issues linked to data use across countries when they present different data regulations (Chen et al., 2021). Furthermore, in countries that differ in privacy preferences, unilaterally imposed regulation and lack of international cooperation generate negative externalities for data usage across said countries (Chen et al., 2021). To add to this already intricate scenario, stricter regulations will soon be enforced, like the EU AI Act (European Union, 2023), the first regulation on artificial intelligence that will discipline, among many other aspects, the theme of authorized data used for AI training.

International firms face the challenge of navigating this complex global legal landscape to find a reliable source of data that complies with the privacy protection of customers. However, with the current state of international requirements, the “data oasis” they are seeking may reveal to be a mirage, thus making synthetic data a more attractive solution.

The issue does not only limit the market research of firms, but it is more widespread than it may seem. Even national statistical institutions, whose activities are to analyze and release data for public use, find themselves constrained by the challenge of privacy. Skinner (2009) reports that the National Statistics Code of Practice in the UK asserts the presence of a “confidentiality guarantee” for surveys. This ensures that no statistics that can be traced back to real individuals will be released, thanks to statistical disclosure control (SDC) or statistical disclosure limitation (SDL) methods. However, with the increase in privacy concerns, it has been shown that getting access to real dataset, which has always been challenging for researchers, will become even harder (Drechsler, 2011).

While some studies approached with skepticism the extent of the effectiveness of the synthetic anonymization process (Stadler et al., 2020), modern methods seem to have already convinced relevant authorities. The 2022 G7 Data Protection and Privacy Authorities communiqué states that technologies like synthetic data: “*can facilitate safe, lawful and economically valuable data sharing that may otherwise not be possible*" (G7, Germany 2022).

Among legal experts, synthetic data has been recognized as a promising solution to legal issues surrounding real data, like compliance to GDPR (Kurapati et al., 2023) and cross-border data flows (Voss, 2020). However, it is still needed to verify whether these benefits are reported by real international organizations.

Data customization

Synthetic data still poses the unanswered economic question of how this new data is useful from a strategic outlook. Currently, many of the world’s most successful firms are employing synthetic data: Amazon uses it to train Alexa’s language system (Slifka, 2019); Google’s Waymo trains their self-driving cars with synthetic data (Bronstein et al., 2022); J.P. Morgan is using them to advance their financial services and improve fraud detection (J.P. Morgan, n.d.).

Synthetic data is clearly still primarily widespread among the largest tech companies like Nvidia, who become one of the pioneers of this technology. When discussing synthetic

data, the Sr. Product Marketing Manager at Nvidia stated "*data is the new oil in today's age of AI, but only a lucky few are sitting on a gusher. So, many are making their own fuel, one that's both inexpensive and effective*" (Andrews, 2021). This statement is proof that data scarcity is a concern even for a company that holds 80.4% of its industry market share (graphics AIB shipments JPR report, 2023).

But, while the innovation motives behind a big tech company using this new and innovative type of data are easier to grasp, an important question remains: what are the advantages for all the firms with an international customer base? Several studies demonstrated that no significant difference is found in the results obtained by data scientists who used synthetic data as opposed to real data (Patki et al., 2016). This suggests that synthetic data could serve as a substitute for global market research purposes, such as achieving a "quasi-perfect price discrimination" (Yang, 2022). Consequently, there are several recognizable benefits that the process of synthetic customization could add to the strategic decisions of a firm.

Domingo-Ferrer & Saygin, (2008) offer insight on one of the key factors that makes synthetic data an exceptional option in terms of depth and quality of insight, which is the ability to generate multiple datasets through the synthesis process, whereas there is only one real dataset. This unique characteristic allows for a variety of analyses, unmatched by real data analysis, marking it one of the earliest significant features discovered in the field of synthetic data (Rubin, 1993).

Additionally, synthetic data could provide new advancements in data processing capacity of firms. Particularly, when in real datasets some information is missing, it is common to just discard incomplete observations, potentially losing accuracy and, in the most severe cases, datasets could be unusable. However, synthetic data can artificially fill the gaps in real datasets (Poudevigne-Durance et al., 2022), allowing for levels of analysis' efficiency without precedents.

A recent study about available datasets for firms (Nikolenko, 2021), showed that the biggest issue was their size and generality. This impacts both the strategic value of these data as well as the costs associated with their analysis, because "*while capturing unlabeled data is relatively easy, the costs of manual labeling are prohibitively high. Synthetic data is an important approach to solving the data problem by either producing artificial data from scratch or using advanced data manipulation techniques to produce novel and diverse*

training examples" (Nikolenko, 2021). Furthermore, the usefulness of data is directly linked with its updatability (Constantiou & Kallinikos, 2016), which determines the time span within which said data can be considered relevant. Synthetic data could be an important step towards allowing firms to customize and tune data for specific analysis purposes. Research (Ghasemaghaei et al., 2017) shows that in order to bring true added value to businesses, data analysis requires a fit between many different elements, including analytics tools, firm's goals and data itself. This "*role of the fit*" (Ghasemaghaei et al., 2017) is often overlooked when using traditional data because of its inherent limitations. It needs to be verified whether synthetic data does provide international firms the unique possibility of ensuring unlimited access to tailored and legally compliant datasets, to match their need for highly specific data on a global scale.

Practical examples of privacy challenges in international firms: digital platforms

An effective example, representative of the type of international firms that would benefit from synthetic data, are digital platforms, which need data to perfect their services, whether it is to provide a better match of supply and demand, offer customized recommendations, or display targeted advertising based on users' activity. Particularly when it comes to marketing, studies have shown that data-based efforts can lead to an improvement in the conversion rate of online advertisements of 100% (Beales, 2010). While data is at the core of a platform's operations, the exploitation of users' data, as well as the type of data collected itself, represent one of the major arguments in favor of more austere restrictions. Tang et al. (2023) show that the power dynamics and information asymmetries that exist between platforms and users have been a concern of many authors (Liu et al., 2017; Zhou & Zou 2023; van de Waerdt, 2020). Platforms seem to have numerous obstacles when it comes to ensuring data protection of their user, and the simplistic truth behind these issues is that data protection comes at the expenses of platforms' performance: this is because "*data sanitization is a complex problem in which hiding private information trades off with utility reduction*" (Zheleva & Getoor, 2011). Compliance with regulations means self imposing limits to the data that platforms collect from their users, and restrictions to the employment of the data they do collect. Synthetic data would represent not only a legal and ethical approach to data analysis, but could actually allow to collect more data than what was even possible prior to the regulations. Therefore, synthetic data represents an interesting option for efficiency improvement in digital platforms.

Furthermore, privacy concerns are particularly prominent in digital platforms, like social media, which relies on the collection of a heterogeneity of data, from texts to images and sensors (Aggarwal, (2011).

To refine their offering and enhance the user experience, platforms use different algorithms for big data analytics (Athmaja et al., 2017); a typical example is the prediction of how users will link with each other (Wang, et al., 2007; Liben-Nowell & Kleinberg, 2003).

Platforms do have privacy protection mechanisms, however there have been instances in which these traditional methods have not been enough, and many researchers have analyzed real cases of privacy breach in platforms like Netflix (Barber & Liu 2011): to improve their recommendation algorithm, the streaming company released a dataset of 100 millions ratings from 480.000 users, whose real IDs were replaced with random ones, as common practice (Backstrom et al., 2007). However, this technique proved inadequate, as the de-anonymization attempts of researchers like Narayanan & Shmatikov (2008) proved effective in retrieving their identities.

Additionally, certain platform's business models, like Meta's, are based on providing free access to users while selling their data and attention to advertisements agencies and other companies, as a source of revenues (Evans, 2009). Since advertisers are the primary source of funding for the platform, it is evident that data is a non-negotiable asset. Graef et al. (2015) highlights that it is particularly the case with a "pay-per-click" model, where advertisers pay exclusively if a user clicks on their advertisement. To make sure that the advertisement attracts users, platforms highly depend on data collection. As highlighted by Estrada et al. (2019) a major risk for privacy is the practice of real-time bidding (RTB). During RTB, auctioneers compete to display their advertisements to users that are most likely to be interested based on their data (Yuan et al., 2013). Users' activities are sold to the highest bidder, but even those who do not win still get access to their data to decide whether or not to bid. Therefore, Some agencies may be collecting data without the intention to bid in the first place, thus obtaining free data (Estrada et al. (2019).

With such dynamics privacy protection is a strenuous challenge. Replacing real data with synthetic ones could possibly ensure that both platforms and advertisers continue to access quality data while safeguarding user's privacy.

Synthetic data risks

As it is true for every new technology, considering synthetic data an all-ecomposing perfect solution would be superficial, if not dangerous. Synthetic data can still pose security, quality and ethical concerns (De Wilde, 2024). One of the main risks is bias amplification (Giuffrè & Shung, 2023): when there are biases in the original dataset, these can be passed to the synthetic dataset and even exaggerated. While methods to reduce these threats, such as Explainable AI (XAI) (Giuffrè & Shung, 2023), are being tested to ensure synthesis transparency and valid results, it is evident that currently there are no guarantees of accuracy and, therefore, a cautious approach is necessary. Moreover, from a quality perspective, it is also necessary that the synthesis generates results that are not just a "noisy version of real-world data" (De Wilde, 2024).

Ruiz et al. (2018) present an overview of other relevant risks, starting from the claims that there are no disclosure concerns with synthetic data (Drechsler, 2011; Drechsler et al., 2007; Reiter, 2005; Reiter, 2002). While fully synthetic data may present this benefit, the same conclusion is not necessarily true when dealing with partially synthetic or hybrid data, which retains part of original data (Hundepool et al., 2012; Drechsler et al., 2007). Upon further investigation, an interesting paradox arise: Ruiz et al. (2018) propose the example of a perfect synthesizer that completely replicate the statistical information of the real data; when this dataset is released, from the perspective of the real individuals, whether the data is synthetic or not no longer matter, since anyone using it would have the same degree of information as if they had been given the real data. Moreover, one of the key features of synthetic data, the capability of creating multiple synthetic dataset from a singular original one (Domingo-Ferrer & Saygin, 2008), could increase this risk of re-identification. These threats exist in many other circumstances. An example is the presence of outliers, like a dataset with one patient with a rare disease (James et al., 2021). In such a case, maintaining the observation would make the patient easily recognizable, but removing it would reduce precision or even introduce bias (James et al., 2021).

Interestingly, these concerns would not fall under the scope of privacy regulations since, from a technical standpoint, the data is formally anonymised, however it is clearly an ethical issue. Ruiz et al., (2018) report the existence of solutions to obtain ex-ante privacy guarantees (Soria-comas & Domingo-Ferrer, 2017; Hu et al., 2014) as well as proposing a new ex-post test based on the concept of "maximum-knowledge attacker" (Domingo-Ferrer & Muralidhar, 2016), which is an attacker that has knowledge of both the original and synthetic

datasets. According to the authors, every dataset able to pass this intensive testing procedure would be able to pass any other test, hence ensuring anonymization.

Nevertheless, It is worth highlighting that the very need for testing whether a synthetic dataset is truly anonymous, symbolizes the importance of carefully studying this technology to understand both the advantages and the disguised risks that it can pose when implemented in real life firms. It could paradoxically emerge that, as synthetic data becomes more advanced, they could heighten the risks associated with privacy concerns compared to real data. Due to the formal de-jure adherence to legal requirements through synthetic data, they could lead to an allowed de-facto loss of privacy.

Synthetic data limitations: ineffectiveness to external privacy attacks

It is essential to not only consider the benefits and risks associated with emerging technologies, but also to comprehend the boundaries of their capabilities, to prevent proliferation of unrealistic expectations. Synthetic data serves as an effective means for anonymizing data held by entities such as firms, to allow them to collect and analyze data while protecting privacy. Furthermore, compared to handing out real data, they are useful in shielding customers' privacy when allowing external entities to collect information, like third party developers whose applications can receive privileges to access users' data on platforms (Ahmed & Shaikh, 2011).

However, synthetic data Achilles' heel is represented by the attempts to cause privacy breaches perpetrated by malevolent external attackers. To better visualize the problem, digital platforms are again a perfect example. Zheleva & Getoor (2011) classify the traditional anonymization strategies for social networks in four categories: edge modification (Liu & Terzi, 2008), randomization (Ying et al, 2009), network generalization (Liu & Yang, 2011), and differentially private mechanisms (Gao et al., 2018).

Synthetic data is a strong ally to anonymize data that platforms need to analyze themselves to improve performances, but it is inapplicable to protect information that users publish themselves by willingly having a public profile (Zheleva & Getoor, 2009). In these instances, it pertains to the network configuration itself to avoid breaches of the real data. To highlight the complexity of the problem, misconfigurations have been observed even in big tech companies, which should posses all necessary resources to prevent them, like the cases of Amazon Web Services (Zuo et al., 2019; Engström et al., 2023) and Microsoft Azure (Zuo et al., 2019).

Consequently, in real business settings, relying exclusively on synthetic data to address and fix the problems related to data protection does not appear a feasible approach, since there are risks associated with handling data that arise for the very presence of data in online databases.

UN Recommendations on the Use of Synthetic Data

Synthetic data is rapidly gaining traction, attracting the interest of every organization looking to integrate AI in their operations and need data for training these systems. The rise in interest towards synthetic data brought this new technology to the attention of the United Nation, leading to the recent release of thirteen recommendations (De Wilde, 2024) applicable in both the private sector and for public policies:

1) Mitigate bias
2) Use a range of generating mechanisms for synthetic data
3) Ensure transparency
4) Calculate and disclose quality metrics for synthetic data, and validate the data.
5) Synthetic data should preferably be open access and always watermarked to disclose their origin
6) Develop and maintain cybersecurity measures to protect synthetic data
7) Model validation and evaluation
8) Establish global quality standards and security measures
9) Locally enforce quality standards and security measures
10) Create ethical guidelines that take synthetic data into account
11) Balance the relationships between experts, curators, and generators of synthetic data
12) Promote global research networks on the safe and ethical use of synthetic data
13) Create policies to make sure synthetic data reduce the divide between the Global South and Global North

The depth of these recommendations is still generic due to the uncertainty surrounding the full spectrum of synthetic data applications, but they represent important guidelines to understand what are the necessary precautionary steps to ensure a safe and controlled use, in order to positively impact both firms and society. Another relevant theme of analysis of this work will be investigating whether these formal recommendations are being integrated into

the practices of international firms that are adopting synthetic data, and what is their practical value.

Theoretical Framework

Synthetic data presents itself as a new resource that could confer great benefits to the firms adopting it, however it is unclear whether it can truly lead to a sustained competitive advantage compared to firms that decide to not adopt it. Therefore, this study will use the resource-based-view (Barney, 1991) to understand if synthetic data respect the requisite of the VRIN framework, being valuable, rare, imperfectly imitable and non substitutable. The computer science literature already explicitly mentions the valuable aspect of synthetic data, which is mirroring the analytical value of real data (Lucini, 2021), while ensuring data privacy. This is particularly valuable for international firms that have to deal with cross-border data regulations. This study will try to verify whether synthetic data also presents the remaining characteristic to be considered a source of sustained competitive advantage.

Furthemore, the dynamic capabilities theory (Teece et al., 1997) suggests that adaptability and flexibility are two key measures to ensure business success, especially in the context of rapid technological changes. Synthetic data will be analyzed in light of this theory to understand how much more flexibility it gives firms. This can be especially relevant for international firms that need to be responsive enough to adapt to the ever changing global scenario, and deal with cross-border data regulations. A firm's ability to reconfigure their competences is key to surviving the international competition, and this study will use this theory to verify whether synthetic data is a relevant source of dynamic capabilities for firms that operate internationally.

Empirical methodology

As shown by the literature review, the novel field of synthetic data is continuously evolving due to the rapid and unpredictable technological advancements, and this poses a challenge for an economic interpretation of this new phenomena. As a result of the novel and still opaque implementation of synthetic data in an economic environment, a deductive method may not fully capture the complexities of this phenomena. Therefore, in order to effectively address the challenges of understanding this highly technical and dynamic subject through an economic perspective, a qualitative approach seems the best fit. This study presents an interview-based approach that allows for depth of insight and first hand

observations. Specifically, this research follows a thematic analysis method to identify and interpret themes (Cassell et al., 2005); Clark & Braun, 2017).

Due to the distinctiveness and potential competitive advantages associated with leveraging this new technology, firms that are starting to explore synthetic data may be wary of sharing information when answering questions in a standardized survey. This would have represented a challenging limitation to data access for this study and would have compromised the depth and accuracy of the results. Therefore, an interview setting, which requires direct interaction and collaboration with firms, has been the right method to build trust and willingness to participate.

To capture a broader spectrum of applications and associated challenges, the sample includes ten among international firms that have implemented synthetic data, and firms that act as providers of synthetic data, generating and selling it to other firms. The diversity in the sample serves the purpose of obtaining a diverse collection of points of view, observing the technology from more than one perspective. Additionally, the inclusion in the sample of international providers of synthetic data, facilitate the identification of the current best practices surrounding the technology, thanks to their superior expertise.

The international firms interviewed all operate in different industries or sectors, and this avoids the identification of industry-specific factors, fostering grounded generalization. The specific figures interviewed within these firms were the managers and data engineers leading the synthetic data generation and analysis. The interviewees were selected for the research based on their relevant experience with the technology and with how the synthetic results are then utilized.

Data collection primarily revolved around open-ended questions, to also avoid possible influence of the interviewer on the respondents. The questions were used as flexible guidelines to aid a comprehensive discussion in semi-structured interviews of the duration of one hour. The full questionnaire is available in the appendix.

The interviews have been transcribed verbatim and all transcripts were anonymized to ensure protection of the participants' privacy. The obtained unstructured qualitative data that have been categorized using Nvivo in order to be analyzed for the presence of patterns and themes. Through an iterative process the codes have been refined to reflect the main insight discovered and the core aspects of the findings, ensuring well-grounded theory development.

Expected findings include detailed insights into the value that synthetic data brings to international firms, defining whether it does represent a source of sustained competitive advantage following the definition of the Resource-based view (Barney, 1991), and whether it is a source of dynamic capabilities (Teece et al., 1997), ensuring firms' flexibility and adaptability. This study also aims to clarify to what extent synthetic data grants easier access to data and knowledge compared to traditional data. The focus on real life cases could contribute to the practical understanding of the overall impact of synthetic data on international firms.

Results and Discussion

This section will present and discuss the results extrapolated from the interviews conducted with the international firms included in the sample. The qualitative results have been organized in tables and summarized in graphs for ease of reading, the complete tables are available in the appendix. The interpretation of the findings will aid the discussion and help to later address the research questions.

Industries and Use Cases for Synthetic Data

The interviewees reported the use of synthetic data in a variety of different industries and sectors (Table.1).

In the **healthcare field** synthetic data is being used to create synthetic medical records (Table.1: Interviewee 4) and managing the supply chain in the pharmaceutical industry (Table.1: Interviewee 10); Synthetic data is also used for refining decision-making for **financial services**, as bank credit scores, risk management and financial modeling (Table.1: Interviewee 2); In the **defense and security** field synthetic data is being implemented for application in defense systems (Table.1: Interviewee 1; 3), surveillance, and monitoring of public safety (Table.1: Interviewee 3); The technology is being used in **retail** to improve transportation (Table.1: Interviewee 3), customer relationship management (Table.1: Interviewee 7) as well as purchasing prediction (Table.1: Interviewee 5); Synthetic data has also been adopted in **quality assurance** in the insurance field (Table.1: Interviewee 9), in fraud detection (Table.1: Interviewee 7) and money laundering detection (Table.1: Interviewee 4) in different industries; **environmental monitoring** is another field in which synthetic sensor data is being used for agriculture (Table.1: Interviewee 1) and wildlife conservation (Table.1 Interviewee 3). Synthetic data is being adopted also for **infrastructure**

developments for urban management, specifically in intelligent traffic (Table.1: Interviewee 3) and traffic models (Table.1: Interviewee 6); the **research field** is using synthetic data for testing in the aeronautical and space domain (Table.1: Interviewee 8), in the satellite industry, as well as for nuclear (Table.1: Interviewee 7) and particle accelerator testing (Table.1: Interviewee 9); Additionally, Synthetic data are used for the developments of **new technologies**, such as facial anonymization, computer vision and and ADAS (Advanced Driver Assistance Systems) (Table.1: Interviewee 3).

A diverse array of applications and fields of use is showcased by each interviewee, proving the high flexibility and adaptability of this technology. It was not possible to circumscribe the areas in which synthetic data is being used, since it appears that there are no defined boundaries to the current uses of synthetic data. This vast applicability and lack of constraints represent a key characteristic of this new technology, testifying that its full potential is still undiscovered or that, perhaps, hard limitations do not exist at all. This suggests that its applications will continue to expand in the industry in which it is already used, as well as additional ones. The results clearly indicate synthetic data versatility, which can be also a crucial driver of innovation, enabling new opportunities and technological developments. Moreover, the possibility to use synthetic data for wide applications in a variety of contexts suggests that it can represent a relevant source of flexibility for firms.

Relevant Technicalities

Defining the current applications of synthetic data gives an overview of what the technology is capable of in its current state, however this still does not clarify why it is possible in the first place to adapt synthetic data to these many applications. Therefore, it is important to identify how each firm define synthetic data (Table.1), as well as the differences in the generation process (Table.3), and the mechanism used to ensure quality and relevance of the synthetic data (Table.4)

Synthetic data definition

The results in Table.2 show that the respondent does not present unanimous agreement on a precise definition. However it is possible to identify several common features and characteristics. Synthetic data is generally defined as data generated artificially that represents realistically or mimics real data, preserving its statistical properties. It is also reported the

existence of a "*kind of a spectrum*" (Table.2: Interviewee 4) in relation to how much of the underlying real data can be encountered in the synthetic dataset. Two interviewees explicitly mentioned Artificial Intelligences as the source of the synthetic data (Table.2: Interviewee 2; 7). Additionally, two respondents did not give a definition, but rather expressed the main goal reached thanks to this technology in their specific cases, namely engineering a data set (Table.2: Interviewee 1) and testing (Table.2: Interviewee 10).

