FINANCIAL CONDUCT AUTHORITY

Feedback Statement

FS23/1

Synthetic Data Call for Input Feedback Statement

February 2023

Contents

	Foreword	3
1.	Introduction	5
2.	Data access and innovation	10
3.	Assessment of synthetic data	15
4.	Synthetic data use cases	22
5.	Role of the regulator	27
6.	Next steps	31

Annex 1
List of questions in the Call for Input 33

Annex 2
Glossary of terms used in this document 35

Annex 3
List of non-confidential responses to the Call for Input 37

Moving around this document

Use your browser's bookmarks and tools to navigate.

To **search** on a PC use Ctrl+F or Command+F on MACs.

Sign up for our **news and publications alerts**

See all our latest press releases, consultations and speeches.

Request an alternative format

Please complete this form if you require this content in an alternative format.

Foreword

Technological innovation has the ability to disrupt existing structures at speed and scale. The cycle of innovation has resulted in the proliferation of data, which in turn has become a key driving force for financial innovation.

Data has the power to propel advancements in the use of artificial intelligence (AI) and machine learning (ML), which could unlock significant value in financial markets and lead to better outcomes for consumers, firms and the wider economy. However, in order to protect consumer privacy, it is important that data sharing occurs under certain conditions and with an appropriate legal basis.

One of the most significant questions for innovators is therefore how to access quality data to drive the development of novel products and services whilst respecting consumers' right to privacy.

Privacy Enhancing Technologies (PETs), including synthetic data, is one promising avenue developed in recent years to address this question. For the past five years, we have explored synthetic data through various innovation initiatives and have monitored the increased adoption of this technology in the broader market. Our research to date indicates that synthetic data can potentially make a significant contribution to beneficial innovation in UK financial markets by expanding opportunities for data access and sharing.

This Feedback Statement sits within our broader synthetic data and PETs work programme and provides a response to the feedback we received to our Call for Input published in March 2022. We are committed to working with industry and to encouraging broad-based discussion with stakeholders on the challenges to innovation in financial services. We would like to thank all of the organisations for their feedback.

Based on respondents' feedback, we have identified several key themes to refine our thinking on next steps, and to ensure that we continue to be at the forefront of innovation in financial services. Respondents unanimously agreed that data is crucial for innovation, however there are challenges to accessing and sharing data in financial services. Although respondents indicated that data protection regulation places specific conditions on the data they can share and access, they also reiterated the importance of consumer privacy, and that data access should have privacy built in by design at every stage of the process.

Feedback also strongly indicated fraud and anti-money laundering as a key use case for synthetic data, in part due to its ability to augment rare patterns of behaviour in a dataset. We are interested in exploring this use case further, building our own internal capabilities and working together with industry to employ synthetic data as a novel regulatory and compliance tool.

Given our position to convene industry, academia and the broader regulatory community around common initiatives, respondents indicated that the regulator could perform an intermediary role in the provision of synthetic data. Feedback also strongly

indicated the need for guidelines, standards and/or governance frameworks to build trust in synthetic data and encourage wider adoption.

Building trust is a crucial milestone in the development of an emerging technology. We will continue to work with industry, academia, regulators, and other stakeholders to ensure that where this technology is adopted, this is done so in a responsible manner and in the interest of consumers. We outline our future plans for engagement in more detail in the final section of this Feedback Statement.

We hope that this Feedback Statement contributes to a continued open dialogue between the public and private sectors to promote innovation in the interests of markets, firms and consumers.

Jessica Rusu

Chief Data, Information, and Intelligence Officer (CDIIO),
Financial Conduct Authority

Chapter 1

Introduction

Why are we issuing this paper

1.1 Evidence gathered from our Digital Sandbox pilots has demonstrated the challenges of accessing and sharing data in financial services, particularly for new market entrants. For the past five years, we have explored the potential for synthetic data to expand data sharing opportunities in a privacy compliant manner; through our TechSprints and Digital Sandbox, through internal projects, and through engagement with industry and academia.

1.2 In March 2022, we published a Call for Input to further our understanding of the market maturity of synthetic data within financial services, and its potential to expand data sharing between firms, regulators and other public bodies.

1.3 The Call for Input sought feedback on the broader challenges in accessing high quality data for innovation in financial services. We also wanted to gather specific feedback on synthetic data, including its advantages and limitations in resolving the data sharing challenge, and the use cases where synthetic data can provide substantial benefits for innovation. Finally, we wanted to gather views on the role of the regulator in the provision of synthetic data.

1.4 The Call for Input forms part of a broader FCA work programme on synthetic data and Privacy Enhancing Technologies (PETs). The aim of this work programme is to explore how technology can enable data sharing in financial services in a privacy compliant manner. By taking steps to address challenges with data sharing in this industry, our ambition is to enhance digital markets and promote responsible innovation in the interest of consumers.