Overall, all the definitions highlight the balancing of realism with privacy and the capacity to provide valuable data for diverse purposes. While the results do not show a singular definition, there seems to be a common background of concepts that is shared among the respondents, like the objective of representation of real data. The real data, however, is not presented as the input of the generation process, but rather seems to be approached almost as an inspiration for the synthetic dataset, as shown by some of the interviewees: "*...does not directly come from real data, but we create it artificially. Of course, we base it on real data…*" (Table.2: Interviewee 5); "*...It is not measured or derives from real data directly…but statistically it represents the data*" (Table.2: interviewee 6). These descriptions of synthetic data fit well the idea of a new type of data that can protect the privacy of real data underneath, as reported by the literature (Ruiz et al., 2018). There are two main themes that emerge, on one hand the concepts of having protected real data available, on the other obtaining brand new data that holds the same statistical power of the real data. One of the participants gave a particularly incisive way of describing synthetic data, presenting it as "*being created rather than being collected*" (Table.2: reference 5). The potential of this new technology shines through this insightful definition thanks to the allusion to the vast concept of creation, which makes synthetic data appear as a promising tool to address the reported issues of using traditional real data, like privacy and ethical concerns (Richards et al., 2014). Furthermore, this definition directly ties into the idea that synthetic data does represent an internally generated source of data that can supplement tailored knowledge to firms, thus representing a source of adaptability to a variety of specific circumstances and a source of dynamic capabilities (Teece et al., 1997).

Synthetic data generation

Understanding the generation process of synthetic data is another key step towards clarifying the processes that this technology undergoes to become a valuable asset for firms.

Figure.1 showcases the techniques used to generate synthetic data and, again, there is no universal approach. The three predominant themes include: **simulation environments** (Table.3: Interviewees 1; 3; 9) which are based on 3D spaces that replicates the needed conditions and assets; **AI assisted generation** (Table.3: Interviewees 2; 7), that uses the computational power of artificial intelligence to extract the relevant information form the underlying data; **statistical modeling** (Table.3: Interviewees 4; 5; 6; 8; 10) to identity the statistical characteristics to generate *ad-hoc* synthetic data.

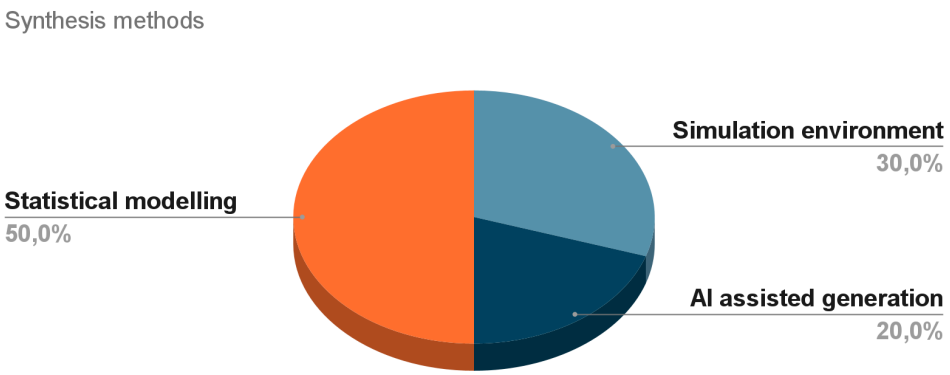

Figure.1: synthesis methods

The diverse approaches to synthetic data generation described by the interviewees reflects again the technology's flexibility. There is no defined best way to generate it, which gives firms the freedom to explore more possibilities and find the most appropriate methods for each specific use case. Additionally, while some advanced techniques are mentioned, like those based on artificial intelligence (Table.3: interviewees 2; 7), it is also reported the existence of more approachable ways of obtaining valuable synthetic data (Table.3: Interviewee 5). This indicates that even simpler techniques, that require less prior research, can be useful in providing efficient ways to start integrating synthetic data in the firms' operations and start generating internal knowledge. This evidence is promising because the learning curve to understand and introduce new technologies can sometimes be quite steep. Furthermore, the availability of different sources of a given technology is a proven way to facilitate higher entry rates in certain industries (Lieberman, 1989). This means that the wide variety of synthetic data generation techniques could have an impact on internationalization or expansion decisions, potentially becoming an important tool for firms assessing and entering new markets.

Ensuring quality and relevance of synthetic data

In light of the findings about the absence of universal definition and generation process of synthetic data, a question about how firms can ensure that synthetic data is suitable and of high-quality arises. When discussing the practices in place to ensure the utility and accuracy of the obtained synthetic dataset several different methods have been reported by the interviewees. Table.4 presents the main methods used during synthetic data generation to check the quality and relevance, while figure.2 summarizes these techniques and figure.3 shows the percentage of participants adopting one or them.

A first approach discussed relies mainly on the **integration of the final users of the data in the verification process** (Table.4: references 1-3). Specifically, the findings show that there can be a distinction between the professionals that generate the data, usually data scientists and statistics experts, and the professionals utilizing said data for different purposes. This distinction can exist for two reasons: on one hand, within a single firm, these functions can be assigned to different employees; on the other, some firms decide to outsource the generation of synthetic data to companies that specialize in synthetic data production. Therefore, the firm that purchases the data participates marginally in data generation, for example providing the necessary real data, and it only has the main function of evaluating whether the data works adequately or needs further refinement. In both these cases, the methods based on the user's feedback require an interactive process of tuning the data until the data reaches the needed quality. Another method that relies on collaboration is based on integrating into the synthetic solutions itself a **system to report when the results are satisfactory and request changes** when needed (table.4: reference 4). There is an additional method that shifts away from extensive communication and collaboration and is based on utilizing **AI systems** to help capture all the relevant statistical characteristics and distinguish good data (Table.4: references 5-7). A popular method is based on thoroughly **comparing the results obtained after analyzing the synthetic data with the results they would get after using the real data** (Table.4: references 8-12). Additionally, another method is based on simulating precisely all the details of the exact environment and assets taken into consideration, to achieve the **highest possible degree of customization** (Table.4: reference 13). This approach aims to create a database perfectly tailored to the specific use case of the firm, ensuring higher effectiveness. Lastly, the results also show that there are instances in which the quality of the data **cannot be verified beforehand** (Table.4: references 14-16),

which can happen if, for example, the size of synthetic data processed in certain multinationals is too large to be systematically checked.

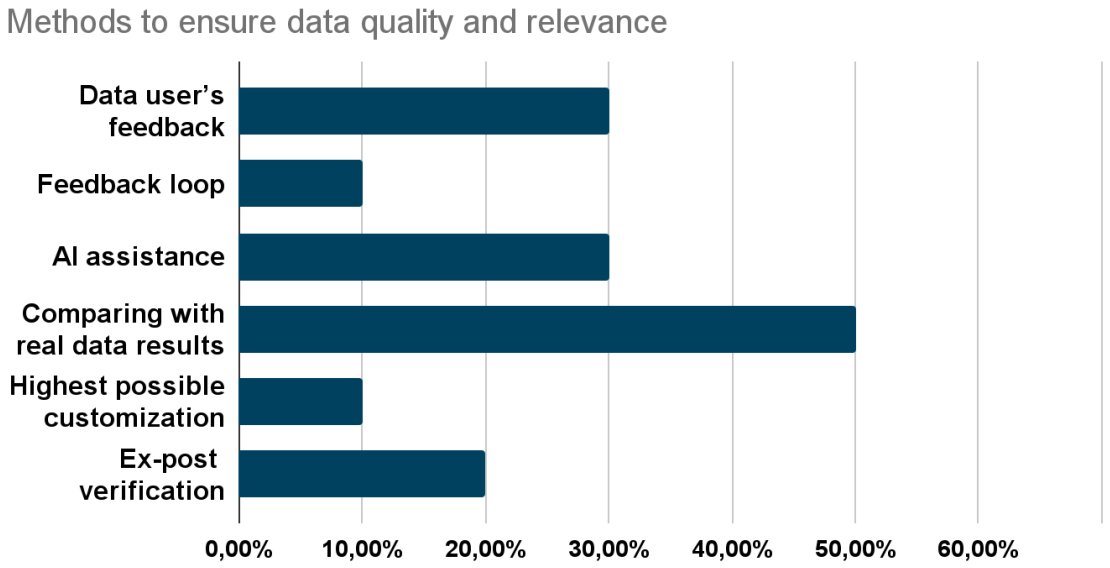

Figure.2: methods to ensure quality and relevance

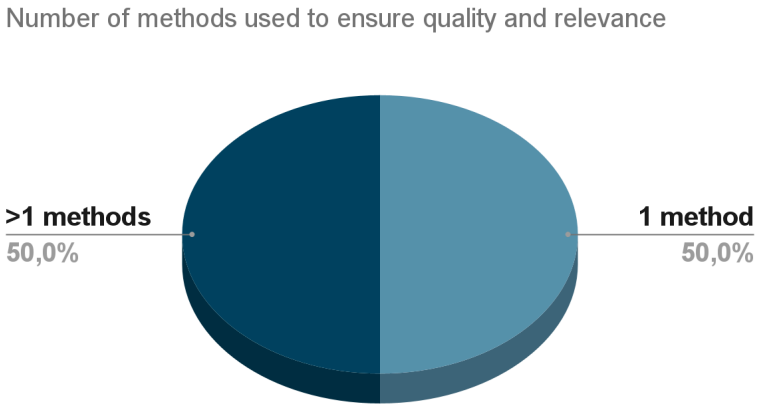

Figure.3: number of methods used by single interviewee to ensure quality and relevance

As shown in Figure.3 50% of the respondents reported using more than one method to control synthetic data validity, stressing the importance of having quality data for efficiency and performance. This commitment towards ensuring synthetic data quality for optimal performance is in line with the literature suggesting that, in general, higher quality data impact firm performance directly and indirectly (Côrte-Real et al., 2020). The existence of this variety of methods may also be explained by the fact that some of these techniques are inapplicable in the cases in which real data is completely absent. To give an example, the method based on comparing the results of the synthetic and real analysis implies the necessary

existence of real data, however, there are many instances in which real data is either unavailable or inexistent. This complication will become even more common as real data becomes more protected, increasing data scarcity. Notably, respondents reported that there are already methods designed to function when detailed real data is not sufficiently available, like the highest customization approach (Table.4: reference 13). This provides evidence in support of synthetic data representing a valid alternative to missing or sensitive real data that a firm may specifically need, therefore proving the rarity of the technology, following the VRIN framework.

Interactions between synthetic and real data

It is not possible to completely separate the concept of synthetic data from real data because, even when real data is not available, the ultimate goal of synthetic data is to replicate it. Therefore, it is essential for the comprehension of this technology's impact to understand the degree of compatibility between real and synthetic data (Table.5), and analyzing how they interact with each other (Table.6).

Compatibility with real data

Figure.4 shows that a number of interviewees, accounting for 50% of the respondents, reported the adoption of synthetic data both as a complement to real data and as a substitution. When synthetic data was used only as a complete substitution for real data (20% of respondents) different reasons have been invoked. In one case synthetic data was preferred because the results were equally valid (Table.5: Reference 6); In another case synthetic data was chosen because using real data was ill-advised due to the high sensitivity of the data (Table.5: Reference 7); the other 30% of the respondents cited the use of synthetic data exclusively as a complement, in one case this choice was dictated by the necessity of correcting an unsuitable level of imbalance present in the real data (Table.5: Reference 1).

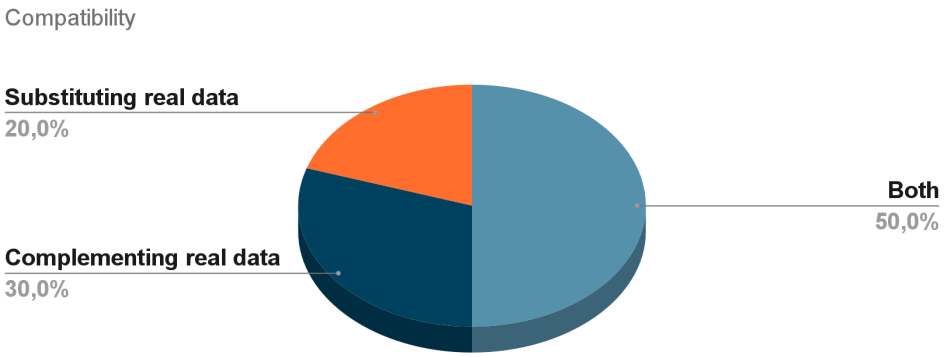

Figure.4: compatibility between synthetic and real data

Interaction issues

Figure.5 highlights that 60% of the interviewees report no major interaction issues arising from combining real and synthetic data. In these cases, the only needed foresight mentioned was related to ensuring that both types of data presented the same format (Table.6: reference 8; 9), and being aware of the possibility of biased real data (Table.6: reference 5). The respondents that did have problems integrating the two types of data reported that they were mainly due to imbalance in the real data (Table.6: reference 1) and complexity of synthetic data modeling (Table.6: reference 2; 4).

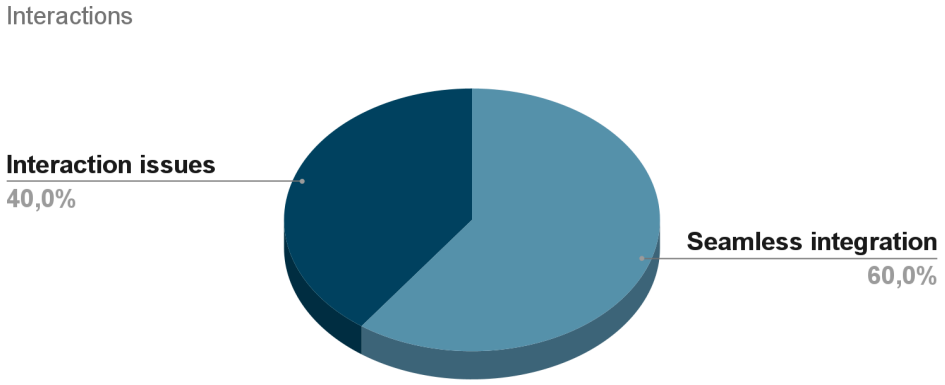

Figure.5: interactions between synthetic and real data

Both the questions around the compatibility and interactions between synthetic and real were not directed towards expressing a value judgment regarding which type of data proved to be better. However, some interviewees spontaneously highlighted that synthetic data, when used on its own, brought better performances (Table.5: reference 16; 18), while others expressed that it helped refine the overall results (Table.5: reference 9). In some cases,

even when possible to use a combination of real and synthetic data, it was expressly decided to not include real data, mainly for privacy reasons (Table.5: reference 13; 14). This shows that synthetic data features are allowing it to emerge as a standalone technology that firms can choose to adopt as an alternative to limited real data, which is in line with the literature suggesting that synthetic data results are equally comparable to real data results (Lucini, 2021). Synthetic data value was considered substantial, in some cases even to the point that *"...synthetic data can completely overshadow or even replace the real data…*" (Table.5: Reference 5). Furthermore, in all of the cases of interaction issues reported in Table.6, the mentioned problems did not represent major complications, signaling that the technology, despite its novelty, is already stable enough to be applied in a controlled manner.

Strategic Impact

Once understood what synthetic data represents for the international firms that are currently adopting it, the priority becomes to focus on what exact implications this technology has on these firms' competitive advantage and strategic decision making. The interviews have shown that the strategic impact of synthetic data can be divided in three main categories: cases of **synthetic data indispensability** (Table.7), of **enhanced power of data** (Table.8) and of **cost reduction** (Table.9).

Synthetic data indispensability

When it comes to the concept of synthetic data representing an indispensable technology, the reasons seem to be linked to two main themes as shown in figure.6. On one hand, there are cases in which synthetic data represent the only viable solution to a specific need or goal, to the point of being considered **unavoidable or irreplaceable** for certain objectives by 80% of the interviewees (Table.7. references 1-10), especially when access to real data represents a significant obstacle. For example, an interviewee noted: "*It's a tool that sometimes makes a project feasible…*" (Table.7: reference 1), Similarly, another interviewee reflected that "*it's something that can't be stopped, because if we don't adopt it…we will succumb, because our competitor from another company, from another country, whatever it is, will adopt it. So it can't be stopped*" (Table.7: reference 8), and another interviewee also stated: "*It was the only thing feasible, because the data, real data, to us, it was really impossible to get there…*" (Table.7: reference 10).

On the other hand, in certain circumstances, synthetic data can offer **unique insight** and depth of information that would not be achievable relying on real data, as reported by 50% of the respondents (Table.7: references 11-18). For example, synthetic data "*allows you to make models on risk scenarios that you do not actually have, or because you cannot use the data for questions by law or simply because you don't have them*" as noted by an interviewee (Table.8: reference 13). Another case of unique insights is also the testing of extreme conditions, as mentioned by other interviewees: "*in the satellite industry, to simulate a shock or some tests, you cannot do them at all….Well, with synthetic data, you can simulate a sensor that is telling you: oh, now I have a shock, now I'm not working anymore or now the temperature is reaching whatever, etc. And you test the behavior of your automations*" (Table.8: reference 15).

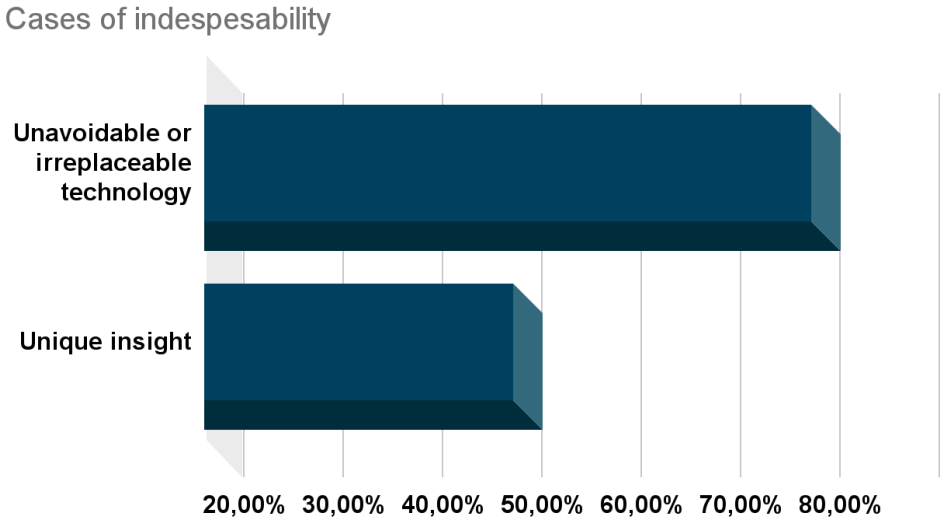

Figure.6: percentage of respondents reporting synthetic data indispensability

It is worth noticing that 80% of the respondents, as shown in figure.6, manifested cases in which synthetic data is now considered indispensable, and the interviewed firms all operate in different industries and use synthetic data for very different purposes. Therefore, the question arises about why exactly this technology seems to be so well applicable in all these diverse contexts. The answer may be the profound changes brought by the 4th industrial revolution (Schwab, 2017) that the world is living. Even though new technologies demand an effort to adapt to it, there are certain innovations that become the norm. It seems that synthetic data may have the requisite to become one of those in certain data demanding sectors; in the words of a respondent: "*It will become a necessity…when the internet was created everyone started working with the internet, it was no longer possible not to work without the internet.*

When someone invented social media, everyone started marketing on social media data…when the synthetic data is truly of proven quality and safety, it will become impossible not to rely on this type of additional information" (Table.7: reference 4).

From a managerial perspective, it is important to consider not only the potential of integrating new technologies, but also the uncertainty associated with them. Schumpeter (1994) highlighted the existence of a destructive transformation that happens with the introduction of innovations, but while synthetic data's impact still presents gray areas, it is evident that the organizations who are adopting it are quickly recognizing its value. Furthermore, the ability to simulate rare or extremely complicated tests holds great power in terms of strategic value. This means that synthetic data is a new technology that actually reduces uncertainty. This is extremely relevant for firms that internationalize since "*more uncertainties might be revealed and risk might increase with international experience, thereby inhibiting internationalization, and in some cases inducing de-internationalisation*" (Liesch, et al. 2011). At least in the first stages of internationalization, utilizing synthetic data could help reduce said uncertainty. Moreover, reducing uncertainty responds to the call for adaptive capabilities that allows organizations to be resilient in the face of disruptions and unpredictability (Adobo et al., 2021). Some interviewees gave examples of this characteristic: "*...it's something you can't really test in the real world. You know, you can't damage a main artery, for instance, just to see whether the system deals with this gracefully…*" (Table.7: reference 12); "*[you do not have to] wait for the lightning, the volcanic ash, instead of extreme phenomena that we don't get*" (Table.7: reference 16). Furthermore, we know from contingency theory that in management there is no universal strategy that fits all firms (Donaldson, 2001), but that they are dependent on the specificity of the internal and external circumstances. Therefore, the ability to make unique and highly specific simulations that would not be possible otherwise suggest that, compared to firms that are not using it, synthetic data is not only valuable, but indeed also a rare, imperfectly imitable and a non substitutable resource, following the resource-based view (Barney, 1991).

Enhanced analytical power of data

The second category linked with the strategic impact of synthetic data relies on the enhancements to the analytical power held by data thanks to synthetically engineered datasets. In this case, three major advantages of using synthetic data have been identified (Figure.7).

The first benefit is **improvements in data analysis results and performance** (Table.8: references 1-20), mentioned by 90% of interviewees. For example, an interviewee reported higher efficiency compared to real data: "*We use it to improve the performance of our models…you're also going to provide more robust, more reliable solutions because you're able to throw at that algorithm many different environments, many different instances… a lot more different scenarios than you can when you capture real data*" (Table.8: reference 3). Another interviewee noted that, using synthetic data it was possible to carry out analysis much more complex, stating that: "*it allows us to use really large models that require a lot of data. The larger your model is, the more parameters you need to train, the more data you need. And so, having synthetic data allows you to use very fancy heavyweight models*" (Table.8: reference 19).

Secondly, another advantage reported is the ability to **expand an existing database** thanks to synthetic data augmentation (Table.8: references 21-30) reported by 70% of the respondents. This practice allows to increase the volume of data for analysis purposes, also ensuring faster analysis; an interviewee reflected: "*We have populated our simulation engine with so many different assets…in all different angles…The list was endless. Add as much variety as you can to expose your neural networks, the better they will perform…We're able to generate enough imagery to be able to retrain, to refine and improve the performance of that solution. And we get it out there faster than anybody else*" (Table.8: reference 23). Data augmentation is also possible starting from meager real data: "*from the little satellite they have, the test satellite, we detect all the parameters, all statistical characteristics, and we start simulating hundreds or thousands of fake satellites, basically. And this is how they can simulate or can verify that the new software will be working with the real satellites*" (Table.8: reference 25)

Lastly, 60% of the interviewees also reported the possibility of making **better data predictions** (table.8: references 31-37). Specifically, some respondents reported the opportunity to use synthetic data to be able to foresee future data patterns. For example, regarding improvements in sales prediction, an interviewee stated: "*When predicting future customer behavior, well, before using synthetic data the predictions were pretty bad…now that we have included the synthetic data, it is going a lot better. So in terms of business impact, of course, knowing what your customers are going to do in the future is incredibly important for*

all aspects of business...it has definitely allowed us to gain a lot of better insights into what our customers are going to do" (Table.8. reference 32).