1.5 In this Feedback Statement we:

- Summarise the feedback we received from the Call for Input
- Set out our response to the feedback received
- Explain our next steps

1.6 This Feedback Statement will be of interest to:

- Academics
- FCA regulated firms
- Start-ups, RegTechs and FinTechs
- Technology and data firms
- Regulators and policy-making bodies
- Consumer groups

Context

1.7 Data is increasingly driving innovation in financial services. Advances in data, analytics and AI could unlock significant value in financial services and beyond, including automation of decision-making that serves the interests of consumers and markets, algorithmic trading, and the prevention of financial crime.

1.8 Since 2014, 25% of firms accepted into the FCA's Regulatory Sandbox have had advanced analytics and data analytics at the core of their business proposition. We established Innovation Pathways in 2019, with 32% of accepted firms leveraging advanced analytics and data analytics in their products. Common use cases include automated advice (particularly in wealth advisory) and back-end process automation for services such as IT service desks and employee payroll management.

1.9 Accurate and effective AI models and systems require large volumes of high-quality data for training, validation, deployment and evaluation. To protect consumer privacy, financial data (and other forms of personal data) are subject to data protection laws that place conditions on data sharing. The subsequent challenges associated with accessing financial data, specifically for new market entrants, can inhibit the development of new products and services in the market and potentially slow down beneficial innovation in financial services.

1.10 Synthetic data is a privacy preserving technique that could expand opportunities for data sharing by generating statistically realistic, but 'artificial' data, that is readily accessible. The field shows significant promise in enabling privacy-compliant data sharing where real data is too sensitive to share, as well as for generating the high-volume and quality data needed to test and train AI models. In particular, synthetic data is useful for augmenting rare patterns and 'edge cases' that are scarce in real datasets, to better train models to respond to infrequent events. Gartner predicts that by 2030, the use of synthetic data will overshadow that of real data in AI model development.

1.11 Since 2018, the FCA has explored the potential use of synthetic data through our TechSprint and Digital Sandbox programmes. We have sought to accelerate the validation and testing of participants' products and solutions by making synthetic data available to participating firms and individuals. The Call for Input built upon our previous efforts by gathering broad input on the challenge of data sharing in financial services. We also wanted to understand industry views on the potential of synthetic data, including its benefits, limitations, and key use cases.

1.12 The Call for Input sits within a suite of recent initiatives to engage industry on technology-related topics. The FCA has recently published the following Discussion Papers:

- DP22/3: Operational resilience: critical third parties to the UK financial sector
- DP22/4: Artificial Intelligence
- DP22/5: The potential competition impacts of Big Tech entry and expansion in retail financial services

1.13 These publications signal our desire to shape digital markets to achieve good outcomes through proactive engagement with industry, academia and other public sector organisations.

1.14 In the broader context, Governments, regulators and industry (both in the UK and globally) are prioritising initiatives to embrace data sharing, to drive experimentation and new growth. In recent years, the Financial Action Task Force has conducted extensive engagement with the public and private sectors to examine emerging technologies that enable collaborative data analytics between financial institutions, whilst respecting data privacy. A core pillar of the UK government's National Data Strategy is to encourage 'better coordination, access to and sharing of data of appropriate quality between organisations in the public, private, and third sectors'. In doing so, the strategy seeks to position the UK as the forerunner of the next wave of innovation.

1.15 Internationally, regulators, private firms and third sector organisations are collaborating to expand data sharing in the interest of innovation. In July, the UK and US governments launched a series of prize challenges in PETs to encourage innovation to tackle financial crime and public health emergencies. The FCA is an observer in these challenges to gather the latest insights on how innovators are utilising these technologies.

1.16 In 2021, the Monetary Authority of Singapore (MAS) announced the creation of a digital platform (and supporting regulatory framework) to permit financial data sharing for risk discovery and collaboration. Participating financial institutions can share information on customers and transactions to prevent financial crime, including money laundering and terrorism financing. In Estonia, the 'AML Bridge' supports private-private information sharing through an end-to-end encrypted messaging platform.

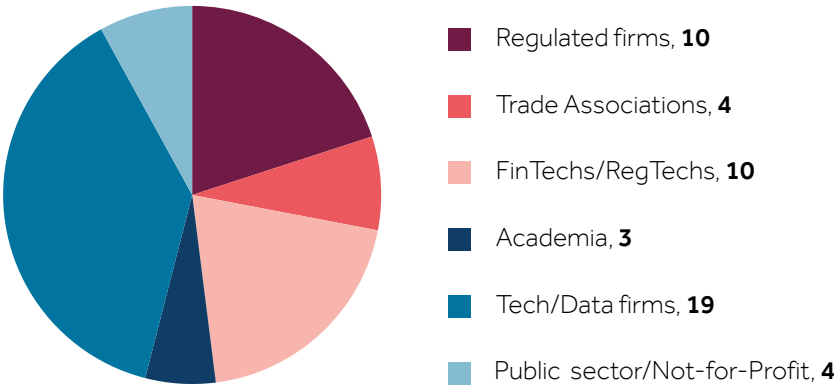
1.17 These global initiatives to enhance data sharing indicate the potential of data-driven solutions to push the boundaries of data analytics and innovation, whilst respecting privacy. We hope that this Feedback Statement will provide a view of current practices in the field of synthetic data and indicate where the most promising use cases and applications are.