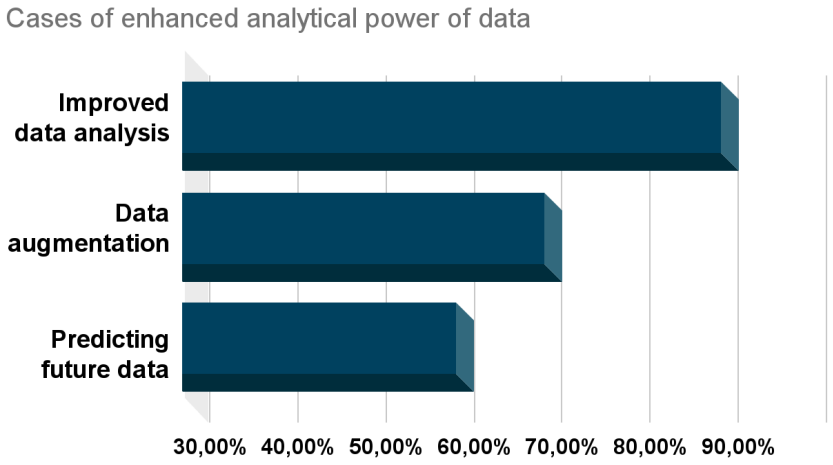

Figure.7: percentages of respondents reporting enhanced analysis

Data availability and data quality do not necessarily go in pairs, and the quality of data is an extremely relevant factor for the performance of a firm that relies on data collection and analysis. Some theories suggest that firms that have positive experience in using external sources of that data (defined as the benefits perceived) increase the willingness in adopting data analytics, while a positive experience in using internal sources reduces it (Kwon et at., 2014). However, all interviewees reported some degree of enhancement in the value of their data analysis, thanks to synthetically generated data, which is essentially artificial data created inside the firm. Thus, contrary to this view, it could be argued that a positive experience with synthetic data, being it an internal source, could actually foster the adoption of data analytics. This also shows evidence in support of synthetic data being a valuable internal source of quality knowledge through data augmentation when the necessary external knowledge is inaccessible.

The impact of synthetic data on the efficiency of an international firms is clearly visible in the words of an interviewee reflecting on the practical benefits of the technology: "*Efficiency, because the fact of being a synthetic data means that you have all and only the necessary information…Performance, because things are done in due time…the goal is to optimize traffic management…Safety, for people who must communicate, manage to communicate without errors and this can be done with the standardization of information, so a synthetic data has a standard format, understandable by everyone, everyone within that*

data manages to take their portion of information to perform their task properly" (Table.8: reference 14).

Furthermore, if we consider Knight's (1921) definition of uncertainty as "*the lack of knowledge about the probabilities of the future state of events*", the ability of synthetic data to allow predictions on future data, like consumers' behavior (Table.8: reference 32), is a key argument in favor of considering the technology as a tool to effectively address uncertainty and gain a competitive advantage.

Cost Advantages

The third category of strategic impact addresses the cost advantages linked with adopting synthetic data, which were reported by 80% of the interviewee and can be divided into three main benefits (Figure.8).

The first cost reduction, reported by 50% of respondents, is associated with the possibility of using synthetic data to conduct **pre-performance testing**, before manufacturing a product, making a strategic change or operational decisions (Table.9: references 1-12). An interviewee presented an intuitive example: "*You want to change your production process and you say how much it costs me to understand if I can change the production process in this way. On the one hand you would have to produce the robot, train it, put it on the assembly line, observe it for a month before understanding, and maybe this will cost you 100 million. If instead you do it in this synthetic, completely virtual environment, maybe it costs you a million. Those are numbers…I don't think they are that far from reality*" (Table.9: reference 3). Other respondents reported cases in which doing real testing would require prohibiting costs: "*In the nuclear fields, for example… some tests, it may cost them billions*" (Table.9: reference 5).

Another cost advantage emerges because the adoption of synthetic data, in many cases, **reduces the need to employ numerous skilled workers** to do tasks with the same level of efficiency and accuracy (Table.9: references 13-17) as mentioned by 30% of the interviewees. For example, an interviewee noted that synthetic data "*Allows you to simplify processes that otherwise…would require you to hire specialized personnel or would take a long time. So cut operating costs"* (Table.9: reference 15); while another interviewee stated

"*It was less expensive because we didn't have to pay humans to make the special training data sets all the time*" (Table.9: reference 17)

Finally, generating synthetic data was also reported as being decidedly less expensive than collecting real data (Table.9: references 18-20), thus further cutting costs for 20% of the interviewees. This advantage is linked to the main fact that the **synthetic generation process itself does not require prohibitive costs** as an interviewee noted: "*It's orders of magnitude cheaper…we're not going to charge you a crazy amount of money, we're not going to charge you an arm and a leg, it's not going to cost you thousands of dollars, it shouldn't, there's no reason*" (Table.9: reference 19).

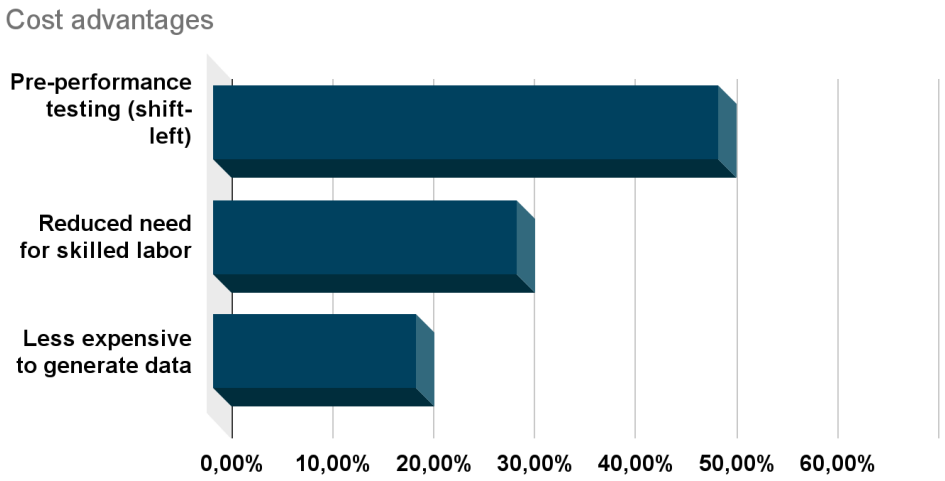

Figure.8: percentages of respondents reporting cost advantages

International firms are especially concerned with the costs of acquiring data, since data collection across different countries can easily become too demanding in terms of costs with the future regulation on data privacy that will potentially restrict data access even more and increase the cost of collecting the necessary information. Particularly, one interviewee stated: "*by 2040, if I'm not mistaken, 70 percent of the data will be synthetic, and I have no difficulty believing it because data is expensive anyway*", further confirming that data scarcity represents a relevant issue and that synthetic data is a viable tool to address it (Table.9: reference 18).

The positive impact on operational costs represents one of the most immediate strategic advantages of implementing synthetic data in firms that rely on data, but this benefit is not only limited to synthetic data generation being merely cheaper than real data collection.

Allowing to test a process or product in advance, synthetic data helps save costs in terms of prevententing malfunctioning or failing after the release on the market. There is a specific concept in the field of software management called shit-left which is based on “taking action early on” (Vaddadi et al., 2023), moving the testing phase from the end of a project to the beginning. This approach serves the purpose of ensuring that design flaws and major issues are discovered and fixed as they get coded, instead of being addressed at the very conclusion when it could be too late to reform the core structure of a software. A respondent highlighted that pre-production testing may be the single most significant advantage of this new technology: *“the principle of -fail big, fail soon-... all of these are principles of agile development, of being fast. If you have something that blocks you that happens at the end, you can't be fast by definition. You have to discover the bad things as soon as possible…So, in my opinion, there is the advantage, there is the real advantage*” (Table.9: reference 12). Synthetic data characteristic of ensuring realistic simulations in advance represents fertile ground to allow shift-left outside of the realm of software management, potentially becoming a strategy that can be applied on a larger scale at organizational level. This idea can be compared to the managerial theories on persistence and transfer of learning (Argote et al., 1990), which suggest that knowledge acquired through production degrades more easily, while starting production later, after acquiring knowledge from other organizations that are already producing, improves productivity in the early stages. This may suggest that, by providing a tool for pre-production testing and optimization, synthetic data may bring early benefits similar to those obtained relying on the experience of already operating firms, with the advantage of not needing external organizations because firms can generate this knowledge from the inside. For the firms that are adopting this technology, This is another case of synthetic data not only being a source of great adaptability and therefore dynamic capabilities, but also presenting the requisite of being valuable, rare, imperfectly imitable and non substitutable, for uniquely allowing to move the testing phase at the very beginning, before the production of even a single unit of product.

Additionally, even though we are in the early stages of this technology, 30% of the respondents, as previously shown in Figure.8, already reported a decreased need for skilled workers with the implementation of synthetic data. This can open the relevant and ongoing discussion on the implication of new technologies on employment: on one hand the risk that machines will replace workers (Leontief, 1952), on the other the idea that machines complement human workers forming together a “supermind” (Malone et al. , 2020). Some

interviewee reported an example of synthetic data impact on employment: "*[Synthetic data] frees them from having to hire ten experts to be able to set something up, and we do this through synthetic data so, let's say, there is artificial intelligence that understands a synthetic given process produced for those who don't have the skills to fully understand that proces*s" (Table.9: reference 14). Moreover, the increased processing power as opposed to human capabilities, ties into the theories of decision making that discuss the concept of rational inattention (Sims, 2003) which suggests that "*organizations allocate their limited cognitive resources to process only the most relevant information due to the costs associated with acquiring and processing information*". While there is no way of knowing with certainty what the full impact on work of technologies like synthetic data will be, the task-based model proposed by Acemoglu & Autor (2011) represents a useful framework for better understanding the possible future dynamics. The authors suggest that technological developments, including those that cause substitution of employees, lead to job polarization. This entails a reduced need for moderate-skilled jobs with an increased need for high-skilled and low-skilled jobs. This approach shifts the focus from only reflecting on whether there will be replacement of workers, to better understand the overall impact of the technology, also considering the new job opportunities linked to the rise of synthetic data.

Legal Compliance

When it comes to data scarcity, privacy regulations are a necessary factor to address, due to their effect on restricting access to real data, therefore representing one of the sources of the problem. The results on how synthetic data can help compliance with data privacy regulation show that there are several advantages (Table.10) linked with the implementation of synthetic data as well as some drawbacks (Table.11).

Legal Advantages

Even those interviewees that are not using synthetic data for compliance purposes in their current projects still expressed how relevant its application in this area is. Four main themes emerged when investigating synthetic data benefits for legal compliance (figure.9).

The biggest advantage was the possibility to **use and share sensitive data** reported by 80% of the respondents (Table.10: references 1-23). An interviewee mentioned the fact that the earliest use of synthetic data specifically addressed privacy issue: "*The first uses that were*

made, synthetic data, were probably already ten years ago in the field of compliance, and they increased considerably with the approval of the European law GDPR…it is a matter of replicating information that cannot be used because it is too private" (table.10: reference 5). Another respondent highlighted that synthetic data was used to reduce the consequences of a potential data breach. "Y*ou have much less of a worry of cybersecurity and being hacked…if your models are being developed by synthetic data…the underlying foundation of that information is coming from essentially an artificial place*" (Table.10: reference 6);

40% of the interviewees (Table.10: references 24-28); mentioned cases in which synthetic data represents the **only viable solution** to compliance. This happens in circumstances in which traditional techniques are no longer effective: "*Companies, they need to share data, and they know that they cannot. So, the old techniques are not working anymore…Data anonymization or data masking, it used to work before. But now, in the context of big data, it's very heavy*" (Table.10: reference 26); and circumstances in which not using synthetic data would represent a burden too heavy for an international firms with many employees handling the data: "*It was the only feasible thing, I mean, there are no other possibilities…alternatively you had to have very high levels of NDA that were not compatible, you know, with the number of employees"* (Table.10: reference 28);

Improvements in **data retention,** were stated by 10% of the interviewees (Table.10: reference 29); Particularly, real data is legally required to be deleted in many instances but synthetic data, being artificially generated can be kept: "*you have to delete the data periodically, and you want to retain a synthetic version, so that you can still look back*" (Table.10: reference 29);

A more **ethical approaches to data use** thanks to synthetic data was reported by 10% of respondents: "*You can also use it from more of an ethical motivation, say I'm already allowed to do this with the real data, but I could answer the same question with synthetic data*" (Table.10: reference 30);

Finally, 10% of the interviewees also mentioned the possibility to **use sensitive data in real time** with synthetic data without concerns for privacy: "*And also, in terms of real-time applications, [real data] does not fit with the requirements. So, the only solution is to allow*

developers to create the data they want, when they want, and they load it to the applications they want" (Table.9: reference 31).

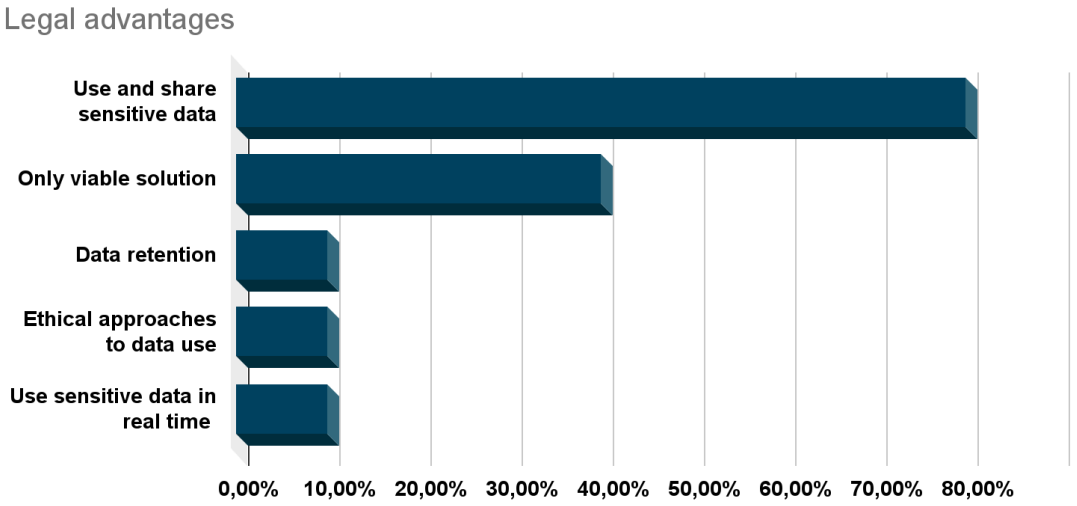

Figure.9: percentages of respondents reporting legal advantages

The results confirm that synthetic data represent an important tool to aid compliance especially for international firms that have to transfer sensitive information across different countries: "*Our biggest clients do have a big Indian team. And due to GDPR, they cannot share the real data with developers*" (Table.10: reference 14). With the tightening of privacy regulations, the ability of ensuring safe transferability of information and data within a single firm and across a number of countries could prove essential in operating efficiently on an international level, thus demonstrating that, also from a legal point of view, synthetic data can be a source of sustained competitive advantages (Barney, 1991), compared to firms that share real data traditionally. Additionally, ensuring a compliant channel for accessing and sharing sensitive data of local consumers internationally could become an innovative way of decreasing the "liability of the foreigners" (Zaheer & Mosakowski, 1997) of international firms that want to enter a new market and have difficulties in accessing the relevant information due to regulations.

Notably, an interviewee reported that synthetic data is starting to be used also by formal institutions for safely sharing data with the public: "*The Dutch ministry of education…use it for legal compliance, so they would not be allowed to share the real data, but they share the synthetic data instead*" (Table.10: reference 12). If other institutions start to adopt synthetic data it could lead to a case of mimetic institutional isomorphism (DiMaggio & Powell, 1983), with firms adopting the practice of successful organizations. Over time, this

could also potentially evolve into coercive or normative isomorphism as the laws and standards on data privacy evolve.

Legal Disadvantages

Some degree of complications when utilizing synthetic data for compliance purposes were reportedly linked to three main causes (figure.10).

The main issue mentioned is linked to the fact that synthetic data is not always fully synthetic, but can have the **underlying presence of real data** (Table.11: references 1-6). 40% of respondents identified this as an issue, for example an interviewee stated: "*we talk about the privacy utility trade-off, where the utility is the quality of the data…you want your data to be similar enough that it's, that you can use it and that you can answer questions with it that you would have wanted to answer with the original data, but not so similar that it starts leaking information about the original individual*" (Table.11: reference 4). Some particular concernes with small dataset were also reported: "*if the group is smaller than 10, you are not allowed to use the exact number…small number of people satisfying a certain criteria…it would be almost possible to determine which people those are*" (Table.11: reference 5);

Another problem is linked to the **uncertainty of the legal scenario** surrounding the new technology, mentioned by 20% of the respondents (Table.11: references 7-10) On one hand, some doubts on the future regulatory environment for data protection were expressed "*There's always some interpretation of the guidelines and laws concerning synthetic data*" (Table.11: reference 7); on the other there is the question about how private synthetic data based on real data really is: "*The biggest issue at the moment because it is difficult to say when synthetic data really are sufficiently private*" (Table.11: reference 9);

A last issue mentioned by a single interviewee is related to the highly specific request of "**opting-out**" of a dataset and its implication for the synthetically engineered dataset: "*If somebody wants to opt out of the original data, they should also no longer be represented in the synthetic data…if you really want to completely be removed from the data set of an organization, then you shouldn't still be kind of half represented in the synthetic data*" (Table.11: reference 11).

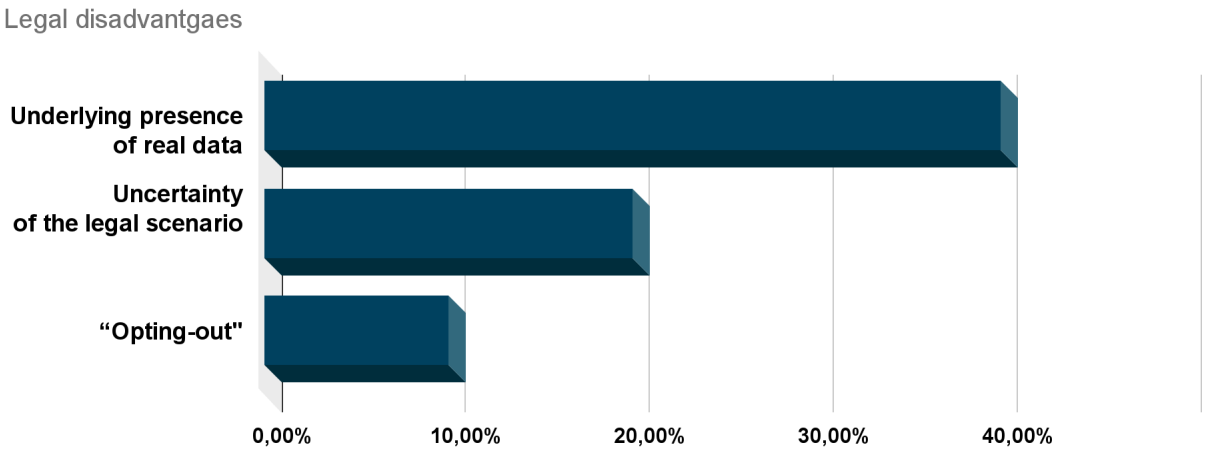

Figure.10: percentages of respondents reporting legal disadvantages

The results show that the climate of legal uncertainty surrounding synthetic data is not attributed to the technology itself, but rather to how the legislator defines privacy. Before commenting on whether this new technology can be a true safe alternative to inaccessible or sensitive real data, from a compliance point of view, one must reflect on what exactly is privacy. One of the interviewees acutely noted: "*how to measure privacy is probably the biggest challenge…because privacy is, of course, it's a human construct. It's a legal construct. And even if you can measure it, it is sometimes difficult to make the measurements interpretable. And there's sometimes a trade-off between to what extent you can measure it very quickly for large groups of people and to what extent you can actually interpret it*" (Table.11: reference 10). It appears that this uncertainty is unrelated to what the technology can actually do, and it has more to do with what can be considered legally safe: "*The legislator essentially has the absolute power to prohibit you from using any type of data and when we talk about synthetic data we are talking about data that is synthetic up to a certain point because in any case you have to know something, otherwise you cannot produce any type of data synthetic*" (Table.11: reference 1). Therefore, the uncertainty is not caused by the novel nature of synthetic data, but it derives from the unpredictability of what privacy regulations will prescribe.

Until regulators worldwide set clear guidelines and requisite, organizations, especially international ones, must decide on how to react to data scarcity that is already an established issue. Looking at strategic management theories, Freeman recognized: "*good theories of management are practical...Not only they predict what may happen and allow managers to adjust to those predictions, but they explain the existence of certain phenomena and the relationship which these phenomena bear to other phenomena*" (Freeman, 2010) In this

context, strategic management could be incentivized to consider the adoption of synthetic data even more so, since it does not only provides a practical approach to react to current data privacy concerns, but it also allows to anticipate the future regulatory environment, proving again a source of dynamic capabilities (Teece et al., 1997) to adapt to the uncertain legal developments.

Challenges and Limitations

To offer a comprehensive understanding on the current implications of synthetic data for international firms, it is essential to investigate what are the main challenges associated with its adoption and what are the limitations of this technology. Particularly, the results have shown that three main categories of relevant problems can be defined: The first is related to **applicability and effectiveness issues** (Table.12); the second entails the **approaches towards bias** in the data (Table.13); the last is linked with **initial reservations** (Table.15) manifested by firms when first integrating synthetic data.

Applicability and effectiveness issues

When it comes to applicability and effectiveness of the data, several specific circumstances have been reported (Figure.11):

In certain cases the synthetic data **model was not sufficiently complex** (Table.12: references 1-6) and did not allow for effective data analysis. This was noted by 60% of the interviewees. For example, an interviewee stated "*The development process is an iterative one. So, in the first iterations we typically encounter some cases that are not covered well by our synthetic data sets, and then we expand the data sets to include more examples*" (Table.12: reference 1);

40% of the respondents mentioned that **domain expertise** represented an obstacle to the generation of reliable synthetic data (Table.12: references 7-14). Specifically, some interviewees noted that they were not able to generate certain data due to lack of relevant knowledge: "*We've been asked many times to generate data where we also beforehand say "We can't possibly do this, We don't have the domain knowledge and also that knowledge is not available in the world*" (Table.12: reference 7), while other highlighted the value of having expert to validate the data: "*The humans that were doing that job… they're former mechanics. So, they know when they see car damage "oh, this component underneath is*

definitely damaged". But in the synthetic data, and even from novices looking at these images, we don't know that insight" (Table.12: reference 12);

In other cases it was the **initial research** necessary to set up synthetic data for the first time that represented the main challenge. This was reported by 40% of the respondents (Table.12: reference 15-20). In particular, an interviewee highlighted the iterative process necessary to obtain satisfactory results: "*there would be instances where solutions wouldn't work and you have to go back to the drawing board and it was a very iterative process between data scientists and our customer side*" (Table.12: reference 16).

30% of the interviewees noted that synthetic data's potential is sometimes exaggerated to the point of being advertised as an all encompassing solution to too many problems, leading to an **overselling of its real capabilities** (Table.12: reference 21-24). For example an interviewee noted: "*it was a little bit harder than we thought to get it right…So, our first idea was that it was rather straightforward to generate a synthetic population, but it turned out to be more difficult than we thought*" (Table.12: reference 18);

Another issue reported by 20% of the respondents was linked to the necessity of **updating the synthetic dataset with real data** (Table.12: references 25; 26). Specifically, it was mentioned the need to rejuvenate synthetic dataset after a certain amount of time to ensure that it was still representative: "*they must continue to be fed by reality, there is nothing to do, because otherwise we risk fossilizing and entering situations that are not representative of reality, because reality evolves, it brings us to the unforeseen. So, if I think of taking a photograph of reality, I transform this photograph into synthetic data, then at a certain point I need to update this photograph after a while, because otherwise that data is no longer representative of reality, because the external scenario evolves*" (Table.12: reference 25).

Lastly, an **increased computation power** (Table.12: references 27-29) for synthetic data was also noted by 30% of the respondents. It was especially a concern when dealing with cases of data augmentation with a combination of real and synthetic data: "*A bit of a problem was the computing power that we needed to be able to handle all of the big volume of data. Because, yeah, with the synthetic data, in addition to our real data, we had quite a large sample. And that caused some strains on our computing power*" (Table.12: reference 27).

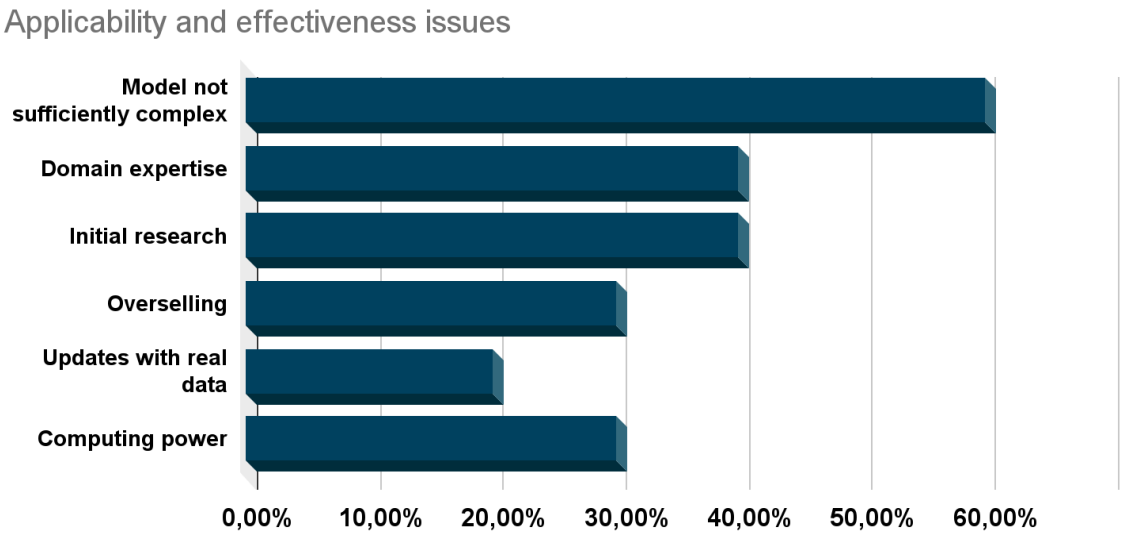

Figure.11: percentages of reported applicability and effectiveness issues

It is essential for every firm to invest resources into understanding what is the best course of action for a given goal. This reflects the mentioned need for an iterative process in order to obtain quality synthetic data through sufficient initial research. Furthermore, the reported difficulties in reaching the needed model complexity is also a reflection of the inherent complexity of the real world. Initial efforts for understanding new circumstances are always required to businesses, especially international ones that should be used to monitor many new markets in different countries. The adoption of synthetic data is no different: a new technology requires new knowledge and adjustments, and the need to conduct deep initial research could be seen as a physiological requirement to introduce new technologies.

Some respondents mentioned that synthetic data required heavy initial research and more computational effort than just processing collected data with traditional methods. It must be asked whether the value of synthetic data really does make a difference, from a strategic point of view, to justify the applicability and effectiveness issues reported. The results suggest that this more demanding approach is met with an increase in the analytical power of data, thus compensating these complications by bringing new depth of analysis. Synthetic data seem to offer an improved way of processing data and this aligns with the studies in the field of international knowledge management which highlights that, for the growth of an international firm, it is relevant not only the access to knowledge in other countries, but also the "*appropriate processing of knowledge from other nations and cultures*" (Kuemmerle, 2002). However, it must be also noted that all the international firms interviewed successfully

invested in the R&D necessary to research and refine the technology to optimal level, and this study could not test what is the percentage of firms that attempted and failed the initial research stage. This data could prove relevant in assessing how worrisome this phase should really be for firms evaluating the technology for the first time. Furthermore, the firms included in the sample all eventually conquered the initial stage and mastered the technology with confidence, but some still incurred the problem of having to refine the synthetic model to adapt it to a specific use later on. It remains unclear how demanding the occasional refinement of the synthetic dataset is.

Additionally, just as it is important to recognize when concepts mentioned as issues represent more a normal required effort than a problem, it is also essential to identify when the technology starts to become idealized as a perfect solution. This was effectively captured by a respondent that identified this tendency in the request of the management to include in their operation an AI trained on synthetic data: "th*e executive leaders are like, "Oh, AI, we need to include AI". And, and it's like, no, nobody that was just in a car accident wants to chat with the chat bot, like, get real*" (Table.12: reference 23). It is clear that synthetic data is rising interest in every field, but it is also important not to lose track of what are its limitations.

Approaches to bias

As discussed in the literature review, bias is a major concern when it comes to synthetic data. This section will attempt to discuss how real international firms approach bias when integrating synthetic data. Specifically, three main approaches were discussed (figure.12).

20% of the interviewee believes that synthetic data, in certain circumstances, actually helped reduce bias, to the point of considering the synthetic dataset **less biased than real data** (Table.13: references 1-2). Namely, an interviewee reported: "*There were other instances with other customers who had tons of real data, but it was biased because the majority of the real data happened only [a few times]. We had one defense company who was in the maritime industry. And the majority of their image datasets were of sunny days. And it's like, you need a little bit of different variety in the weather, and different variety also of the ocean, the rolling waves, for example, not just calm waves. so there, in that instance, they were able to combine our synthetic with their real*" (Table.13: reference 2);

In 40% of instances, **bias was actually desirable** (Table.13: references 3-9). This was the case when the firm was employing synthetic data to better predict consumers behavior: "*we're also not trying to eliminate bias necessarily. This use case was about predicting customer behavior and, of course, customers are people, and people are inherently biased. So to a certain extent, it is fine that there is bias in the data*" (Table.13: references 4);

Some degrees of **effort to correct certain types of bias** (Table.13: references 10-17) were reported by 70% of the interviewees. Even though synthetic data and AI go often hand in hand with each other, in some cases it was recognized that using AI could represent an additional risk when it comes to bias, as opposed to more traditional statistical methods: "*if you're using more statistical approaches, I think the probability is lower that your bias will be amplified…if you're using AI, because that's even why when we use data augmentation. We use that specifically because AI doesn't represent every group equally*" (Table.13: reference 13).

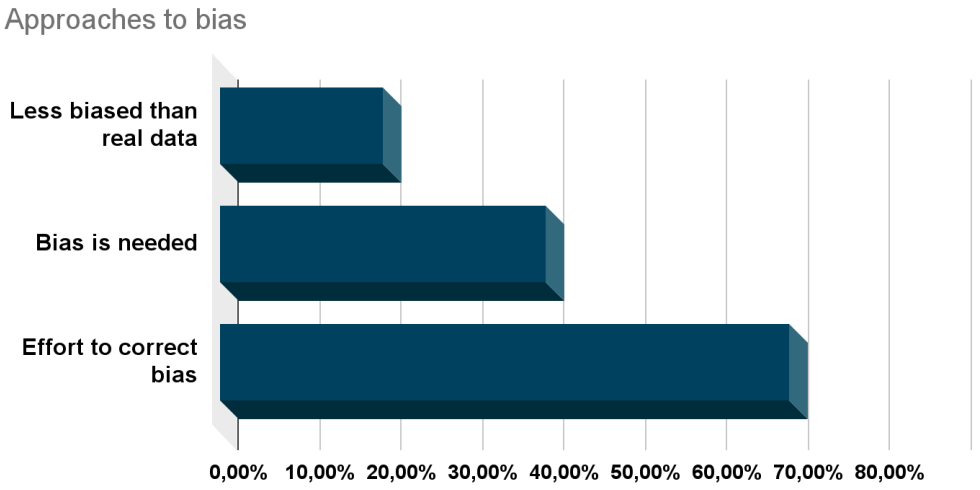

Figure.12: percentages of reported approaches to bias

Bias amplification was described by the literature as one of the main concerns (Giuffrè & Shung, 2023), with no clear path to ensure with certainty its containment. However, the results of this study show that none of the interviewed international firms reported that bias was a major issue that could not be effectively addressed through different strategies and methods, or that it was a specific concern related to synthetic data more than with real data.

Interestingly, while recognizing and addressing bias was still relevant, there were instances in which bias needed to be preserved: "*when I went to see the algorithm with which*

they made these purchase predictions… there was definitely a bias, but it's an informed bias…you, with your bias, lead the market…it never occurred to me the idea that a bias should be corrected" (Table.13: reference 9). This can be linked with the behavioral economic theories that suggest that humans have bounded rationality (Simon, 1990), which is a behavioral bias that impacts their decision-making. Therefore, when using synthetic data to predict customer behaviors, it appears clear that bias is a vital component of these predictions as an interviewee reported: "*Humans were always biased, by the way. So, when you need data, you need data that's mimicking your reality*" (Table.13: reference 7).

Initial Reservations

80% of the respondents reported some initial reservations when first approaching the decision on the adoption of synthetic data (Table.15). The reservations mentioned can be classified in three main themes (Figure.13).

40% of the interviewees noted that the biggest obstacle was represented by the need of showing **proof of effectiveness** (Table.15: references 1-4) of synthetic data. Specifically, a respondent noted the difficulty to prove synthetic data value in advance, especially in certain sectors: "*There's definitely some convincing required…for new markets, almost always, we have to create some examples ourselves, just to prove we can do this. It's also very dependent on the market itself. So, for instance, in semicon, there are more larger companies that are used to taking some risk on this new type of technology. So, it's a bit easier to get in as a proof of concept or to prove yourself within an initial project, within a research project. For medical projects, that is more difficult…there we need a real, very, very solid track record and prove how well this works*" (Table.15: reference 1);

In 40% of cases it was the idea that **synthetic data is not authentic data** (Table.15: references 5-8) the main issue; Some interviewees reported the difficulty in accepting synthetically generated data as equally valid compare to real data : "*Real data is always more, how to say, complete and also unpredictable than a synthetic one. The synthetic one must tend to encompass this nature of unpredictability*" (Table.15: reference 8);

Some interviewees, accounting for 20% of the respondents, stated that the initial concerns were related to **security and privacy** (table.15: references 9-10). Specifically, it was

reported the uncertainty regarding how much of the real data can be identified in the synthetic dataset: "*The privacy part is a big one, so kind of the legal side, also the ethical side, but yeah, how can we really know that it's not leaking anything?*" (Table.15: reference 10).

The remaining 20% of the interviewee reported **no initial reservations** (Table.15: references 11-12): there were cases in which no concern was present at all: "*I immediately thought it was a very good idea. And it is also not completely new. It is also used by other research programs*" (Table.15: reference 11).

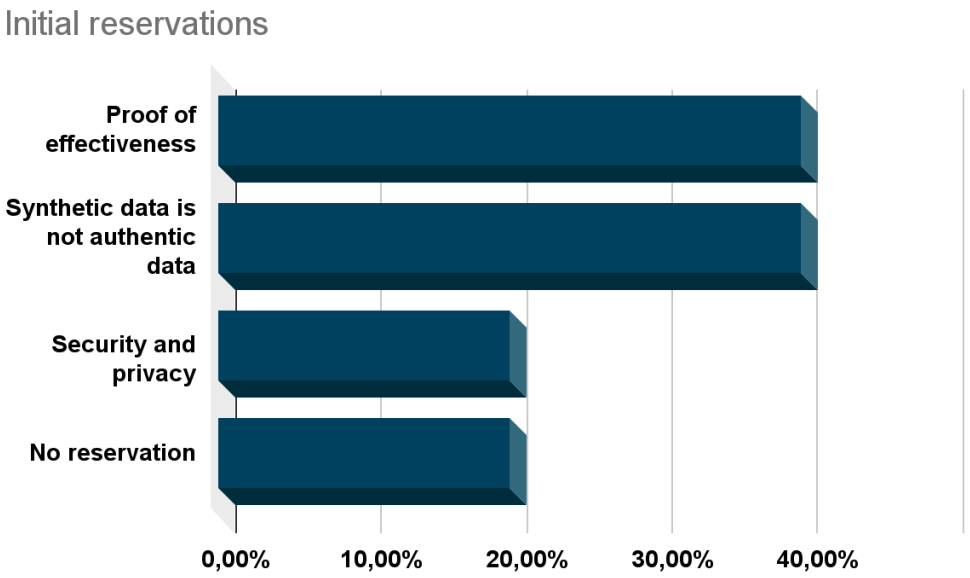

Figure.13: respondents reporting initial reservations

When investigating the results related to the initial reservations surrounding synthetic data, it is important to first understand at what stage of diffusion this technology actually is. Synthetic data seems to be currently between the introduction and growth stage of its life-cycle (Levitt, 1965). In some industries, and for some applications, its potential is already well established, but synthetic data notoriety is expanding further, also thanks to the diffusion of AI trained with it. These initial reservations can be interpreted as the natural uncertainty that comes with adopting a new technology, rather than being perceived as a signal of a well-founded distrust towards it. An interviewee conveyed this message with a brilliant metaphor: "*When we go to a pharmacy and we buy vitamin C, for example, if it is synthetic or if it is natural, you will go to the natural one, you know, even if it is more expensive…Resistance to change is what's always here*" (Table.13: reference 7). Furthermore,

20% of the interviewees reported no initial reservations at all, in some cases also describing synthetic data as the "*...only thing feasible*" (Table.14: reference 12).

Future Perspectives

Synthetic data is not a new discovery in the field of computer science, on the other hand its use in real business settings is much more recent. Therefore, it can be relevant to understand what firms that are starting to adopt it foresee for the future development of this technology. The interviewees highlighted two main themes when discussing the future of synthetic data: on one hand all respondents reported **ongoing research efforts** (Figure.14) related to further applications of synthetic data; 50% also reported another focus on **improving reliability** (Figure.15) of the data.

When it comes to research efforts, the results show that two main areas of research were mentioned. The first revolves around synthetic data capability of **enabling new technologies**, reported by 80% of the interviewees (Table.16: references 1-13). Particularly, one interviewee mentioned that synthetic data is to be considered as just the starting point of new discoveries to come: "*In the end, synthetic data is a tool. It's not an end goal. It's a tool to enable development of other algorithms down the pipe…Reinforcement learning is typically something that is really greatly helped by simulation synthetic data*" (Table.16: reference 1). The second category was focused on achieving **wider applicability** of synthetic data, mentioned by 80% of the respondents (Table.16: reference 14-23). An interviewee reported that their customers also joined in the exploration of new possible applications: "*We see that even with our clients, they are keen to understand more and more the technology…It's also on the agenda of many companies. Clients, not only willing to use the technology, but also…would like to get their hands even dirty. They would like to develop and participate in some use cases, which is good*" (Table.16: reference 20).

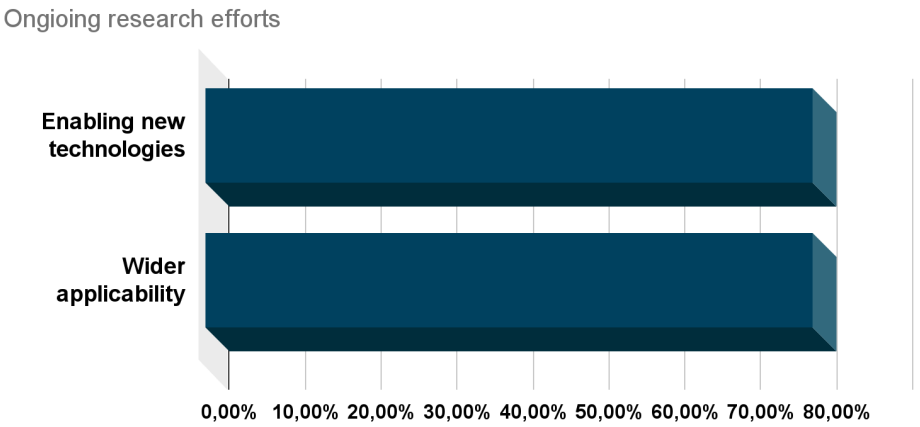

Figure.14: respondents reporting ongoing research efforts

Among those interviewees who reported improvements in reliability of synthetic data as a programmed future goal, the two main aspects of security and quality arose. Specifically, 40% of the interviewees reported an effort to **increase security** (Table.17: references 1-4) of the data, for example an interviewee commenting on storing customers' data: "*We might not need to store as much data ourselves, which is, of course, also better for security and privacy reasons that you don't store all customers' data directly*" (Table.17: reference 2). Another 30% of respondents mentioned an effort to **increase quality** (Table.17: references: 5-7) of the data. For example, an interviewee reported a possible decrease in the amount of data that firms need to collect: "*As we get more and more accurate ways of generating synthetic data, it will make just the entire process of collecting data much easier. Because you might not need to collect thousands of responses from customers if you can use 100 and then create a representative sample using synthetic data… So that'll obviously make a big difference in just survey costs*" (Table.17: reference 6).

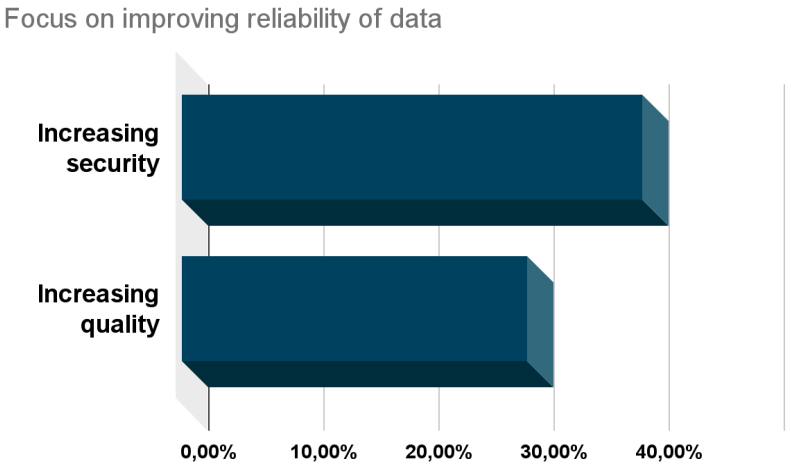

Figure.15: respondents reporting focus on improving data reliability

Even when there are several evidence pointing in a specific direction, the future is always uncertain to a certain degree. Therefore, it is important to take into account that the interviewees' answers surrounding synthetic data's future potential may be dictated by the hope that this technology will bring them new and significant benefits. It is not possible to determine how realistic the future innovations mentioned are. A respondent gave a particularly comprehensive and optimistic overview of some of the biggest goals that could be potentially reached thanks to synthetic data: "*detecting litter in the seas and the oceans and cleaning that in a more efficient, affordable way…cancer detection…climate change predictions*" (Table.16: reference 4). A notable reflection of one of the interviewees was related to the adoption of synthetic data is starting to spread even in those industries in which introducing new technologies usually requires a long process of verification: "*If the aeronautical domain, as you can see today, is fully engaging in [these] applications… even though it is so safety critical, it means that we are really witnessing a revolution*" (Table.16: reference 21).

Interestingly, when discussing the future of synthetic data, some interviewee noted the necessity to direct this new technology towards the common good: "*If the driver of synthetic data generation, if the drivers are inclusiveness, environmental sustainability, but in general, then surely synthetic data could be a tool that guarantees, over time, sustainability, equity, inclusiveness and so on*" (Table.16: reference 22). As of now, it is only possible to wait to see how the regulators worldwide will decide to regulate this technology that is described by the interviewee as one of the most promising of our times.

Further studies will be necessary to see whether the mentioned future applications of synthetic data will be in line with what the respondents highlighted in this research.

Response to UN recommendations

The findings have shown three main approaches towards the UN recommendations (Figure.16). 40% of the respondents is **unaware** (Table.18: references 1-4) of their existence; 30% regarded them **useful as generic guidelines** (Table.18: references 5-8), an interviewee commented a luck of practical suggestions to aid synthetic data safe use: "*Those 13 points are absolutely reasonable, nothing to object to. For the moment, from my point of view, as often happens, they are also very obscure to implement, they are a bit philosophical if you like,*

there is no clear roadmap on what companies need to do to be able to do so" (Table.18: reference 5); the remaining 30% stated their **indifference** towards the recommendations (Table.18: references 9-11), for example, a respondent highlighted that these guidelines would be taken into account only if requested by a client: "*Unless our customers come to us and say, look, the UN said A, B, and C, and I want to make sure that that's in there, no problem. Then we will make sure that our solutions are adhering to whatever regulation is necessary*" (Table.18: reference 9).