Overview of responses

1.18 We received 50 responses from a wide range of organisations including regulated firms, trade associations, technology and data firms, FinTechs and startups, and academia. Whilst the Call for Input predominantly focused on financial services, we received responses from organisations operating in other sectors, including health and telecommunications. Some respondents provided feedback on all questions, whereas others gave detailed responses on specific questions. Several respondents provided more general feedback on synthetic data and data sharing. The breakdown of responses by organisation type is below.

1.19 We would like to thank all of the organisations for their feedback. The views expressed have helped to refine our thinking on next steps and future areas of focus.

Responses by organisation type



'Regulated firms' refer to firms registered on the Financial Services Register.

1.20 The structure of this Feedback Statement follows the themes and structure of the Call for Input:

- **Chapter 2: Data access and innovation,** where we assess the challenges of accessing and sharing data in financial services
- **Chapter 3: Assessment of synthetic data,** where we explore the benefits, limitations and generation techniques for synthetic data
- **Chapter 4: Synthetic data use cases,** where we outline our findings on the main synthetic data use cases, and the requirements to meet these use cases
- **Chapter 5: The role of the regulator,** where we explore the role of the regulator with regards to synthetic data

Next steps

1.21 The FCA has a primary objective of promoting competition in financial markets in the interest of consumers. In our three-year strategy outlined in 2022, we committed to becoming a more innovative, assertive and adaptive regulator. This includes: 'shaping digital markets to achieve good outcomes', 'preparing financial services for the future', and 'reducing and preventing financial crime'.

1.22 Based on the feedback to the Call for Input and previous research, our current position is that synthetic data can potentially make a significant contribution to beneficial innovation in UK financial markets. We believe further research (specifically use case identification and understanding its utility as a regulatory tool) is required before the benefits of this technology can be fully realised.

1.23 We have and will continue to engage with domestic and international regulators, industry, and academia to explore synthetic data and data sharing further. Since the Call for Input closed, we have engaged with the Information Commissioner's Office (ICO) to discuss the regulatory questions highlighted by respondents. In addition, we

hosted two roundtables (domestic and international) with regulators and government bodies in December 2022 and January 2023. The purpose of the roundtables was to communicate the Call for Input findings and assess potential areas of overlap to drive greater collaboration and efficiencies in public sector approaches to data sharing.

1.24 Several respondents to the Call for Input referenced the challenge of validating synthetic data from both a privacy and utility perspective. In response, we will host a joint industry-academic roundtable in early 2023 in partnership with the Alan Turing Institute and the ICO to understand this challenge further and discuss the various methods to validate synthetic data. We will publish a paper in the coming months outlining our key findings and next steps.

1.25 We will continue to explore potential partnerships to address key use cases in the future and leverage the Digital Sandbox and our other firm-facing services to engage with industry and academia. Our previous work through the TechSprint and Digital Sandbox initiatives, the responses to the Call for Input and subsequent engagement have increased our understanding of the potential value of synthetic data. We recognise the opportunities of initiatives to expand data sharing – including the National Data Strategy and Open Finance initiative – and we will continue to explore how synthetic data can enable innovation in financial services.

1.26 In addition, we are establishing a Synthetic Data Expert Group to create an effective framework for collaboration across industry, regulators, academia and wider civil society on issues related to synthetic data. This group will explore key issues in theory and in practice with the use of synthetic data in UK financial markets and identify best practices for adoption. It will also provide a sounding board on specific FCA synthetic data projects, for example our upcoming project to utilise synthetic data to test the effectiveness of transaction monitoring systems in identifying money laundering.

1.27 Applications to join the group will open in February, and we will hold the first session in the spring.

Chapter 2

Data access and innovation

2.1 In the Call for Input, we stated that data can unlock significant beneficial innovation in financial services, specifically through the development of AI models. Whilst the development of this technology in financial services is still modest, it has huge potential across:

- Financial crime and fraud prevention
- Credit scoring
- ESG performance and reporting
- Sales and trading
- Customer engagement