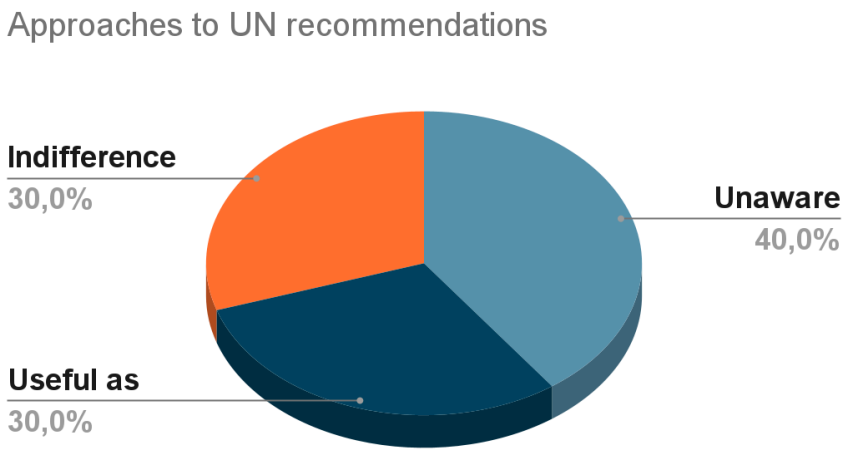

Figure.16: respondents' approaches to UN recommendations

The data collection for this study takes place three months after the release of the UN recommendations, and this factor could explain why some firms are still unaware of them. More relevant seems to be the views of those who had the chance to review the recommendations and stated that they do not represent relevant points for the firm's operations. One, specifically, is the answer that stands out: "*[the UN] are still not aware about what even is synthetic data... the recommendation from the United Nations is well, you know, to help in argumentation*" (Table.18: reference 11). This answer stresses an existing detachment, as often occurs, between new technology users and formal approaches by regulators. The approaches toward these recommendations can represent a proxy for the international firms' perception of policy that regulate new technologies. Even though it is not possible nor correct to infer that UN recommendations and official legal prescriptions are similar, the results of Table.18 can offer some insight on the perception that international firms have on regulators' knowledge and expertise regarding this technology or, perhaps, new technologies more generally. Another interesting perspective is the idea that regulators are willingly holding off official regulations because of the high potential that synthetic data

could have combined with AI: "*The impression that we in the sector have that for the moment the legislator is rightly concerned with the interests of citizens without wanting to put too many spokes in the wheels in a technology that of AI which is perhaps the only opportunity to have that production leap at an economy of scale level that we have been waiting for for probably 20 years and which can in turn be perhaps the only solution to many problems such as that of costs of social safety nets, social security, the aging of the population, the lack of specialized personnel*" (Table.18: reference 6).

Conclusions

Key findings and Contribution

The possibility to use synthetic data for a wide variety of applications and the explored technicalities suggests that it can represent a key element to enhance a firm flexibility and adaptability. Therefore, it does represent a relevant source of dynamic capabilities. Evidence suggests that having diverse sources to access a technology can increase entry rates in certain industries (Lieberman, 1989) and, in relation to this, synthetic data can count on a number of generation techniques, from the most advanced to more simple and approachable methods. This could facilitate its successful adoption as a means to aid entry in other countries; moreover, there is the innovative possibility of generating knowledge internally, which could have a relevant impact on internationalization success in cases of inaccessibility or absence of real data. Specifically, synthetic data could change the way firms experience internal and external sources of data (Kwon et at., 2014), offering the value of having a customized internal source of knowledge, as opposed to the traditional view that stresses the necessity of external knowledge (Chung & Yeaple, 2008; Awate et al., 2015). This new approach of internal knowledge acquisition also challenges the main theories on the persistence and transfer of learning (Argote et al., 1990), which pose the accent on the importance of acquiring knowledge from external organizations that are already operating. Additionally, although many of the methods mentioned to ensure quality and relevance of synthetic data require real data, techniques that do not require it are already existent, testifying that synthetic data can indeed be a reliable and quality solution in cases of limitations to data access.

Another set of considerations is related to the comparison between real and synthetic data, The results confirm that synthetic data can solve many of the privacy issues related to using real data (Ruiz et al., 2018), while offering the same analytical value (Lucini, 2021).

Even when possible to use both real and synthetic data together, in many cases it was desiderable to not include real sensitive data for privacy reasons. Furthermore, allowing the safe transferability of information across countries, synthetic data could be seen as an additional tool to help reduce the “liability of the foreigners” (Zaheer & Mosakowski, 1997) when entering new markets. The possibility to safely acquire and share data across borders confirms that, from a legal point of view, synthetic data does serve the purpose of solving the data scarcity caused by the tightening of privacy regulations, therefore representing a valuable resource, following the resource-based view (Barney, 1991).

A relevant aspect for studying the future legal scenario surrounding synthetic data is also the role of formal institutional approaches toward this technology. Some institutions, like the Dutch Ministry of education, are already adopting it to safely publish sensitive information and, following the institutional theory (DiMaggio & Powell, 1983), it could be argued that more firms, or perhaps even other institutions, will follow this successful example. Using the approaches of firms to the thirteen recommendations of the UN as a proxy for perception that international firms have on regulators' expertise on synthetic data, it is also possible to notice a lack of trust of the regulators knowledge of this new technology.

Contingency theory (Donaldson, 2001) highlights the absence of a universal strategy for every firm and the importance of circumstances, and synthetic data ability to simulate highly specific, extreme or unique cases for testing purposes also means that synthetic data is a new technology that actually reduces uncertainty, and presents the feature of rarity, imperfect imitability. Additionally, Synthetic data is even considered indispensable or the only approach possible in some cases, confirming also the non substitutability of the technology. From a strategic point of view, it can be argued that synthetic data can change the overall impact of data on a firm's performance, and its adoption is associated with relevant enhancement to data analysis. When it was mentioned that synthetic data had higher value compared to real data, it was also reported an increase in analysis results and firm performance, which is in line with the idea that higher quality data has a direct and indirect effect on performance (Côrte-Real et al., 2020). Synthetic data also represents a tool to not only access data, but also to allow international firms to effectively grow by processing these data more effectively, in agreement with the theory on the importance of processing data adequately, suggested by Kuemmerle (2002). Synthetic data, thus, proves to be a new source

of dynamic capabilities that gives international firms the flexibility to effectively process the data of a diverse and international customer base.

Another important reflection is on the interaction between technology and human workers. The idea that machines will replace humans (Leontief, 1952) is not new. However, synthetic data provides mixed evidence, on one hand confirming a reduction in the need of skilled workers; on the other, following the task-based model (Acemoglu & Autor, 2011), it could be argued that a shift in labor dynamics due to a new technology will bring new job opportunities as well, therefore suggesting that synthetic data can fall in the category of technology that complement human capabilities (Malone et al. , 2020).

Unexpectedly, even though bias amplification was described as one of the major risk of synthetic data (Giuffrè & Shung, 2023), the results have shown that in real business context, bias can actually become a key feature to preserve for analysis purposes, since it represent a measure of the inherent bias and bounded rationality (Simon, 1990) of human nature. However, other practical issues not yet mentioned by the literature arose. Specifically, the initial research stage to set-up the technology was particularly demanding for some organizations. All the firms interviewed successfully overcame this issue, therefore it was not possible to test the failure rate, which could be crucial to assess the risks for firms that are evaluating the adoption of this technology. Additionally, while the sampled firms all eventually mastered these initial requirements, some still faced the occasional challenge of insufficient model complexity and the need to refine the synthetic dataset. These complications could be eventually overcome by some firms, like the ones interviewed, but there is the risk that they could determine a failed attempt for others.

Practical relevance

This study's research questions aimed to investigate synthetic data as an internally generated resource that can impact managerial decisions of international firms, and as a resource that can allow them to address circumstances in which real data is not readily available or is even non-existent, both from a strategic and legal perspective. Through the lens of dynamic capabilities theory (Teece et al., 1997) and the resource-based view (Barney, 1991) it was possible to analyze this technology, as shown in the key findings, and now is

possible to summarize its practical implications for real international firms and their management.

From a strategic point of view, the results recognize that synthetic data does improve the way an international firm can react to changes and new circumstances thanks to its inherent ability to predict future data, mirror realistically existing real data in unique simulations or substitute it when access is limited. Therefore, managers could consider adopting it to improve a firm's flexibility and adaptability to uncertain circumstances or obstacles to data access. From a legal perspective, since synthetic data allow the safe transfer of information thanks to its anonymization ability, and managers could decide to adopt it to continue to share data globally among subsidiaries, while complying with the present regulations on data privacy that reduce data access, as well as planning in advance for future and more restrictive international and national laws.

The instances in which synthetic data was recognized as indispensable, unavoidable, irreplaceable, or the only viable solution, highlights that the technology could become a relevant source of competitive advantage compared to firms that do not adopt it. Therefore, international firms should consider implementing it especially in fields where data analysis is very relevant or fields where the data collected is particularly sensitive or scarce. Even regulators may want to consider synthetic data as an important resource to facilitate data protection, as formal institutions are already starting to adopt it for this very purpose.

All things considered, since the technology is still not yet widely used, firms that act as first movers in its adoption can experience a sustained competitive advantage, which will decrease as it grows in popularity and more firms start to adopt it.

Final considerations

The results confirm the validity of the benefits suggested by the literature, also suggesting new applications and advantages, both from a strategic and legal point of view. But the findings on the challenges and limitations also uncover many practical concerns that were not discussed by the literature, which is still mainly focused on the technical aspects of this new technology. These issues could potentially prove too difficult to overcome and impede the successful adoption of the technology. Indeed, this work does not suggest that

synthetic data is a perfect alternative to real data and a solution to all data access problems and legal compliance, or that it represents the ideal strategic resource. Real data is and will always be relevant. The findings show that synthetic data does not represent a way to completely eliminate real data, but a way to utilize real data in the most efficient and safe way, complementing it and substituting it when real data is insufficient, too sensitive or absent.

Moreover, synthetic data is not only a "tool-technology", intended as a technology that only exerts its utility for a specific function, but it is also an "enabling technology" that makes the development of other innovations possible, and many are the examples shown in the results. Therefore, from the practical perspective of managers of international firms, synthetic data could be a resource to address business problems like compliance and inaccessible real data, as well as presenting an interesting way to approach testing, data analysis, future prediction of consumers behaviors or automations based on AI trained with it.

It seems wise for managers to promptly start evaluating whether synthetic data is going to truly become a necessity for their international business, because **in our times of extremely rapid technological developments the future was actually yesterday.**

Limitations

Extensive measures to ensure the highest methodological rigor were not possible because of single authorship. Additionally, even though the sample includes international firms operating in different industries and that are using synthetic data for different purposes, the total of ten interviewees may not be sufficient to grant sufficient generalizability. Moreover, the rapidly evolving technological scenario makes it difficult to predict for how long the results of this study will be relevant. The results also suggest that the advantages associated with the technology are far more than the drawbacks, but this may be influenced by the fact that the interviewees that are providers of this technology have interests in portraying it positively since their firm's performance is tied to success of the technology. Additionally, the sample did not include international firms that attempted to adopt synthetic data but ultimately failed, which could be a relevant aspect to assess more comprehensively the overall disadvantages of the technology and could also impact the generalizability of the results.

Future Research Suggestions

Due to the novel and dynamic nature of synthetic data and the rapidly evolving technological developments, there are several possibilities for further studies. Firstly, studies that include testing of real and synthetic data using international firms' data could offer better insight and more quantitative proof of the advantages and challenges associated with this technology. Secondly, only as regulations become more tight, the legal compliance implications of synthetic data could be tested more extensively. Additionally, future studies will be able to determine the true impact of synthetic data on labor. Further studies are necessary also to determine what is the failure rate of international firms that succumb to the initial or later challenges associated with synthetic data. These research would contribute to a more comprehensive understanding of the impact of synthetic data on international businesses.

Appendix

Table.1 Synthetic Data Applications

Interviewee 1	Synthetic sensor data
	Science and health
	Semi-con
	Defence systems
	Agriculture
Interviewee 2	Bank credit score
	Risk management
	Integrate AI in quantitative finance
	Understanding financial models
Interviewee 3	Facial anonymization
	Computer vision models
	National and international defence
	Security, surveillance and transportation for commercial sectors
	ADAS (Advanced Driver Assistance Systems)
	Wildlife monitoring and conservation
	Intelligent traffic and smart cities
Interviewee 4	Synthetic medical records
	Money laundering detection
Interviewee 5	Retail
Interviewee 6	Representation of the 17 millions individual of the dutch population
	Traffic models
Interviewee 7	Data copy and migration
	Customer Relationship Management Solution.
	Testing in nuclear and satellite industry
Interviewee 8	Testing in aeronautical and space domain
Interviewee 9	Quality assurance in insurance
	Testing in particle accelerator
Interviewee 10	Management for healthcare and pharmaceutical industry

Table.2 Definition of synthetic data

Interviewee 1	[with synthetic data] we can engineer a data set.
Interviewee 2	Artificial intelligences that we train to describe real processes and to be able to create realistic data.
Interviewee 3	Data that is synthetically made inside our engine…then we add our secret sauce, and we use that to train our models.
Interviewee 4	Data that are generated from real privacy sensitive data in order to protect privacy while also making the data available. You can also look at purely synthetic data…it's generated from scratch in the sense that you define parameters or you specify a model and you generate it from there. You can also make a distinction between if you're adding random noise to the real data, or if you're anonymizing the data, or if you're actually generating completely new data…So that's kind of a spectrum on how much those two can, yeah, how much the original data is still visible in the synthetic data.
Interviewee 5	Anything that does not directly come from real data that we have, but we create it artificially, Of course, we base it on real data...but it is created by us as a way to augment our data... It's being created rather than being collected.
Interviewee 6	It's artificially generated. It is not measured or derives from real data directly…but statistically it represents the [data].
Interviewee 7	Synthetic data is realistic data made by artificial intelligence…fake but realistic data that is mimicking the statistical characteristics of your real data.
Interviewee 8	Is data that represents the phenomenon under examination.
Interviewee 9	synthetic data comes from a model about the real data.
Interviewee 10	I don't have a precise definition…in all areas of technical and scientific nature, language sometimes becomes important…sometimes it happens that you use the same word, but then behind there are different things…We talk about testing, in the sense of the word, so in the environment of software testing.

Table.3 Generation of synthetic data

Interviewee 1	We generate 3D environments and we simulate sensors within those environments.
Interviewee 2	We have a proprietary AI that we have trained for the generation of synthetic data.
Interviewee 3	We used Unity as the base. Then you need to populate or bring as many different what we call 3D assets…anything that lives in our simulation environment. And then you need to augment that.
Interviewee 4	If you're generating it from real privacy-sensitive data, then you kind of describe the characteristics of those data that you want to be represented in your synthetic data…you'll have your original data, then you create a model of those data, and then with the model, you generate new records that follow the relations that you specified in the model. And it's also possible to not base the model on real people, but to define the kind of relations that you already know to be true, or that you know that you want to be in your synthetic data.
Interviewee 5	[We use] the most simple way of generating synthetic data, and that is we just take the existing data points and we copy them. So we don't make any changes. We don't do any combinations

	between different data points. It is literally just grabbing the small group and then taking each sampling…Just randomly, keep randomly sampling from the small group with replacement.
Interviewee 6	It is based on the population data of the real population that we got from the Central Bureau of Statistics, and you cannot directly copy that data. You have to anonymize it and aggregate it. And from the aggregated data, we have regenerated as good as possible this synthetic population. So, each member of the synthetic population does not represent a real person in the Netherlands. But statistically, it should show the same age distribution, gender distribution, and income distribution It starts with generating 17 million individual people who have no property yet. And then, according to the statistics, we assign them…And then, we do some special processing to also make sure some of the correlations are okay. For example, that a person of 5 years old does not get a driver's license. Because driver's licenses are normally only for people of 18 years old and older. And, yeah. So, there's a model that makes sure that unrealistic combinations are avoided.
Interviewee 7	You allow your team to create fake but realistic data that is mimicking the statistical characteristics of your real data…You have two ways of generating synthetic data. The first way is you connect your data, your databases, with the synthetic data factory that will go and train the model, basically, upon your real data. And once it's trained, it will create your synthetic data. It's an iterative approach…The second technique, and this is what we are using, is to go to the real data. We don't train the model on this real data. AI go[es] to the real data to detect the statistical characteristics, to detect the patterns of your data, to detect the relationship between the fields.
Interviewee 8	There are various parameters. We can say that it brings information on the state of the system that manages it and or that generated it, on the environment in which this system is operating and the temporal reference in which this data is valid.
Interviewee 9	[syntehtic data] would come from all of these beamline simulations for the accelerator and particle codes that kind of give a prediction of what particles and a particle accelerator beam should be doing in certain environments, certain electrical fields and magnetic fields.
Interviewee 10	You can divide generation for functional testing and non-functional testing, in both sectors synthetic data are used… to see if the software does what it should do.

Table.4 Methods to ensure quality and relevance of synthetic data

Data user's feedback	*Reference 1:* *(Interviewee 2)* Data is perceived as the fuel…whether it is synthetic or real. If at the end of the fair it doesn't appear that you manage to have an advantage over at least the majority of your competitors, customers clearly tell you. We do the customer's job to understand if they can actually help the customer to do his job better and then there is also the other thing, that the customer comes back to us and tells us this works, this doesn't work.
	Reference 2: *(Interviewee 3)* The value of synthetic data really is confirmed or validated by the customer All we need is one image from our customer's environment. And we're able to then retrain really, really quickly.
	Reference 3: *(Interviewee 7)* our AI is detecting everything, but it's [also about] showing to the end user, or the expert, these are parameters: "Are you happy? Click next", Would you like to fine tune something?"
Feedback loop	*Reference 4:* *(Interviewee 9)* Having the human in the loop, like a feedback system…"we predict the damages here" and be able to toggle that on and off. We wanted to add a button that was like "oh, this did a good job,

	or oh, this did a bad job".
AI assistance	*Reference 5:* *(Interviewee 2)* If you want to simulate market data with all the good properties of market data, you have to go and see if these properties actually exist. We have another AI that is trained to distinguish good data from bad data and then we have studies on the use of synthetic data for the creation of the final product of our users which are themselves trading algorithms. So we have studies where we see that the synthetic data actually helps in the development of the trading algorithm.
	Reference 6: *(Interviewee 7)* AI can detect the statistical characteristics and you decide "do I need static data? Means that I need data for the 10 years ago? Or do I need real-time data? Or do I need the data to predict for me the future?"...the more iterations the AI is doing, the best is the quality.
	Reference 7: *(Interviewee 9)* you have this idea of a train set, a validation set, and a test set. The train set trains your model. It's kind of like your model is learning all of those-..And then I'll try out…on the validation set. And then, as the prediction gets better on the validation set, then you know that you don't have to train your model anymore.
Comparing with real data results	*Reference 8:* *(Interviewee 1)* So, we don't use real data for training, but we can use it to compare our synthetic data to see in a latent space how close we get. There's some mathematics for that, to do that comparison. Of course, the biggest test is always just to generate the data, train a network, and test on real data….We evaluate using the real data, if any is available.
	Reference 9: *Interviewee 4)* see what would the outcome have been, what would the decision have been with real data versus original data. But of course you could also have a look at what the risks are that have been decreased.
	Reference 10: *(Interviewee 5)* We did some tests with, well, the old version without synthetic data. The results were pretty bad. It turns out, indeed, it just wasn't very good at predicting, indeed, what customers were going to do. And then we did the same test again after introducing the synthetic data. And, yeah, it was much more accurate in predicting what people were going to do.
	Reference 11: *(Interviewee 6)* the process of verification and validation…we compare that with the original input that we had…We compare the statistics of both tables and see if they are close enough. There's, of course, always a small error, but they must be close enough...is very much different from what we had as an input, then…quality is not so good.
	Reference 12: *(Interviewee 8)* The dirty, evil real data from outside…can infringe the behaviour of a system. Then I have to make it robust compared to external interferences.
Highest possible customization	*Reference 13:* *(Interviewee 3)* a number of our customers, what happened was we would produce the synthetic data that they needed in the exact environment that they were looking for, that they would have their AI solution operating in.
Ex-post verification	*Reference 14:* *(Interviewee 4)* Usually, you measure the quality afterwards, so I think it would be really nice if we can kind of make statements about to what extent the quality will be preserved before we generate the data, and that is something that I also research
	Reference 15: *(Interviewee 10)* when you have this huge amount of data, you don't start doing the double check, you tend to trust yourself…It happened that there were some material errors…you realised…because,

	precisely, you see strange results, sometimes the results were simply the results that you did not expect. *Reference 16:* you can go based on experience, and say "yes, these results seem plausible to me". If the results don't seem plausible to you, you start asking questions. The results are not plausible because, in fact, they are strange or the data were not good. So, in some cases, you say "let me have a look at this data" so you go and investigate a little more, you go and check a little more.

Table.5 Compatibility between real and synthetic data

Exclusively complementing real data	*Reference 1:* *(Interviewee 5)* In my case, it's more complementing the existing data because, like I mentioned, the real data was very imbalanced.
	Reference 2: *(Interviewee 2)* Our synthetic data is sold precisely as a complement to the real data.
	Reference 3: *(Interview 10)* we had a smaller database…we made arguments.
Exclusively substituting real data	*Reference 4 :* *(Interviewee 6)* For many of our applications, it can substitute the real data…We can do (calculations) based on the synthetic population, just as well as we would on real data.
	Reference 5: *(Interviewee 7)* We recommend to never connect a synthetic data factory with production data. You need the first step to profile your data. Of course, you need to provide your production data. And once you have what we call the metadata or the data model, we load it to the factory. And this is where you can play with it and create your synthetic data…all your synthetic data is based on your parameters and based on your metadata only…They create whatever data, you know, that is stored, but it's all fake at the end of the day. It is replacing completely the real data. When we talk about dev and test…synthetic data can completely overshadow or even replace the real data.
Both complementing and substituting real data	*Reference 6:* *(Interviewee 1)* And we've had projects where [synthetic] algorithms were developed on real data. *Reference 7:* We've done many projects where we trained on just 100% synthetic data. *Reference 8:* it's also not a given that there is real data. Sometimes there is no real data at all, there's just nothing.
	Reference 9: *(Interviewee 3)* There are, I would say, instances where synthetic data absolutely do compliment…we would take our synthetic data piece and just use that as a refinement tool to make the algorithm better, to perform better. *Reference 10:* There were some customers of ours…who literally had one to two images. that's it. either of the environment that their product was going to be inside…And also of the actual object that they needed to detect or classify in the environment. Right…take our synthetic data and go to a test environment, run it through with their algorithms. They obviously train their algorithms on our synthetic data…and they were able to detect the object in real time very well.

	Reference 11: *(Interviewee 4)* It's possible to use them combined, especially in the case of data augmentation. *Reference 12:* You can also use combinations of synthetic data and real data, but it's possible to use them combined *Reference 13:* Usually you create synthetic data so that you do not have to use the original data. Reference 14: Most of our use cases…you do not want to share the original data, or you prefer not to share the original data because it contains sensitive information. So in that case it doesn't make a lot of sense to also add original data to it, but more to generate synthetic data only.
	Reference 15: *(Interviewee 8)* Before entering into operation, these systems feed on data, which can be both synthetic, artificial data, and real data, taken from the real world. *Reference 16:* [compared to the real data] It manages to be certainly more efficient, more optimized in conveying all the necessary information.. it's a matter of optimization of information.
	Reference 17: (Interviewee 9) We had this set of real data that we used that used a box around the damaged area. And then we had synthetic data, which actually had like the contours of the damaged area... we got finer… damage detection with the synthetic data. *Reference 18:* Usually our model did a lot better on the synthetic test set and a little bit worse on the real data test set.

Table.6 Interaction between real and synthetic data

Interaction issues	*Reference 1:* *(Interviewee 2)* There are always interaction problems because, especially when we talk about data that is acquired from real processes, it is never as good as you expect. For example, technically we talk about drift or bias, this means that they are a little bit unbalanced…they don't give you a completely transparent view of what's going on…the issue is that many times they are small data sets, the way in which they are acquired, the way in which they get processed, they tend to lie.
	Reference 2: *(Interviewee 6)* Yeah, sometimes we do have some troubles…There is some publications about how many cars are driving around in the Netherlands…at first we found that in our synthetic population, we had not the same number of cars and it was significantly different, not just one or two, but 10 to 20% less cars…but it was because we did not count the leased cars…In the end we found a different way to also assign the leased cars.
	Reference 3: *(Interviewee 7)* Basing synthetic data on the real data, you need always to connect the factory to the production [real] data, which is problematic for many companies.
	Reference 4: *(Interviewee 9)* We had to do some, I would say very detailed computer science work under, underneath to get, to make use of both this synthetic and the real data and bring them, um, kind of bring them together.

Seamless integration	*Reference 5:* *(Interviewee 1)* So, real data sets often have some sort of bias, and there's limited tools available to address that...But from using synthetic data, we do have the tools to do that.
	Reference 6: *(Interviewee 3)* [no major issue reported]
	Reference 7: *(Interviewee 4)* I haven't encountered any issues there.
	Reference 8: *(Interviewee 5)* It's a smooth process…we just combine the real and the synthetic data together. And both of them are in the same format. It's only the specific values that are different between the samples. But they're all in the same format, so there's no issues.
	Reference 9: *(Interviewee 8)* There are no problems with interaction. It is important that there is a single, standard format that represents both real data and synthetic data. It is the same format. Then maybe there is a flag that says this is real, this is synthetic, but they must be represented in the same language. Two things. To be able to relate them, to compare them, to use them for the same reason, for the same function, they must have the same format,
	Reference 10: *(Interviewee 10)* No, in our case, there wasn't really a problem.

Table.7 Synthetic data indispensability

Codes	Quotes
Unavoidable or irreplaceable technology	*Reference 1:* *(Interviewee 1)* It's a tool that sometimes makes a project feasible. So there's a lot of technology now that can't be developed just because the data is too much of a stopping block. And this can be an enabling technology for a lot of development. Reference 2: In the end, synthetic data is also just a tool. And for some jobs, it's a tool that is very hard to ignore. You will need it.
	Reference 3: *(Interviewee 2)* Synthetic data…allows you to simplify processes that otherwise would not you would be able to do *Reference 4:* It will become a necessity…when the internet was created everyone started working with the internet, it was no longer possible not to work without the internet. When someone invented social media, everyone started marketing on social media data…when the synthetic data is truly of proven quality and safety, it will become impossible not to rely on this type of additional information.
	Reference 5: *(Interviewee 3)* It is like the next 21st century oil when it comes to AI.
	Reference 6: *(Interviewee 6)* It was not possible to use, like I said, the CBS (Central Bureau of Statistics) data. CBS has all the data, but it's not allowed to use that directly.
	Reference 7: *(Interviewee 7)* For big companies, they keep trying synthetic data…In the majority of the projects in terms of data, when they are having a lot of delays, when they know that they may fail, they need to simulate data, to do data augmentation, they don't have another option.

	Reference 8: *(Interviewee 8)* So it's something that is now out of control, because it's something that can't be stopped, because if we don't adopt it… we will succumb, because our competitor from another company, from another country, whatever it is, will adopt it. So it can't be stopped.
	Reference 9: *(Interviewee 9)* The ground truth of where the damage is in the photo…That would not be possible. I mean, we mimicked that by having experts draw on real photos.
	Reference 10: *(Interviewee 10)* It was the only thing feasible, because the data, real data, to us, it was really impossible to get there. I repeat, it was really PII, so we developers had absolutely no possibility of having access… they are very sensitive data, you can't.
Unique insight	*Reference 11:* *(Interviewee 1)* Train an algorithm on some very rare edge cases that are otherwise either very hard, expensive, or just not possible to get. Reference 12: Scenario generation, basically digital twinning, where we generate all kinds of unwanted scenarios. So, if you're doing surgery, you'd like to test how the system responds if something happens unexpectedly, and it's something you can't really test in the real world. You know, you can't damage a main artery, for instance, just to see whether the system deals with this gracefully. you can do whatever you want, because it's a bit more of a sandbox in that respect. So, you can do some more extreme scenarios, and that is appreciated by, for instance, the FDA.
	Reference 13: *(Interviewee 2)* From the point of view of risk management…it allows you to make models on risk scenarios that you do not actually have, or because you cannot use the data for questions by law or simply because you don't have them, because they have never been registered, because the application is too new, you don't have them. *Reference 14:* essentially means new scenarios that you have never seen before and which can be interesting for understanding the future.
	Reference 15: *(Interviewee 7)* In the satellite industry, to simulate a shock or some tests, you cannot do them at all….Well, with synthetic data, you can simulate a sensor that is telling you: "oh, now I have a shock, now I'm not working anymore or now the temperature is reaching whatever, etc." And you test the behaviour of your automations.
	Reference 16: *(Interviewee 8)* [you do not have to] wait for the lightning, the volcanic ash instead of extreme phenomena that we don't get.
	Reference 17: *(Interviewee 10)* you want to guarantee catastrophic events, let's say, not common, so you want to test how much your software is resilient compared to events, let's say, not very likely, but that can happen. And so, in this case, you have to generate synthetic data that is able to solicit that thing. *Reference 18:* it's the only thing to do, the miracle is, there is no other way.

Table 8. data enhancement through synthesis

Improved data analysis	*Reference 1:* *(Interviewee 2)* Synthetic data available allows you to simplify processes, especially when the synthetic data comes through a platform like the one we created which has different layers of analysis…having better results.

	Reference 2: When we started developing the first test cases we realized that the approach worked better than we hoped, which was very surprising because this stemmed from the fact that the artificial intelligence could understand these very complex and somewhat chaotic processes of financial markets better than we expected.
	Reference 3: *(Interviewee 3)* We use it to improve the performance of our models…you're also going to provide more robust, more reliable solutions because you're able to throw at that algorithm many different environments, many different instances… a lot more different scenarios than you can when you capture real data. *Reference 4:* we used to take pre-trained models on real data…we would take our synthetic data piece and just use that as a refinement tool to make the algorithm better, to perform better…You can take a foundational model trained on real data and use synthetic to just make it perform better in certain instances *Reference 5:* The unexpected surprise was customers' [reaction] to...reducing their bottom line, generating more sales.
	Reference 6: *(Interviewee 4)* you could have access to the synthetic data but to [access] the original data, that would take very long, much longer processes or would be much more difficult. *Reference 7:* If you're able to do many more analyses or use AI more efficiently or more effectively because you're using synthetic data, then that could be a competitive advantage. Or if you're able to gain, gather more insights maybe from external parties because you're using synthetic data, that could also be an advantage.
	Reference 8: *(Interviewee 5)* we could make the quality of the data as like the data set as a whole, we could make it better by introducing these synthetic samples. *Reference 9:* It just makes it easier to collect the data and it allows you to make better use of the data that you do have. Like in our case, our predictions are a lot better because of the synthetic data that we include.
	Reference 10: *(Interviewee 6)* We use it in our models so we can improve our models and make them more accurate. Now, having more accurate models gives also more accurate outcomes. And then in the end, it is also used for decision making, like investments in infrastructure and this sort of thing. *Reference 11:* instead of an abstract number that you say, there are 500 people living here, and 10% of them, I don't know, get stuck in traffic jams. But now we can say, agent 12 and agent 13 and agent 15, they get stuck in traffic jams. *Reference 12:* We can, for the time being, distinguish ourselves from other people who do traffic modeling, because we do it on this level of detail. Yeah, it can certainly be a competitive advantage.
	Reference 13: *(Interviewee 7)* [Synthetic data] could speed up the process of development of what we call time to market. When you have an idea, it could speed up the process of making your products and speed up the development lifecycle.
	Reference 14: *(Interviewee 8)* Efficiency, because the fact of being a synthetic data means that you have all and only the necessary information…Performance, because things are done in due time…the goal is to optimize traffic

	management…Safety, for people who must communicate, manage to communicate without errors and this can be done with the standardization of information, so a synthetic data has a standard format, understandable by everyone, everyone within that data manages to take their portion of information to perform their task properly. *Reference 15*: Making the results of a more rigorous analysis more robust *Reference 16:* From a technical point of view…decision-making system for the management of air traffic behaves better, has a better efficiency than the flight controller, maintaining the appropriate distances between the planes, avoiding collisions…It performs better.
	Reference 17: *(Interviewee 9)* We didn't have to take a picture of a car, drive it to a whole bunch of different locations, take a whole bunch of different pictures with different cameras and different resolutions. *Reference 18:* The end result was to speed up the process of estimating how much our company should pay for the damages. *Reference 19:* It allows us to use really large models that require a lot of data. The larger your model is, the more parameters you need to train, the more data you need. And so, having synthetic data allows you to use very fancy heavyweight models.
	Reference 20: *(Interviewee 10)* it's faster, sometimes, you do the same thing, but, if you do it quickly, then, you can do other things at the same time.
Data augmentation	*Reference 21:* *(Interviewee 1)* Have run into the wall of not having the right data for the job.
	Reference 22: *(Interviewee 3)* We call it model drift…with synthetic data, you can create enough of [data], so that…you're not going to have any false positives or false detections, which doesn't sound like a big deal. But when that camera is installed at an intersection, and it is being used to for traffic control actuation for vulnerable road users, like at a pedestrian, like an old an elderly person, or a little kid or someone on a bicycle, that is huge…because that could be life or death. *Reference 23:* We have populated our simulation engine with so many different assets…in all different angles…The list was endless. Add as much variety as you can to expose your neural networks, the better they will perform…We're able to generate enough imagery to be able to retrain, to refine and improve the performance of that solution. And we get it out there faster than anybody else.
	Reference 24: *(Interviewee 4)* Synthetic data are also increasingly being used to augment the data. So if you don't have enough data, you add more to it, or to compensate for imbalance in the data set. If you have minorities that are underrepresented, and you're using AI to make predictions, then they may also have lower quality predictions by the AI, so you can supplement your data there as well.
	Reference 25: *(Interviewee 7)* from the little satellite they have, the test satellite, we detect all the parameters, all statistical characteristics, and we start simulating hundreds or thousands of fake satellites, basically. And this is how they can simulate or can verify that the new software will be working with the real satellites

	Reference 26: *(Interviewee 8)* It represents real data perfectly, indeed it extends it, because this process, starting from the sensible world, let's say, and extrapolating its essence, it's ontology…of the sensible world, and then synthesising it into a data that is all-encompassing.
	Reference 27: *(Interviewee 9)* The model that we were using at the time had so many parameters that it required a lot of data…we were able to…create fake images of cars in different environments and in different lighting and different colours. And then additionally, we were able to impose damage onto them…in order to feed into the engine and generate these different photos from different perspectives and angles and different lens aberrations, so there was just a lot more variety of data augmentation that we could do with the synthetic setup Reference 28: It saved time labelling and we got a huge amount of data augmentation and variation. Reference 29: We just got a much larger variety of data than we would be able to collect just using our clients' damaged photos.
	Reference 30: *(Interviewee 10)* You have to imagine situations, you don't often have historical data…so you have to test new functionalities, of which you don't have data, and so you have to invent them a bit, obviously they are educated guesses…so you have to make an imaginative effort…Making synthetic data often means creating new data. Obviously, with characteristics, statistics…that copy [real data], or sometimes, instead…they were made ad hoc to solicit specific things, but always on the basis of [real data].
Predicting future data	*Reference 31:* *(Interviewee 2)* We work in…quantitative finance, therefore people who in turn use models that consume data to make predictions on what to buy,what to sell, and when…
	Reference 32: *(Interviewee 5)* When predicting future customer behaviour, well, before we were using synthetic data the predictions were pretty bad…now that we have included the synthetic data, it is going a lot better. So in terms of business impact, of course, knowing what your customers are going to do in the future is incredibly important for all aspects of business...it has definitely allowed us to gain a lot of better insights into what our customers are going to do.
	Reference 33: *(Interviewee 7)* Future is basically predictions…you detect basically the patterns of your data…so [synthetic data] will try to mimic the statistical characteristics of 5 or 10 years ago. So whatever you give as inputs to the AI, the more you give, the best is the output, basically.
	Reference 34: *(Interviewee 8)* Drawing up the synthetic data, you do it when you don't know the phenomenon. You don't know how, for example,...they buy them when there is a drop in inflation or they don't…you slowly draw it up.You say: "I understood from what I have collected that in the months that go from Christmas to Easter people buy more this instead of this thing", and you draw up the rules.
	Reference 35: *(Interviewee 9)* We would have…modelling mixtures and transfer learning techniques, to try to get a model that could predict real data.
	Reference 36: *(Interviewee 10)* We also dealt with warehouse data, of the pharmacies, so… we built the engine that… brought to the store the drugs, in such a way that the customer came in and said: "I need an aspirin" and we didn't said "no, I don't have it, come tomorrow"... you don't want to…keep stuff in the warehouse for too long, but, obviously, you don't want to…lose the customer, no, you lose money, and so, you need the historical data,

	we had two years of historical data to make the prediction…[with synthetic data] you wanted to create new data, in a way that you could then inject into your databases. *Reference 37:* It gives you the possibility to do a whole series of tests if you are smart enough…to make predictions, to do educated guesses. You make a series of hypotheses, as verisimilar as possible, on the characteristics of the data.

Table.9 Cost reduction

Pre-performance testing (shift-left)	*Reference 1:* *(Interviewee 1)* You can generate defective or data of defective items before production line is operational
	Reference 2: *(Interviewee 2)* You want to create a robot that assembles two parts of a vehicle…There are many companies that created virtual environments where it is no longer necessary to have the physical robot but there is a physical model of the robot that lives in this simulation environment and learns to do that task, once it has learned…the robot is then translated directly into the real environment. This cuts costs impressively. *Reference 3:* You want to change your production process and you say how much it costs me to understand if I can change the production process in this way. On the one hand you would have to produce the robot, train it, put it on the assembly line, observe it for a month before understanding, and maybe this will cost you 100 million. If instead you do it in this synthetic, completely virtual environment, maybe it costs you a million. Those are numbers…I don't think they are that far from reality.
	Reference 4: *(Interviewee 7)* Some companies can have, for example, one or two small satellites. So, they have some data, but they would like a new IT system to manage the legacy satellites and new satellites. And they need to test this application with the volume of the data expected for all…either they assume it's working, and they go live, and they discover the issues after that, or they do what we call data augmentation. *Reference 5:* In the nuclear fields, for example, to have to simulate some tests, it may cost them billions. *Reference 6:* And it could also enable companies to shift-left…You allow the developers to test while they are developing, so they discover the bugs and the issues in time. If they discover them late in the process, it will cost a lot to the company because you need to go back to the initial case.
	Reference 7: *(Interviewee 8)* I'm making an airplane, I'm making a space platform, sending it in flight costs me an eye. So...I generate data with a simulator…from the simulated data I try to say "this simulation is very close to reality because I did it with the same means, with the same conditions it encounters, everything" and it turns out that in 10,000 hours of flight this gets three incidents…it's crazy. So, here you don't do tests because tests cost an eye and you do it with synthetic data. *Reference 8:* We have the ice gallery…you do it in the gallery, but before doing it in the gallery you do it by software, simulated. Why so? Because doing it in the gallery we spend a lot of money every time we do the tests. *Reference 9:* if you introduce changes in the treatment of employees…[with synthetic data] you see in a fairly congruent period of time the effects on the productivity of this change. You can measure how it is.

	Reference 10: *(Interviewee 10)* in the pharmaceutical industry, so…you have 10,000 stores that, at the same time, access that software…you have to test that your software respond to 10,000 people at the same time…it is even more important that you have a volume of data that is comparable to those that you expect that your software should… so, you need 10,000 fake users…with a geographic distribution that is similar…so, you have to have a series of data that respect all these characteristics. *Reference 11:* If you discover an error at the end, it is a disaster, because you have to do everything over again…There are companies that have failed because they failed the performance tests… there was a company that… had a website and this page literally takes 20 seconds to open, because you didn't size it by the number of tests, you don't go there anymore…even if it technically works, as a user, you say "I'm not going there anymore because it doesn't work". So you lose business, you get a terrible name…So the synthetic side allows you to do these tests as soon as possible. it's a monstrous advantage, monstrous. Because, I repeat, you know, in modern software there are precisely these Shift-Left principles, which are not trivial. *Reference 12:* the principle of "fail big, fail soon"... all these principles of agile development, of being fast. If you have something that blocks you, that happens at the end, you can't be fast, by definition. You have to discover the bad things as soon as possible…So, in my opinion, there is the advantage, there is the real advantage.
Reduced need for skilled labor	*Reference 13:* *(Interviewee 1)* we're now talking about a medical project and there's synthetic data. The alternative would be to hire a lab to work on cadavers, human cadavers, to create data and that would involve 40 people for one year of work…that's quite a budget.
	Reference 14: *(Interviewee 2)* [Synthetic data] frees them from having to hire ten experts to be able to set something up, and we do this through synthetic data so, let's say, there is artificial intelligence that understands a synthetic given process produced for those who don't have the skills to fully understand that process. *Reference 15:* Allows you to simplify processes that otherwise…would require you to hire specialized personnel or would take a long time…So cut operating costs.
	Reference 16: *(Interviewee 9)* [synthetic data] prevented a whole bunch of human hand labelling of damaged sections of photos. So that was a huge thing. When we were doing the transfer learning and different techniques to combine, to bridge the gap between synthetic and real data, we did hire a bunch of people to circle damage. And it was…a business problem… some of the insurance adjusters would draw a very fine outline that tightly hugged all of the damage. Some of them would like to draw little dots all over for the damage. Some would just circle the whole thing. So finding a metric… was hard. *Reference 17:* It was less expensive because we didn't have to pay humans to make the special training data sets all the time.
Less expensive to generate data	*Reference 18:* *(Interviewee 2)* By 2040, if I'm not mistaken, 70 percent of the data will be synthetic and I have no difficulty believing it because data is expensive anyway.
	Reference 19: *(Interviewee 3)* It's orders of magnitude cheaper…we're not going to charge you a crazy amount of money, we're not going to charge you an arm and a leg, it's not going to cost you 1000s of dollars, it shouldn't, there's no reason. *Reference 20:* we're really good at being competitive when it comes to costs and price because we can use synthetic data.

Table.10 Synthetic data advantages for legal compliance

Advantages	Quotes
Using and Sharing sensitive data	*Reference 1:* *(interviewee 1)* With a company that does traffic monitoring. And there's quite a bit of regulation in place that makes it hard to use recordings of cameras in the public space. So, also there, you may want to create synthetic data. *Reference 2:* if you want to see if people are walking on the railway tracks. It's not something that happens every day... but you probably won't have a ton of data. And the data you have, there's a good chance you can't use it. So, then there's a ton of different examples…from a regulatory perspective, this also gives some flexibility. *Reference 3:* Any compliance towards privacy or personal data, I think that's always a very key one.
	Reference 4: *(interviewee 2)* This is something between trying to free yourself a little from the evolution of the regulations which will always be more prohibitive and on the other hand to use new information technologies especially AI *Reference 5:* The first uses that were made, synthetic data, were probably already ten years ago in the field of compliance, and they increased considerably with the approval of the European law GDPR…it is a matter of replicating information that cannot be used because it is too private
	Reference 6: *(interviewee 3)* you have much less of a worry of cybersecurity and being hacked…if your models are being developed by synthetic data…the underlying foundation of that information is coming from essentially an artificial place.
	Reference 7: *(interviewee 4)* Make more data available…different parties have information about the same people or information about the same constructs, then it might be easier to combine it if you do it in a private way. *Reference 8:* Making the decisions transparent, because you could share more information about synthetic data than you would about the real data. *Reference 9:* Share a data set…you don't want to share the real data set. *Reference 10:* A hospital that…doesn't want to give the software engineer the real patient data, so they want something that's more private. *Reference 11:* Collaboration, or working with maybe providers of AI or data science, and also while protecting privacy. Transparency is also one of the goals sometimes, so that you can show what the data looks like. *Reference 12:* The Dutch ministry for education…use it for legal compliance, so they would not be allowed to

	share the real data, but they share the synthetic data instead. *Reference 13:* A lot of software testing is now done with real data and they know that they can do it in a more private way
	Reference 14: *(interviewee 7)* Our biggest clients do have a big Indian team. And due to GDPR, they cannot share the real data with developers. *Reference 15:* [synthetic data] reduces the risk of data leaks….eight out of 10 data leaks are due to human errors. And the number one factor of the data leaks is data sharing…because you need to develop a new feature, you clone your whole database. And once you clone it and you share it with your developers, they start storing it in Google Drive, in Dropbox, and so on and so forth. *Reference 16:* it's very difficult to say…to a big company to change the whole process from sharing the real data…When you base your synthetic data on metadata…you share only the description of your data. You never share the data, which is totally fake and therefore totally safe and completely, compliant with any regulation across the group.
	Reference 17: *(interviewee 8)* In our context, make anonymous data beyond the competition because if you say… that competitor A has a defect…if the stock markets fall, a mess happens. *Reference 18:* To exchange information, not to read the competition, the competitiveness of a company, not to expose vulnerabilities…So the synthesis of the data is certainly useful to protect any vulnerability, it can be privacy, it can be another vulnerability *Reference 19:* other things that we also do, support decisions for the choice of mode of transport…lower CO2 emissions... You could start by saying, ok, let's gather people a little…and you gather all these data…obviously you gather sensitive data, sensitive data that you can not manage, because you don't need them to understand how things are moving. So your synthetic data is pure personal information of people, so much so that you only care about knowing the average age, how much they move…you purify it of personal information, of privacy.
	Reference 20: *(interviewee 9)* It's much easier to get data without violating copyright law. So, that's a huge advantage right there.
	Reference 21: *(interviewee 10)* In functional testing, you want to see if the software does what it should do, and so you need a series of data that allows you to perform this test. The problem is that in many circumstances, in almost all the circumstances that I know, you can't use the real data...it was a software that had to manage sensitive data of people… the people who do these tests can't have access to real data *Reference 22:* At the same time, you have a double requirement, one is to protect and not to use real data. And you have the requirement to have to use data that is plausible, because the functionality of the software changes depending on the data that you use, so you want the software to be exposed to data that is as plausible as possible. *Reference 23:* PII (personal identifiable information) they are protected by the federal law, I mean, you're in trouble if you have access to these data…we didn't have the possibility to look at them.