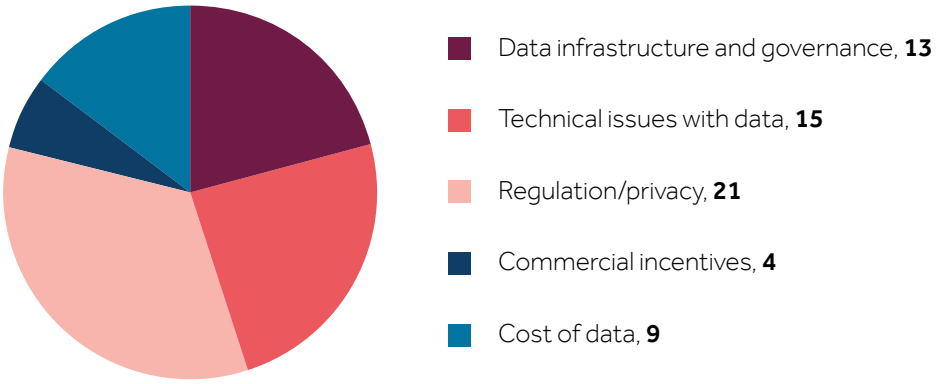
2.2 We also identified challenges with accessing and sharing data in financial services, which are inhibiting the development of new products and services in the market. This insight is supported by evidence gathered during our previous engagement with industry: for example firms in the first and second Digital Sandbox pilots emphasised that accessing financial data is a challenge, specifically for firms in early stages of development.

2.3 In this section of the Call for Input, we wanted to understand whether respondents agreed that access to data is important for innovation within financial services. We also wanted to know whether respondents agreed with our assessment of the difficulties in accessing high-quality financial datasets. In particular, we wanted specific information on the challenges they faced, for example understanding the legal requirements for data sharing, cost, and technology infrastructure.

Summary of responses

2.4 Respondents universally indicated that data is crucial for innovation, regardless of their organisation type. All respondents also expressed that accessing high-quality datasets is a challenge in financial services.

Challenges accessing data



Numbers represent number of respondents. Some respondents indicated multiple challenges in their response.

2.5 16% of respondents stated that purchasing high-quality data can be costly, with one respondent indicating that buying full datasets can be expensive and impractical as a new entrant. Whilst open access datasets exist online, their quality can be uncertain and their utility for testing models is often limited. A further 13% of respondents noted that complex due diligence processes to onboard data, whilst necessary, can be costly and can prohibit new entrants from accessing data. These respondents represented banks, data companies, and trade associations, suggesting that this particular challenge is well understood across the industry.

2.6 41% of respondents indicated that issues with infrastructure and governance are inhibiting their access to data. In particular, respondents indicated legacy IT systems and a culture of operating in data silos as particular challenges. 28% of firms also referenced poor data management practices and a lack of standardisation between datasets – including format, structure, labelling and management – as impeding the quality and consistency of datasets. The lack of data structure standardisation causes issues with technology interoperability, and limits firms' ability to use external datasets to enrich their internal data.

2.7 In addition to data infrastructure and governance, 47% of respondents referenced technical challenges with the real data. 31% stated that datasets are often too small or incomplete to train predictive or machine learning models to a sufficient degree of accuracy. For particular use cases, including novel 'edge cases', the data does not exist in sufficient quantities for training purposes, or does not exist at all. Respondents identified fraud as a specific use case where known 'bad behaviour' comprises too small a percentage of the overall dataset to sufficiently train AI models.

2.8 Poor data quality is also a significant barrier to innovation. Respondents indicated that investment in data quality varies from firm to firm: often, datasets are not kept up to date, lack referential integrity, and can be highly unbalanced due to missing data. Similarly, three respondents referenced data bias as a particular challenge in response to this question. A vague understanding or appreciation of bias in training datasets can

reduce the effectiveness and accuracy of AI models. If not appropriately addressed at an early stage in model development, bias can result in harmful outcomes for consumers. Bias arises as a key challenge and consideration throughout responses to this Call for Input.

2.9 66% of respondents referenced regulation or data protection laws in response to this section of the Call for Input. Respondents emphasised the importance of data protection legislation to ensure that consumers' privacy is protected. Whilst respondents agreed with our assessment that access to data can incur lengthy onboarding and due diligence processes, they reiterated that data access should have consumer privacy built in by design at every stage of the process. According to respondents, the challenges associated with regulation and data protection legislation can be categorised into three themes:

1. **Complex regulatory environment:** the global regulatory environment is complex and there is a lack of harmony in data laws between different jurisdictions. Changing requirements and broader shifts in the global regulatory environment are difficult to implement, specifically for firms operating across multiple jurisdictions.
2. **Lack of clarity in regulations:** respondents also stated that the language used in data protection regulation is unclear, and requires further clarity before organisations feel confident in sharing data.
3. **Data protection legislation and innovation:** Data protection regulation can be used to achieve beneficial innovation that protects consumers' right to privacy, however respondents stated that it can be difficult navigating certain requirements (for example transparency requirements) when using data for training, research and exploration. One respondent stated that data protection legislation should accommodate greater data sharing for use cases where innovation can bring significant societal benefit, including fraud, anti-money laundering and terrorist financing.