Only viable solution	*Reference 24:* *(interviewee 6)* It was totally impossible without it…it would have been impossible for us to use the real data.
	Reference 25: *(interviewee 7)* It's the only approach. It's the only approach that works. *Reference 26:* Companies, they need to share data, and they know that they cannot. So, the old techniques are not working anymore…Data anonymization or data masking, it used to work before. But now, in the context of big data, it's very heavy.
	Reference 27: *(interviewee 8)* synthetic data in a correct, useful, effective, useful and not risky format, not exposing to vulnerability, is certainly the solution.
	Reference 28: *(interviewee 10)* It was the only feasible thing, I mean, there are no other possibilities…alternatively you had to have very high levels of NDA that were not compatible, you know, with the number of employees.
Data retention	*Reference 29:* *(interviewee 4)* you have to delete the data periodically, and you want to retain a synthetic version, so that you can still look back.
Ethical approaches to data use	*Reference 30:* *(interviewee 4)* You can also use it from more of an ethical motivation, say I'm already allowed to do this with the real data, but I could answer the same question with synthetic data.
Use sensitive data in real time	*Reference 31:* *(interviewee 7)* And also, in terms of real-time applications, [real data] does not fit with the requirements. So, the only solution is to allow developers to create the data they want, when they want, and they load it to the applications they want.

Table.11 Synthetic data disadvantages for legal compliance

Disadvantages	Quotes
Underlying presence of real data	*Reference 1:* *(interviewee 2)* The legislator essentially has the absolute power to prohibit you from using any type of data and when we talk about synthetic data we are talking about data that is synthetic up to a certain point because in any case you have to know something, otherwise you cannot produce any type of data synthetic.
	Reference 3: *(interviewee 4)* Although synthetic data can definitely also protect privacy and differs in the sense that it's not real data anymore. So anonymized data is still real data *Reference 4:* we talk about the privacy utility trade-off, where the utility is the quality of the data…you want your data to be similar enough that it's, that you can use it and that you can answer questions with it that you would have wanted to answer with the original data, but not so similar that it starts leaking information about the original individual
	Reference 5: *(interviewee 6)* if the group is smaller than 10, you are not allowed to use the exact number…small number of people satisfying a certain criteria…it would be almost possible to determine which people those are

	Reference 6: *(interviewee 7)* The approach, which…you go to the real data and then you create fake data out of it…they do techniques like data shuffling…you're still playing with the real data.
Uncertainty of the legal scenario	*Reference 7:* *(interviewee 1)* There's always some interpretation of the guidelines and laws concerning synthetic data.
	Reference 8: *(interviewee 4)* we need to make sure that legally…[synthetic data] can be made interpretable for legal experts to kind of have a say about whether [protection] is sufficient. *Reference 9:* the biggest issue at the moment because it is difficult to say when synthetic data really are sufficiently private. *Reference 10:* How to measure privacy is probably the biggest challenge…because privacy is, of course, it's a human construct. It's a legal construct. And even if you can measure it, it is sometimes difficult to make the measurements interpretable. And there's sometimes a trade-off between to what extent you can measure it very quickly for large groups of people and to what extent you can actually interpret it.
"Opting-out"	*Reference 11:* *(interviewee 4)* If somebody wants to opt out of the original data, they should also no longer be represented in the synthetic data…if you really want to completely be removed from the data set of an organization, then you shouldn't still be kind of half represented in the synthetic data.

Table.12 Applicability and effectiveness issues

Code	Quotes
Model not sufficiently complex	*Reference 1:* *(Interviewee 1)* The development process is an iterative one. So, in the first iterations we typically encounter some cases that are not covered well by our synthetic data sets, and then we expand the data sets to include more examples
	Reference 2: *(Interviewee 4)* The model was not complex enough to capture everything that was going on in the original data. So then you're limited in what you can do with the synthetic data.
	Reference 3: *(Interviewee 6)* There are some properties of the population that we did not take into account in our synthetic population [because] in the real dataset, this [needed info] was not listed…Another example is that we assume households to be like consisting of just a few person, like a family household, but there are also other types of households…An example would be a prison or an institute. Also disabled people or handicapped people…We did not make a separate category for that kind of people.
	Reference 4: *(Interviewee 8)* Must be sufficiently exhaustive to then represent reality, and reality always remains something very, very varied, complex, and even unpredictable.
	Reference 5: *(Interviewee 9)* Because of the limitations of the model performance, we moved to just having it be a recommender system, like a helper for human work, rather than automating the process of giving in.
	Reference 6: *(Interviewee 10)* Not to characterize them sufficiently, because then, if you were doing the software design, you had to characterize the test design. You had to characterize the data as it served you statistically. There is the

	delicate part.
Domain expertise	*Reference 7:* *(Interviewee 1)* We've been asked many times to generate data where we also beforehand say "We can't possibly do this, We don't have the domain knowledge and also that knowledge is not available in the world". *Reference 8:* Say you want to have the correlation between some medication and some specific values. We can't do that through synthetic data because we can't make a realistic simulation of the human body in all its complexity. Nobody can… you can only train that if you gather that data yourself in the real world. So you need some domain knowledge to do so. So there's definitely limitations. *Reference 9:* We will need to quite well understand the domain we're working in. So that can be a challenge.
	Reference 10: *(Interviewee 7)* we give it back to the expert to review…Because, you know, the use case specification is known by the human and not by AI. I believe that we cannot ask the AI to do the job and trust that the AI will do the comparison and tell us that my output is having a high quality. No, you cannot do that. So we put the human in the middle of the process. So instead of training the model and saving the model and asking the model to create the data for us and tell us the quality of the data I'm generating. We are putting all our eggs in the same basket. I do believe that AI, while it's very powerful today, is not still not major. It's not 100 percent major. We still need maybe five to 10 year Reference 11: To be realistic, to be humble, we are saying that AI will do a job, we save a model, and we give that model to the expert to validate. They need to validate. We cannot just produce sometimes.
	Reference 12: *(Interviewee 9)* The humans that were doing that job… they're former mechanics. So, they know when they see car damage "oh, this component underneath is definitely damaged". But in the synthetic data, and even from novices looking at these images, we don't know that insight.
	Reference 13: *(Interviewee 10)* My job is in this sector here. Is saying, "Look, I need this data, like this" It depends on your responsibility…Because you have to say, look, with this kind of data I solicit the things I'm interested in soliciting. *Reference 14:* I had to ask for it, it's there, there's my component, the human component. That's the risk. it's not because [the data] gave it to you wrong for a material form, it's because you asked for the wrong thing, that's the risk.
Initial research	*Reference 15:* *(Interviewee 2)* The architecture of this AI model…it was the biggest challenge, it was more than a year of research.
	Reference 16: *(Interviewee 3)* there would be instances where solutions wouldn't work and you have to go back to the drawing board and it was a very iterative process between data scientists and our customer side. Reference 17: It's essentially still a black box. We don't really understand. We don't, we can't say with a hundred percent certainty or confidence why it detects A and it doesn't detect B in the same environment.
	Reference 18: *(Interviewee 6)* it was a little bit harder than we thought to get it right…So, our first idea was that it was rather straightforward to generate a synthetic population, but it turned out to be more difficult than we thought.

	Reference 19: *(Interviewee 9)* Some stuff about the lighting and shadows that, you know, wasn't perfect…that was a lot of work to learn that engine in order to create the synthetic data. *Reference 20:* a lot of customer photos were taken in parking lots, and the painted lines give a certain, like, reflection off of the curve of the car. And so, we realized that we needed to simulate cars in parking lots…people taking photos inside garages, Suddenly, you need a model of the inside of a garage…common environments people take real photos in, you know, is in sunny days underneath trees…previously, our 3D model just did, like, cloudy, diffuse light days, not these, like, bright sunlight, direct rays. So, eventually, when we were able to get that, it turned out better.
Overselling	*Reference 21:* *(Interviewee 7)* A lot of players of AI, not only synthetic data, AI in the market, they keep overselling AI. It's like AI will do coffee for you. Yeah, it will kind of do…AI can do a lot, but we still need the human in the middle. *Reference 22:* The biggest challenge for data, not only for synthetic data…is understanding your data, knowing what you want. Too many companies…they've been excited about synthetic data. And…they don't have a use case or they are not able to know.
	Reference 23: *(Interviewee 9)* the executive leaders are like, "Oh, AI, we need to include AI". And, and it's like, no, nobody that was just in a car accident wants to chat with the chat bot, like, get real.
	Reference 24: *(Interviewee 10)* There is this buzzword, you know, of synthetic data, of artificial intelligence, that, it's a way to talk about statistics in a very fancy way.
Updates with real data	*Reference 25:* *(Interviewee 8)* they must continue to be fed by reality, there is nothing to do, because otherwise we risk fossilizing and entering situations that are not representative of reality, because reality evolves, it brings us to the unforeseen. So, if I think of taking a photograph of reality, I transform this photograph into synthetic data, then at a certain point I need to update this photograph after a while, because otherwise that data is no longer representative of reality, because the external scenario evolve**s.**
	Reference 26: *(Interviewee 9)* the 3d models of cars that were for sale, they're all like… fancy cars, but most of our customers, they have economical cars and those were harder to find…. So that made it a little bit tricky. You have this shift through time, cars 10 years ago don't look like cars we have now. So model drift data…is an issue…they would have to regenerate the synthetic data to train the model again.
Computing power	*Reference 27:* *(Interviewee 5)* A bit of a problem was the computing power that we needed to be able to handle all of the big volume of data. Because, yeah, with the synthetic data, in addition to our real data, we had quite a large sample. And that caused some strains on our computing power.
	Reference 28: *(Interviewee 8)* There can be limits in the hardware, in the machine, in the storage capacity, that is, in the memorization of data or in the computation.
	Reference 29: *(Interviewee 9)* It took a lot of computing to run and generate all of the images.

Table.13 Approaches to bias

Less biased than real data	*Reference 1:* *(Interviewee 1)* The algorithms trained on just synthetic data performed better than because we have the tools to generate much more diversity in data.
	Reference 2: *(Interviewee 3)* There were other instances with other customers who had tons of real data, but it was biased because the majority of the real data happened only [a few times]. We had one defense company who was in the maritime industry. And the majority of their image datasets were of sunny days. And it's like, you need a little bit of different variety in the weather, and different variety also of the ocean, the rolling waves, for example, not just calm waves. so there, in that instance, they were able to combine our synthetic with their real.
Bias is needed	*Reference 3:* *(Interviewee 2)* Our system was created to be able to highlight these biases and keep them under control. What clients can do is actually decrease or increase certain types of bias to see what happens. When we talk about bias, we are not talking about bias with respect to a segment of the population but we are talking about bias with respect to the impact of certain processes. A fairly intuitive example is that the markets themselves have a level of nervousness and you see this in the data…it is interesting to see what happens if you inject nervousness into the data or remove it.
	Reference 4 *(Interviewee 5)* we're also not trying to eliminate bias necessarily. This use case was about predicting customer behavior and, of course, customers are people, and people are inherently biased. So to a certain extent, it is fine that there is bias in the data. In this specific application…it is also okay to have a certain level of bias.
	Reference 5: *(Interviewee 7)* Sometimes you need not only quality data. You will need dirty data. When you are working in data quality projects, for example data cleansing; when you have a big database and you would like to create codes to go and to correct your data, you will need to enrich the data. *Reference 6:* we need to create, to mimic the reality in terms of data completeness, data correctness and so on and so forth *Reference 7:* Human[s] were always biased, by the way. So, when you need data, you need data that's mimicking your reality. *Reference 8:* when you create your synthetic data, it should have some outliers as well. It should be realistic. Realistic doesn't mean clean. It means it is respecting the statistical characteristics of your data... You cannot say: “okay, I need fair data”: You need to have realistic data. It's not only fair or quality data.
	Reference 9: *(Interviewee 10)* when I went to see the algorithm with which they made these purchase predictions… there is definitely a bias, but it's an informed bias…you, with your bias, lead the market…it never occurred to me the idea that a bias should be corrected.
Effort to correct bias	*Reference 10:* *(Interviewee 1)* [in case of fully synthetic data] we have to make some real considerations…because how is there any bias in what we generate ourselves? Is there anything that could skew the results? That same engineering also should take place with the real data. So if there are biases introduced, we did that ourselves. We were not given that bias by the measurements. So it gives some extra tooling and extra responsibility.

	Reference 11: *(Interviewee 3)* That comes down to the expertise of our data scientists or expertise of any data scientist. There will always be biases. It's a matter of how to reduce them to a comfortable level where they don't get in the way.
	Reference 12: *(Interviewee 4)* The first step is to be aware of the bias in your original data, because that's what's going to cause the bias in your synthetic data as well. And I think it's probably an algorithm dependent on how much, whether or not your bias will be amplified. *Reference 13:* if you're using more statistical approaches, I think the probability is lower that your bias will be amplified…if you're using AI, because that's even why when we use data augmentation. We use that specifically because AI doesn't represent every group equally.
	Reference 14: *(Interviewee 5)* I did a lot of investigating to make sure that this was a sound approach. The main reason for that was that I was afraid that we would be amplifying, indeed, the biases that were present in the [real] data, especially because we're literally just copying them over.
	Reference 15: *(Interviewee 6)* Beforehand, you don't always know which attributes are the most significant ones. We have assumed that these ones are probably relevant, but we don't know for sure whether they are. For example, we now do have some discussion in our department whether we should include the migration background. Like I said, it could maybe amplify some bias…the choice you make on what properties of your synthetic data are relevant…it is on our radar that it is a concern.
	Reference 16: *(Interviewee 8)* by design, it must guarantee the construction and generation of synthetic data not affected by bias…so the choice of the source, the verification that there are no preconditions that can generate bias.
	Reference 17: *(Interviewee 9)* we assessed by having test sets that were held out that were just synthetic and just real. So, that's how we evaluated if there was bias…you need to make sure that all of the variables in your synthetic environment are reflective of the real data set.

Table.15 Initial reservation

Proof of effectiveness	*Reference 1:* *(Interviewee 1)* There's definitely some convincing required…for new markets, almost always, we have to create some examples ourselves, just to prove we can do this. It's also very dependent on the market itself. So, for instance, in Semicon, there are more larger companies that are used to taking some risk on this new type of technology. So, it's a bit easier to get in as a proof of concept or to prove yourself within an initial project, within a research project. For medical projects, that is more difficult…there we need a real, very, very solid track record and prove how well this works.
	Reference 2: *(Interviewee 3)* I think that is [obtaining] market validation.
	Reference 3: *(Interviewee 4)* People are more sceptical about the quality. You also see it the other way around, that people are overly confident about the quality and about the technology.

	Reference 4: *(Interviewee 9)* it was hard to communicate with our business partners that "Hey, the model can only do as good as the humans do"... there was a lot of pushback… it's a challenge to get a business to adopt this stuff… they wanted a magical result. And then it's like, "your human adjusters only agree with each other 75% of the time." I think that was a really shocking statistic for them.
Synthetic data is not authentic data	*Reference 5:* *(Interviewee 4)* There are definitely reservations…some of the reservations we've heard, for example, from the medical community is that it's not real data.
	Reference 6: *(Interviewee 5)* It feels kind of like you're just hacking, just coming up with your own data. But of course, that's not entirely what's happening because we are basing it off of existing data. And we're not just creating our own data in a way that we think might be best.
	Reference 7: *(Interviewee 7)* When we go to a pharmacy and we buy vitamin C, for example, if it is synthetic or if it is natural, you will go to the natural one, you know, even if it is more expensive…Resistance to change is what's always here.
	Reference 8: *(Interviewee 8)* Real data is always more, how to say, complete and also unpredictable than a synthetic one. The synthetic one must tend to encompass this nature of unpredictability.
Security and privacy	*Reference 9:* *(Interviewee 2)* Confidentiality is certainly the main obstacle. There is a way to overcome this obstacle which is clearly a matter of cyber security.
	Reference 10: *(Interviewee 4)* The privacy part is a big one, so kind of the legal side, also the ethical side, but yeah, how can we really know that it's not leaking anything?
No reservation	*Reference 11:* *(Interviewee 6)* I immediately thought it was a very good idea. And it is also not completely new. It is also used by other research programs.
	Reference 12: *(Interviewee 10)* In this case it was the only thing feasible.

Table.16 Ongoing research efforts

Codes	Quotes
Enabling new technologies	*Reference 1:* *(Interviewee 1)* In the end, synthetic data is a tool. It's not an end goal. It's a tool to enable development of other algorithms down the pipe…Reinforcement learning is typically something that is really greatly helped by simulation synthetic data *Reference 2:* Convolutional graph networks…synthetic data can really help to train these models.
	Reference 3: *(Interviewee 2)* it's a field where those who have a little bit of intuition about how technology evolves, especially as it has over the last ten years, he understands that there is an opportunity
	Reference 4: *(Interviewee 3)* detecting litter in the seas and the oceans and cleaning that in a more efficient, affordable way…cancer detection…climate change predictions

	Reference 5: We're also known for not just producing AI solutions, but deploying them, basically having them run on very low power, power-constrained devices. And that's really hard.
	Reference 6: *(Interviewee 4)* I can definitely also see it's growing in terms of what the technology can do. So better technologies, technologies for data types that don't exist yet. *Reference 7:* Something that we're looking into is relational data…many other topics as well. I do have a colleague that looks more into the use of generative AI for synthetic text.
	Reference 8: *(Interviewee 5)* more advanced way to generate new data points… I would want to implement that…And, of course, doing a bit of research, like if I can explore this further, doing a bit of research into what new developments there are in synthetic data.
	Reference 9: *(Interviewee 6)* We are looking into machine learning methods or AI methods to do the analysis. Also, for the generation of the synthetic population
	Reference 10: *(Interviewee 8)* This is the goal. So, technological developments, through computer science, because this allows us to be more efficient, to see many facets of things together, faster, in a more optimized way, to take into account many links, because we are faced with complex problems,...synthetic data and the use of data that synthesizes all these links, these goals, and therefore manages them more easily *Reference 11:* You can synthesize a lot of data in a very short time, in the device where you memorize synthetic data. In perspective, you will do it not on the memory cell…you will probably do it with quantum mechanics…which will make the other revolution.
	Reference 12: *(Interviewee 9)* I think you can start playing with these multimodal models like question answers, combining language, video. Yeah, I think. And then the fact that you have the ground truth with it is just, yeah, it's a lot of human effort to make a multimodal data set that has ground truth. Because the human would need a label like so many things about it. But if you make synthetic data, you get it for free, the labels. *Reference 13:* The climate around synthetic data for accelerator particles, particle accelerators is a little bit different. The simulations are highly regarded. And I would say the both operators of accelerator of particle accelerators and kind of more computational lists.They are both extremely interested in using simulation to have to improve operations, like in real time of the particle accelerator
Wider applicability	*Reference 14:* *(Interviewee 1)* I think there's definitely a lot of potential.
	Reference 15: *(Interviewee 2)* [with] simulation environments…you can think of any type of task as a simulation environment and therefore this will probably create hundreds if not thousands of products
	Reference 16: *(Interviewee 4)* there's a lot of needs on the market to apply the technology. There's also a lot of motivation.
	Reference 17: *(Interviewee 5)* I think that there are a lot of applications in which synthetic data can have added benefits,

	both from a business perspective and also from an ethical societal perspective.
	Reference 18: *(Interviewee 6)* We should reap the full benefits of our approach and make sure that it is used in the modeling, and maybe a little more than it is now, because it is actually a good idea and a good approach to do our modeling using this synthetic data.
	Reference 19: *(Interviewee 7)* We innovate with real use cases. We innovate with customers. We innovate while working in real challenges…So, it depends on the appetite of the companies to synthetic…the more use cases we have, the more features or enhancements we can put on our factories. *Reference 20:* We see that even with our clients, they are keen to understand more and more the technology…It's also on the agenda of many companies. Clients, not only willing to use the technology, but also…would like to get their hands even dirty. They would like to develop and participate in some use cases, which is good.
	Reference 21: *(Interviewee 8)* If the aeronautical domain, as you can see today, is fully engaging in [these] applications… even though it is so safety critical, it means that we are really witnessing a revolution. *Reference 22:* If the driver of synthetic data generation, if the drivers are inclusiveness, environmental sustainability, but in general, then surely synthetic data could be a tool that guarantees, over time, sustainability, equity, inclusiveness and so on.
	Reference 23: *(Interviewee 10)* It's true, yes, someone is studying new technologies to do new things.

Table.17 Focus on improving reliability

Increasing security	*Reference 1:* *(Interviewee 3)* if it's running on an edge device, a pure edge device, meaning it's going to be running on a chip and all the information is stored on that chip. And it's that information isn't being sent anywhere. Then you have much less of a worry of cybersecurity and being hacked.
	Reference 2: *(Interviewee 5)* We might not need to store as much data ourselves, which is, of course, also better for security and privacy reasons that you don't store all customers' data directly.
	Reference 3: *(Interviewee 8)* the possibility of synthesizing in electronic impulses, tomorrow quantum, because quantum is already being done, an enormous amount of information can certainly improve the management of many processes, of many phenomena, improve from the point of view of security,
	Reference 4: *(Interviewee 9)* Our data came from audio transcripts of telephone calls. So then you have a transcriber or human transcriber, and then you have PII identifying inter, information that you had a screen out. But if you, if they're able to just generate synthetic data, they don't have to worry about protecting the customer.
Increasing quality	*Reference 5:* *(Interviewee 4)* I would hope that there's more applications, that those applications are smoother so that it's easier for people to use synthetic data…You require less expertise, which also comes with insightful measurements of privacy and data quality.

	Reference 6: *(Interviewee 5)* As we get more and more accurate ways of generating synthetic data, it will make just the entire process of collecting data much easier. Because you might not need to collect thousands of responses from customers if you can use 100 and then create a representative sample using synthetic data…So that'll obviously make a big difference in just survey costs.
	Reference 7: *(Interviewee 8)* from the point of view of time used, because time always remains the greatest good, the synthetic data allows you to reduce time, if we want to conclude the speech. It allows you to reduce time because it synthesizes the information, everything and only what you need to do things right, it manages to keep it in very little space of memory, of computing capacity.

Table.18 Approaches to the 13 UN recommendations on the use of synthetic data

Codes	Quotes
Unaware	*Reference 1:* *(Interviewee 1)* I haven't really read into this yet.
	Reference 2: *(Interviewee 6)* This is news to me.
	Reference 3: *(Interviewee 9)* I'm not familiar with them.
	Reference 4: *(Interviewee 10)* No, absolutely don't know them.
Useful as generic guidelines	*Reference 5:* *(Interviewee 2)* Those 13 points are absolutely reasonable, nothing to object to. For the moment, from my point of view, as often happens, they are also very obscure to implement, they are a bit philosophical if you like, there is no clear roadmap on what companies need to do to be able to do so. *Reference 6:* The impression that we in the sector have that for the moment the legislator is rightly concerned with the interests of citizens without wanting to put too many spokes in the wheels in a technology that of AI which is perhaps the only opportunity to have that production leap at an economy of scale level that we have been waiting for for probably 20 years and which can in turn be perhaps the only solution to many problems such as that of costs of social safety nets, social security, the aging of the population, the lack of specialized personnel.
	Reference 7: *(Interviewee 5)* We will keep the guidelines in mind…like good principles…a good set of guidelines for how you should go about generating your data…common sense around how you use your data.
	Reference 8: *(Interviewee 8)* There is a global movement of thought, I would say, on the need to discipline the information content, the process of generating synthetic data. There is a need felt on a global level…And these rules are all acceptable…So they are absolutely valid indications… They are not prescriptive, there is no indication of saying you have to do this and that. They are more best practices at the moment. But they are fine, they are fine. We will continue to work on this, there will be room for lawyers, philosophers, psychologists.
Indifferent	*Reference 9:* *(Interviewee 3)* Unless our customers come to us and say, look, the UN said A, B, and C, and I want to make

	sure that that's in there, no problem. Then we will make sure that our solutions are adhering to whatever regulation is necessary.
	Reference 10: *(Interviewee 4)* it's not something that's very much on the radar for us at the moment.
	Reference 11: *(Interviewee 7)* [the UN] are still not aware about what is even synthetic data... the recommendation from the United Nations is well, you know, to help in argumentation.

Questionnaire:

Preliminary Understanding

- How do you define synthetic data?
- What are the synthetic data solutions that your firm uses?
- What motivated you to adopt synthetic data?
- What goals do you aim to achieve with synthetic data?

Technical Aspects

- How does the process of generating synthetic data work in your firm?
- How do you ensure the quality and relevance of synthetic data?

Adoption and Integration

- Did you have initial reservations regarding the adoption of synthetic data? If so, how were they overcome?
- Do you have practices to ensure the integration of synthetic data with the traditional real data? (How do the two types of data interact with each other?; Have there been any issues, and how have they been addressed?)
- Can you give examples of cases in which synthetic data well complements the real data or substitutes it, partially or entirely.

Legal Compliance

- Does your firm use synthetic data to help compliance with regulations? Why or why not?
- Was compliance with regulations among the reasons considered when deciding to adopt synthetic data? Why or why not?
- How did the adoption of synthetic data impact compliance with data protection and privacy regulations for your firm?
- Have your firm encountered any challenges while using synthetic data as a means to comply with regulations? If so, how were they addressed?

- Looking forward, will the use of synthetic data impact positively or negatively any legal considerations?

Strategic and Managerial Insights

- What is the role that synthetic data plays in your operations and decision-making?
- How do you assess the value added by synthetic data compared to traditional data?
- Do insights derived from synthetic data influence strategic decisions for you?
- Have there been any unexpected benefits or challenges associated with synthetic data for you?

Challenges and Limitations

- What are the most significant challenges you have faced when generating synthetic data?
- Have your firm encountered specific limitations that have impacted synthetic data applicability or effectiveness? If so, how were they overcome?
- How does your firm address the potential risks of synthetic data, like bias amplification (the risk that existing bias in the real dataset gets magnified in the synthetic dataset)?

Future Perspectives

- How do you see the role of synthetic data evolving in the future? (What future goals do you think synthetic data can help achieve?)
- Does synthetic data represent or will represent a source of competitive advantage compared to firms that haven't adopted it?
- Are there any new techniques for synthetic data generation and analysis that your firm is exploring or plans to explore?

UN recommendations

- The UN released 13 recommendations on the use of synthetic data. Does your firm find them helpful for your operations and/or for your clients? Why or why not?

References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In Handbook of labor economics (Vol. 4, pp. 1043-1171). Elsevier.

Adobor, H., Darbi, W. P. K., & Damoah, O. B. O. (2021). Strategy in the era of "swans": the role of strategic leadership under uncertainty and unpredictability. Journal of Strategy and Management.

Aggarwal, C. C. (2011). An introduction to social network data analytics (pp. 1-15). Springer US.

Ahmed, J., & Shaikh, Z. A. (2011, July). Privacy issues in social networking platforms: comparative study of facebook developers platform and opensocial. In International Conference on Computer Networks and Information Technology (pp. 179-183). IEEE.

Andres, G. (2021) What Is Synthetic Data?. NVIDIA Blogs https://blogs.nvidia.com/blog/what-is-synthetic-data/

Andrews, G. (2021, June 8). What is synthetic data? NVIDIA Blog. https://blogs.nvidia.com/blog/what-is-synthetic-data/

Argote, L., Beckman, S. L., & Epple, D. (1990). The persistence and transfer of learning in industrial settings. Management science, 36(2), 140-154.

Athmaja, S., Hanumanthappa, M., & Kavitha, V. (2017, March). A survey of machine learning algorithms for big data analytics. In 2017 International conference on innovations in information, embedded and communication systems (ICIIECS) (pp. 1-4). IEEE.

Awate, S., Larsen, M. M., & Mudambi, R. (2015). Accessing vs sourcing knowledge: A comparative study of R&D internationalization between emerging and advanced economy firms. Journal of International Business Studies, 46, 63-86.

Backstrom, L., Dwork, C., & Kleinberg, J. (2007, May). Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th international conference on World Wide Web (pp. 181-190).

Barbier, G., & Liu, H. (2011). Data mining in social media. Social network data analytics, 327-352.)

Barney, J. (1991). Firm resources and sustained competitive advantage. Journal of management, 17(1), 99-120.

Bathla, D. (2021). Data analytics the way forward. Chitkara University. Retrieved from https://www.chitkara.edu.in/blogs/data-analytics-the-way-forward/

Beales, H. (2010). The value of behavioral targeting. Network Advertising Initiative, 1(2010).

Bronstein, E., Srinivasan, S., Paul, S., Sinha, A., O'Kelly, M., Nikdel, P., & Whiteson, S. (2022). Embedding Synthetic Off-Policy Experience for Autonomous Driving via Zero-Shot Curricula. Waymo. Retrieved from https://waymo.com/research/embedding-synthetic-off-policy-experience-for-autonomous-driving-via-zero

California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100-.199 (2018)

Cassell, C., Buehring, A., Symon, G., & Johnson, P. (2005). Qualitative management research: A thematic analysis of interviews with stakeholders in the field. Available at https://eprints.ncrm.ac.uk/id/eprint/4905/1/final_research_report.pdf

Chen, Y., Hua, X., & Maskus, K. E. (2021). International protection of consumer data. Journal of International Economics, 132, 103517. https://doi.org/10.1016/j.jinteco.2021.103517

Chung, W., & Yeaple, S. (2008). International knowledge sourcing: Evidence from US firms expanding abroad. Strategic Management Journal, 29(11), 1207-1224.

Clarke, V., & Braun, V. (2017). Thematic analysis. The Journal of Positive Psychology, 12(3), 297-298.

Coche, E., Kolk, A., & Ocelík, V. (2024). Unravelling cross-country regulatory intricacies of data governance: the relevance of legal insights for digitalization and international business. Journal of International Business Policy, 7(1), 112-127.

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. Journal of Information Technology, 30, 44-57.

Côrte-Real, N., Ruivo, P., & Oliveira, T. (2020). Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value?. Information & Management, 57(1), 103141.

De Wilde, P., Arora, P., Buarque, F., Chin, Y., Thinyane, M., Stinckwich, S., Fournier-Tombs, E., & Marwala, T. (2024). Recommendations on the use of synthetic data to train AI models. United Nations University.

Di Minin, E., Fink, C., Hiippala, T., & Tenkanen, H. (2019). A framework for investigating illegal wildlife trade on social media with machine learning. Conservation Biology, 33(1), 210.

Di Minin, E., Fink, C., Tenkanen, H., & Hiippala, T. (2018). Machine learning for tracking illegal wildlife trade on social media. Nature ecology & evolution, 2(3), 406-407.

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. Frontiers in Environmental Science, 3, 63.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American sociological review, 48(2), 147-160.

Dobocan, G. (2020, February 15). Tech fails: BBC's £100M digital media blunder. Medium. Retrived form https://medium.com/north-code/tech-fails-bbcs-100m-digital-media-blunder-ca397bb99ecd

Domingo-Ferrer, J., & Muralidhar, K. (2016). New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. Information Sciences, 337, 11-24.

Domingo-Ferrer, J., & Saygin, Y. (2008). Privacy in statistical databases. Berlin Heidelberg: Springer.

Donaldson, L. (2001). The contingency theory of organizations. Sage.

Dow, R. (2023) AIB shipments climb in Q2 2023, with unit sales increasing from quarter to quarter – Jon Peddie Research https://www.jonpeddie.com/news/aib-shipments-climb-in-q2-2023-with-unit-sales-increasing-from-quarter-to-quarter/

Drechsler, J., Bender, S., & Rässler, S. (2007). Comparing Fully and Partially Synthetic Data Sets for Statistical Disclosure Control in The German IAB Establishment Panel: supporting paper für die Work Session on Data Confdentiality 2007 in Manchester. EUNECE/Programmes.

Drechsler, J. (2011). Synthetic datasets for statistical disclosure control: theory and implementation (Vol. 201). Springer Science & Business Media.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), eaao5580.

Dummy file. (n.d.) McGraw-Hill Dictionary of Scientific & Technical Terms, 6E. (2003). Retrieved March 6 2024 from https://encyclopedia2.thefreedictionary.com/dummy+file

Emam, K. E., Mosquera, L., & Hoptroff, R. (2020). Practical synthetic data generation: Balancing Privacy and the Broad Availability of Data. O'Reilly Media.

Engström, V., Johnson, P., Lagerström, R., Ringdahl, E., & Wällstedt, M. (2023). Automated Security Assessments of Amazon Web Services Environments. ACM Transactions on Privacy and Security, 26(2), 1-31.

Estrada-Jiménez, J., Parra-Arnau, J., Rodríguez-Hoyos, A., & Forné, J. (2019). On the regulation of personal data distribution in online advertising platforms. Engineering Applications of Artificial Intelligence, 82, 13-29.

European Parliament and Council of the European Union. (2016). 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). Official Journal of the European Union

European Union Artificial Intelligence Act (provisional agreement as of December 2023).

Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. Journal of economic perspectives, 23(3), 37-60.

Fernhaber, S. A., Mcdougall‐Covin, P. P., & Shepherd, D. A. (2009). International entrepreneurship: leveraging internal and external knowledge sources. Strategic Entrepreneurship Journal, 3(4), 297-320.

Freeman, R. E. (2010). Strategic management: A stakeholder approach. Cambridge university press

Gao, T., Li, F., Chen, Y., & Zou, X. (2018). Local differential privately anonymizing online social networks under hrg-based model. IEEE Transactions on Computational Social Systems, 5(4), 1009-1020.

Garrett A. Johnson, Scott K. Shriver, Samuel G. Goldberg (2023) Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR. Management Science 69(10):5695-5721. https://doi.org/10.1287/mnsc.2023.4709

Ghasemaghaei, M., Hassanein, K., & Turel, O. (2017). Increasing firm agility through the use of data analytics: The role of fit. Decision Support Systems, 101, 95-105.

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digital Medicine, 6(1), 186.

Graef, I., Wahyuningtyas, S. Y., & Valcke, P. (2015). Assessing data access issues in online platforms. Telecommunications policy, 39(5), 375-387.

Hu, J., Reiter, J. P., & Wang, Q. (2014, September). Disclosure risk evaluation for fully synthetic categorical data. In International conference on privacy in statistical databases (pp. 185-199). Cham: Springer International Publishing.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P. P. (2012). Statistical Disclosure Control. John Wiley & Sons.

Ilieva, R. T., & McPhearson, T. (2018). Social-media data for urban sustainability. Nature Sustainability, 1(10), 553-565.

J.P. Morgan. (n.d.). Synthetic data. Retrieved [2024, March 10], from https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data

James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: exploring use cases to optimise data utility. Discover Artificial Intelligence, 1(1), 15.

Kedia, B., Gaffney, N., & Clampit, J. (2012). EMNEs and knowledge-seeking FDI. Management International Review, 52, 155-173.

Knight, F. H. (1921). Risk, uncertainty and profit (Vol. 31). Houghton Mifflin.

Kuemmerle, W. (2002). Home base and knowledge management in international ventures. Journal of Business venturing, 17(2), 99-122.

Kurapati, Shalini, & Gilli, Luca. (2023). Synthetic Data: Convergence between Innovation and GDPR. Journal of Open Access to Law, 11, 1-12.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International journal of information management, 34(3),

Leontief, W. (1952). Machines and man. Scientific American, 187(3), 150-164.

Levitt, T. (1965). Exploit the product life cycle (Vol. 43). Cambridge, MA, USA: Graduate School of Business Administration, Harvard University.

Liben-Nowell, D., & Kleinberg, J. (2003, November). The link prediction problem for social networks. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 556-559).

Lieberman, M. B. (1989). The learning curve, technology barriers to entry, and competitive survival in the chemical processing industries. Strategic Management Journal, 10(5), 431-447.

Liesch, P. W., Welch, L. S., & Buckley, P. J. (2011). Risk and uncertainty in internationalisation and international entrepreneurship studies: Review and conceptual development. Management International Review, 51, 851-873.

Liu, F., Xiao, B., Lim, E. T., & Tan, C. W. (2017). The art of appeal in electronic commerce: Understanding the impact of product and website quality on online purchases. Internet Research, 27(4), 752-771.

Liu, K., & Terzi, E. (2008, June). Towards identity anonymization on graphs. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 93-106).

Liu, X., & Yang, X. (2011, September). A generalization based approach for anonymizing weighted social network graphs. In International Conference on Web-Age Information Management (pp. 118-130). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lucini, F. (2021, October 20). The Real Deal About Synthetic Data. MIT Sloan Management Review. Retrieved from https://sloanreview.mit.edu/article/the-real-deal-about-synthetic-data/

Lucini, F. (2021, October 29). Synthetic data for speed, security and scale. Accenture. https://www.accenture.com/us-en/insights/artificial-intelligence/synthetic-data-speed-security-scale

Malone, T. W., Rus, D., & Laubacher, R. (2020, December 17). Artificial intelligence and the future of work. MIT Task Force on the Work of the Future. https://workofthefuture-taskforce.mit.edu/research-post/artificial-intelligence-and-the-future-of-work/

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., (2011). Big Data: the Next Frontier for Innovation, Competition and Productivity. McKinsey Global Institute

Mendelevitch, O., Lesh, M. D., & Prince, C. (2021, October 27). Synthetic data: Unlocking data to drive healthcare innovation. NIH HEAL Initiative. Retrieved from https://heal.nih.gov/events/2021-10/reach-seminar

Moore, G. (2012, August 13). Without big data analytics, companies are blind and deaf, wandering out onto the web like deer [Tweet @geoffreyamoore]. Twitter. https://twitter.com/geoffreyamoore/status/234839087566163968

Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.

Nikolenko, S. I. (2021). Synthetic data for deep learning (Vol. 174). Springer Nature. https://arxiv.org/pdf/1909.11512.pdf

OECD (2019), Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, OECD Publishing, Paris, https://doi.org/10.1787/276aaca8-en.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016, October). The synthetic data vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399-410). IEEE.

Poudevigne-Durance, T., Jones, O. D., & Qin, Y. (2022). MaWGAN: A generative adversarial network to create synthetic data from datasets with missing data. Electronics, 11(6), 837.

Raghunathan, T. E. (2021). Synthetic data. Annual review of statistics and its application, 8, 129-140.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. Journal of Official Statistics, 18(4), 531.

Reiter, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. Journal of the Royal Statistical Society Series A: Statistics in Society, 168(1), 185-205.)

Richards, N. M., & King, J. H. (2014). Big data ethics. Wake Forest Law Review, 49(2), 393-432.

Roundtable of G7 Data Protection and Privacy Authorities. (2022, September 8). Communiqué: Promoting Data Free Flow with Trust and knowledge sharing about the prospects for International Data Spaces. Office of the Privacy Commissioner of Canada.

Rubin, D. B. (1993). Statistical disclosure limitation. Journal of official Statistics, 9(2), 461-468.

Ruiz, N., Muralidhar, K., & Domingo-Ferrer, J. (2018). On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective. In Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings (pp. 59-74). Springer International Publishing.

Sallier, K. (2020). Toward More User-centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis. Statistical Journal of the IAOS, vol. 36, no. 4, pp. 1059-1066.

Samiee, S. (1984). Transnational data flow constraints: A new challenge for multinational corporations. Journal of International Business Studies, 15, 141-150.

Savvas, A. (2011, April 6). BBC's uncompeted handover of IT deal to Siemens led to failure – MPs. Computerworld. Retrived from https://www.computerworld.com/article/1676893/bbc-s-uncompeted-handover-of-it-deal-to-siemens-led-to-failure-mps.html

Schumpeter, J. A. (1994). Capitalism, socialism and democracy (2nd ed.). Routledge. (Original work published 1942).

Schwab, K. (2017). The Fourth Industrial Revolution. Penguin UK

Simon, H. A. (1990). Bounded rationality. Utility and probability, 15-18.

Sims, C. A. (2003). Implications of rational inattention. Journal of Monetary Economics, 50(3), 665-690. https://doi.org/10.1016/S0304-3932(03)00029-1

Skinner, C. (2009). Statistical disclosure control for survey data. In Handbook of statistics (Vol. 29, pp. 381-396). Elsevier.

Slifka, J. (2019, October 11). Tools for generating synthetic data helped bootstrap Alexa's new-language releases. Amazon Science. https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases

Soria-Comas, J., & Domingo-Ferrer, J. (2017, August). A non-parametric model for accurate and provably private synthetic data sets. In Proceedings of the 12th International Conference on Availability, Reliability and Security (pp. 1-10).

Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data–anonymization groundhog day. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 1451-1468).

Straits Research report. (2023) Synthetic Data Generation Market Size & Share projection till 2031 https://straitsresearch.com/report/synthetic-data-generation-market (Report code: SRTE54781DR)

Syntegra. (n.d.). Syntegra partnering with National Institutes of Health (NIH) and the Bill and Melinda Gates Foundation to democratize access to the largest set of COVID-19 patient records. Retrieved (2024, March 10), from https://www.syntegra.io/news/syntegra-partnering-with-national-institutes-of-health-nih-and-t

he-bill-and-melinda-gates-foundation-to-democratize-access-to-the-largest-set-of-covid-19-patient-records

Tang, Y., Zhang, Y., & Ning, X. (2023). Uncertainty in the platform market: the information asymmetry perspective. Computers in Human Behavior, 148, 107918.

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic management journal, 18(7), 509-533.

Tiwald, P., Ebert, A., & Soukup, D. (2020, May 8). Diving deep into fair synthetic data generation (Fairness Series Part 5). MOSTLY AI. https://mostly.ai/blog/diving-deep-into-fair-synthetic-data-generation-fairness-series-part-5

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., ... & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. Biological Conservation, 233, 298-315.

United Nations Economic Commission for Europe. (2022). Synthetic data for official statistics: A starter guide. Geneva.

Vaddadi, S. A., Thatikonda, R., Padthe, A., & Arnepalli, P. R. R. (2023). Shift left testing paradigm process implementation for quality of software based on fuzzy. Soft Computing, 1-13.

van de Waerdt, P. J. (2020). Information asymmetries: recognizing the limits of the GDPR on the data-driven market. Computer Law & Security Review, 38, 105436.

Voss, W. Gregory. (2020). Cross-border data flows, the gdpr, and data governance. Washington International Law Journal, 29(3), 485-532.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How ‘big data’can make big impact: Findings from a systematic review and a longitudinal case study. International journal of production economics, 165, 234-246.

Wang, C., Satuluri, V., & Parthasarathy, S. (2007, October). Local probabilistic models for link prediction. In Seventh IEEE international conference on data mining (ICDM 2007) (pp. 322-331). IEEE.

Yang, K. H. (2022). Selling consumer data for profit: Optimal market-segmentation design and its consequences. American Economic Review, 112(4), 1364-1393.

Ying, X., Pan, K., Wu, X., & Guo, L. (2009, June). Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing. In Proceedings of the 3rd workshop on social network mining and analysis (pp. 1-10).

Yuan, S., Wang, J., & Zhao, X. (2013, August). Real-time bidding for online advertising: measurement and analysis. In Proceedings of the seventh international workshop on data mining for online advertising (pp. 1-8).

Zaheer, S., & Mosakowski, E. (1997). The dynamics of the liability of foreignness: A global study of survival in financial services. Strategic management journal, 18(6), 439-463.

Zheleva, E., & Getoor, L. (2009, April). To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In Proceedings of the 18th international conference on World wide web (pp. 531-540).

Zheleva, E., & Getoor, L. (2011). Privacy in social networks: A survey. Social network data analytics, 277-306.

Zhou, B., & Zou, T. (2023). Competing for recommendations: The strategic impact of personalized product recommendations in online marketplaces. Marketing Science, 42(2), 360-376.

Zuo, C., Lin, Z., & Zhang, Y. (2019, May). Why does your data leak? uncovering the data leakage in cloud from mobile apps. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 1296-1310). IEEE.