2.10 Several respondents differentiated between the exact language of data protection regulations versus the risk-averse practices the regulation can encourage. Data protection regulations do permit data sharing under specific conditions, however the legal, monetary and reputational consequences of violating data protection regulations reduce incentives for data sharing across the industry. One respondent highlighted that the FCA's new Consumer Duty – which requires firms to deliver good outcomes for consumers – may further strengthen firms' caution when sharing their customers' personal data, potentially making external access to this data more difficult.

Our response

2.11 These findings complement our view that data is crucial to innovation, however accessing and sharing data in financial services is challenging for various reasons.

2.12 We agree with the assessment that granular data issues, including the size of datasets, data quality and bias, inhibit the utility of data for testing and development purposes. Studies have shown that synthetic data can resolve some of these issues; for example

by addressing bias in real datasets, and by augmenting data to improve the modelling and testing of rare events such as fraud.

2.13 Many of these findings are also supported by our ongoing engagement on machine learning and AI:

- The AI Public-Private Forum's (AIPPF) final report noted that firms' data management and governance is sometimes organised in silos, which can be particularly inefficient for AI systems that require a holistic, cross-functional approach.
- The challenge of legacy systems and data silos aligns with findings with the recent machine learning survey jointly conducted by the FCA and the Bank of England. The survey found that the greatest constraint to machine learning adoption and deployment is legacy systems. Respondents to the survey flagged the same challenges when we ran the survey in 2019, suggesting that legacy systems are a persistent barrier to the integration of new technologies.
- With regards to data bias, the Artificial Intelligence Discussion Paper notes that AI may pick up bias within training datasets and therefore not perform as intended. The Discussion Paper suggests that new data quality metrics, including representativeness and completeness, may be needed.

2.14 With regards to data protection legislation, we acknowledge that certain regulations place conditions on data sharing to meet certain requirements, whether this is in the interest of consumer privacy or competition for example. Whilst we appreciate that expanding data access and sharing could promote innovation in financial services, the safeguarding of consumer privacy through data protection regulations is of the upmost importance.

2.15 Whilst the Data Protection Act places conditions on data sharing, we emphasise that data sharing between different entities is possible under the current regulatory framework if at least one lawful basis for sharing data is met. For example, the FCA has previously shared data externally in a manner compliant with data protection regulation. In September 2022, we hosted an Authorised Push Payment Fraud TechSprint during which we obtained pseudonymised transactional datasets from multiple banks. Participants in the TechSprint had access to the real datasets (as well as synthetic transactional datasets with embedded fraud typologies) in a closed sandbox environment. From both this and our previous experiences with data sharing, the main learning was the importance of a robust process with privacy considerations at the centre to ensure that any data sharing was compliant with regulatory obligations. To do this we needed to comply with Data Protection Impact Assessments, Data Sharing Agreements and other legal requirements. We also placed additional protections in the sandbox environment to prevent participants removing real data from the platform.

2.16 Ultimately, data protection legislation permits data sharing if firms follow robust processes that build in privacy by design, consult independent legal advice where necessary, and educate themselves around the privacy considerations at all stages of the process.

2.17 Considering changes to data protection regulations falls beyond the scope of our regulatory remit. However, we acknowledge that the sensitivity and wealth of the data held by financial institutions can make data sharing in financial services a complex process, perhaps more so than some other industries. We believe it is our role to understand the impact of regulation in promoting innovation in financial services, particularly where this may create challenges for innovation. We will continue to engage closely with the ICO to explore opportunities for data sharing in financial services within the bounds of UK data protection regulation. We will also continue to engage with the private sector and academia to identify opportunities to collaborate on this particular challenge.

Chapter 3

Assessment of synthetic data

3.1 We consider that synthetic data may play a key role in expanding data sharing opportunities and ensuring that data-driven innovation is steered by techniques that respect consumers' right to privacy.

3.2 In the Call for Input, we wanted to explore industry attitudes towards synthetic data, including the benefits and the potential limitations. We outlined 3 overarching benefits for synthetic data:

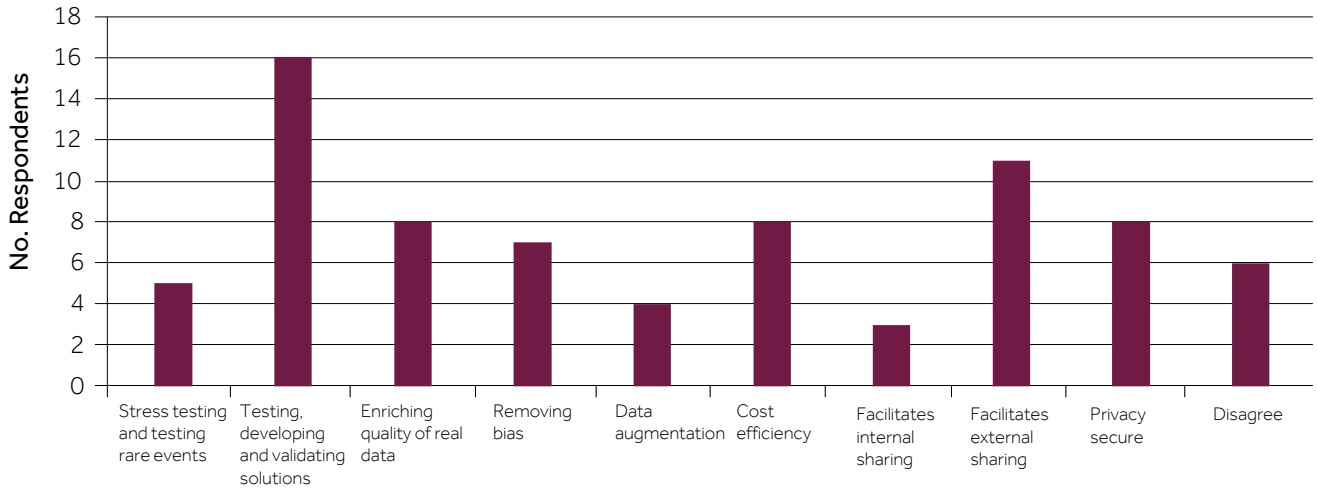
- Data privacy – when data privacy considerations make collecting, sharing, and accessing real data difficult or with prohibitive timeframes.
- Real data is limited or does not exist – where the data required is rare, does not exist in sufficient quantities for training purposes, or it does not yet exist and must be simulated for as yet unencountered conditions.
- Cost efficiency – large volumes of training data are needed for training accurate machine learning algorithms. However, it can sometimes be more efficient to generate high volumes of synthetic data than capture and/or label real data.

3.3 In terms of the limitations, in the Call for Input we referenced the privacy-utility challenge whereby the greater the accuracy of the synthetic dataset, the higher its privacy disclosure risk. We also highlighted that outliers pose a significant risk of de-anonymisation due to the replication of unique or unusual characteristics. Finally, we mentioned the risk that synthetic datasets will replicate and/or introduce bias if this is not accounted for during the generation process.

3.4 We also wanted to explore the maturity of synthetic data generation across small and large firms, regulated firms and FinTechs, technology firms and academia. To date, we have largely experimented with synthetic data more than anonymisation, pseudonymisation or PETs. We wanted to benchmark our internal research to industry practices, and assess the range of technologies that firms use to enable data sharing in financial services.

Key findings

Benefits outlined by respondents



3.5 The majority of respondents agreed with our assessment of the benefits of synthetic data. The most commonly-cited benefit related to testing, developing and validating novel solutions. This finding aligns with much of the wider research on synthetic data, where the ability to train AI and machine learning models without requiring access to real data is often referenced as a key benefit.

3.6 21% of respondents highlighted that synthetic data could enrich the quality of data in various ways, for example by fixing structural deficiencies in the real data, by developing large-scale datasets where labelled data is scarce, and by balancing skewed data. A further 18% reiterated that synthetic data can be designed to remove biases in existing data where a definition of fairness is built into the generation process. Synthetic data can also be used to retrospectively test and re-train algorithms that are proven to be prone to bias.

3.7 10% of respondents indicated data augmentation – the ability to increase the size of a real dataset to train algorithms – as a key benefit. Generating high volumes of synthetic data can be far more cost efficient than capturing and labelling vast quantities of real data. In addition, firms can use synthetic data to model uncommon scenarios and even unencountered conditions, which is particularly helpful for stress testing or testing rare events, for example fraudulent activity.

3.8 Whilst the majority of respondents agreed with our assessment of benefits, two respondents highlighted that the benefits of synthetic data are yet to be proven given the nascency of the technology. Two further respondents indicated that the data privacy benefits of synthetic data are overstated, either due to the real identity disclosure risks associated with synthetic data, or because the processing of real data to produce synthetic data will trigger data protection regulation. Further evaluation of and guidance for this technology is therefore required to realise these benefits across a range of use cases.

3.9 For the limitations, respondents agreed with our assessment of the privacy-utility challenge and that extreme care must be taken with outliers to minimise the risk of de-anonymisation. Several respondents indicated that using a combination of PETs – such as differentially private synthetic data – could mitigate the privacy risk further.

Privacy Enhancing Technologies (PETs)

Privacy Enhancing Technologies (PETs) is an umbrella term covering a broad range of technologies designed to maximise the use of data by reducing risks inherent to data use. According to the ICO, 'PETs are technologies that can help organisations share and use people's data responsibly, lawfully, and securely, including by minimising the amount of data used and by encrypting or anonymising personal information.

Many technologies fall under this umbrella term. Below are several that have grown in prominence in recent years.

- **Homomorphic encryption:** an encryption method that enables computational operations on encrypted data.
- **Secure multi-party computation (SMPC):** similar to homomorphic encryption, SMPC allows users to conduct computational operations on multiple encrypted data sources.
- **Differential privacy:** differential privacy adds 'statistical noise' to a dataset to mitigate the privacy risk, and can be used to statistically quantify the privacy risk of a dataset.
- **Zero-Knowledge proof:** a method whereby one party can prove to another party that a given statement is true without revealing the statement's contents.
- **Synthetic data:** data generated using an algorithm, rather than from real-world events.

3.10 Respondents identified several other challenges involved in synthetic data generation:

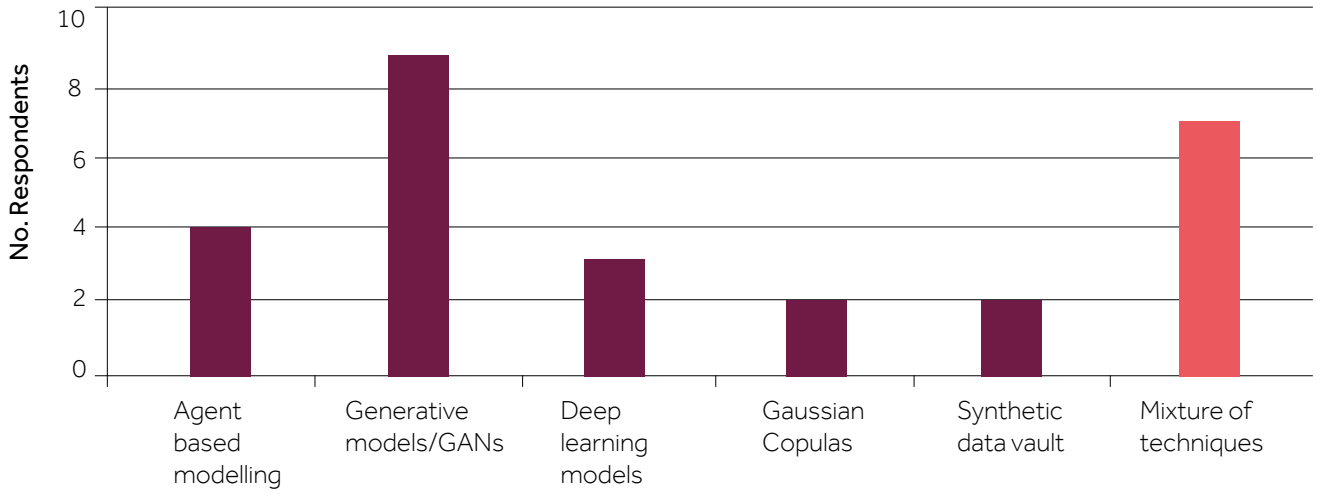
- **Quality** – 21% of respondents stated that synthetic data can replicate bias in the real data, if this issue is not directly accounted for prior to generation. Furthermore, measures to mitigate the risk of de-anonymisation, including intentional data loss or the removal of statistical connections, can introduce additional biases into the synthetic data. 18% also highlighted that synthetic data generation can lead to inaccuracies, especially when real data does not exist and assumptions are made when generating the data.
- **Validation** – validating synthetic data requires access to the real data as well as an understanding of generation techniques and processes used, which can be challenging when a firm has purchased the synthetic data from a third party. In addition, there are no clear benchmarks on what comprises 'good performance' for synthetic data generation methods. There are a multitude of different utility metrics for validating synthetic data, and therefore determining the correct technique or metric can be a challenge for organisations.

- **Ethics –** 18% of respondents highlighted the difference between using synthetic data to test processes and technologies versus making real world decisions. If the synthetic dataset contains biases or is of poor quality, firms could make incorrect decisions resulting in consumer harm, discrimination and exclusion. One firm indicated that consumers should have the right to challenge systems and decisions based on synthetic data. Similarly, two firms highlighted the 'black box' of algorithms as a key challenge. Algorithm-based generation methods are not always inherently 'explainable', which invokes ethical considerations where synthetic datasets are used in decisions with real impacts on consumers.

3.11 Around half of respondents indicated that they generate, are experimenting with, or are exploring synthetic data. Of these, nine respondents generate synthetic data as a core offering of their business model or platform. The majority of firms generating synthetic data fall under the category of FinTechs, RegTechs or data firms. In general, larger firms indicated that they had started to experiment with their synthetic data capabilities; however based on their response, they are at a less advanced stage of development compared to the smaller firms.

3.12 Respondents highlighted a range of different techniques for synthetic data generation. Generative models, including Generative Adversarial Networks (GANs) were the most prevalent technique amongst respondents, however seven participants also indicated that they used a range of techniques.

Generation techniques



3.13 Firms that use agent-based modelling indicated its advantages for generating realistic data at the granular level and for not requiring access to real data. Those that used data-driven methods, including GANs, stated their utility for replicating statistical properties at the macro-level. Other techniques included the Synthetic Data Vault python package, Gaussian Copulas, and other deep or machine learning models.

3.14 Of the 28% of firms that used a variety of generation techniques, a) the complexity of the input dataset, b) the required results or outcomes they were trying to achieve, and

c) the use cases or problems they were trying to solve, were all factors in determining which technique to use.

3.15 We also asked firms whether they had experimented with anonymisation, pseudonymisation and PETs. 35% of respondents had also experimented with anonymisation and pseudonymisation, of which 25% expressed that synthetic data gave better privacy guarantees. 10% of these firms also indicated that achieving the desired level of privacy with synthetic data can be more cost efficient than using traditional anonymisation or pseudonymisation techniques.

3.16 Amongst respondents, experimentation with other PETs was significantly lower than engagement with synthetic data. 15% of respondents have previously combined differential privacy techniques with synthetic data to offer mathematical privacy guarantees, and one firm had explored the use of homomorphic encryption. The latter firm also preferred synthetic data to solve their particular use case.

3.17 Ultimately, the benefits of using synthetic data versus other PETs can vary depending on the use case. Whilst respondents to this Call for Input indicated a preference for synthetic data, it is unlikely that, on an industry-wide scale, synthetic data would be preferred over other privacy-related technologies across all use cases.

Our response

3.18 We believe that the ability to model and augment rare patterns and edge cases is a particularly beneficial property of synthetic data, alongside the privacy benefits already outlined. We have seen this benefit in practice in the autonomous vehicle industry, and believe this could provide significant advantage for financial services. For example, several studies have cited the utility of synthetic data for fraud detection (specifically credit card fraud) and prevention due to its effectiveness in modelling rare malicious activity.

3.19 Respondents indicated that further guidance is required regarding the data processing implications of generating synthetic data. The ICO have released draft guidance on PETs and synthetic data, including the data processing implications of generating synthetic data. In general, processing personal data is most likely to occur where synthetic data is generated by reference to personal data (as opposed to techniques that do not require real data as an input). Firms will need to identify a lawful basis under GDPR to use personal data to a) generate synthetic data in the first place and b) validate the utility of the synthetic data where real data is processed to do so. We continue to closely engage with the ICO on the data protection implications of generating and sharing synthetic data.

3.20 The challenge of validating both the fidelity (the statistical similarity of the synthetic dataset to the input real data) and the utility of synthetic data (evaluating its ability to meet a given use case or purpose) is an area of active academic debate in this field. Broadly speaking, organisations can validate synthetic data using either 'broad' or 'narrow' measures. 'Broad' measures assess the fidelity of the synthetic dataset to the real dataset by quantifying the statistical similarities between the original and generated

dataset. 'Narrow' measures by comparison will compare the differences in model performance, for example inference or prediction, between the original and synthetic dataset. Both forms of measures have their limitations and benefits depending on the purpose of the synthetic dataset. There is therefore no general measure that firms can use to validate synthetic data across all use cases, and users will need a strong understanding of the purpose of the synthetic data in order to validate the dataset.

3.21 However, these validation measures often assume access to the real data, which can be a challenge when organisations purchase data from an external party. This is a situation where an independent third party, such as a regulator, could play a beneficial role. We discuss the role of the regulator in further detail later in the Feedback Statement.

3.22 In early 2023, in collaboration with the Alan Turing Institute and ICO, we will host a roundtable to engage academia and industry on the challenge of synthetic data validation. We will publish the key findings of this roundtable, alongside our research to date on this topic, in the coming months.

3.23 We are encouraged to see that respondents are reflecting on the broader ethical and consumer considerations when generating or using synthetic data. In July 2022, the FCA published a new Consumer Duty which introduced the new Consumer Principle: 'a firm must act to deliver good outcomes for retail customers'. When making real-world decisions based on analysis from synthetic data, firms must ensure that synthetic data is fair, representative and take required steps to remove bias from both the real input data (where relevant) and the generated dataset. Firms should also not use synthetic data as a substitute for real data in every instance, and should consider for every use case whether real or synthetic data is more appropriate.

3.24 We are also encouraged to see engagement with synthetic data across all types of organisations in the industry. We consider this feedback a strong basis for further research into the role we could play to foster innovation through synthetic data. Future avenues for engagement could involve granting organisations access to synthetic datasets hosted on our Digital Sandbox platform. This would help the FCA to gather granular evidence on synthetic data utility over a longer period and for specific use cases.

3.25 To this end, we also recognise that we need to continue collaborating with industry to explore methods to expand data sharing opportunities to drive innovation in financial services. This Call for Input sits within a broader synthetic data and PETs work programme which seeks to understand the risks, benefits and opportunities of these technologies in mitigating data access challenges in financial services. By expanding opportunities for data sharing, our ambition is to drive data-driven innovation in the industry, for example AI model development.

3.26 To date, we have extensively engaged with the public and private sector to explore opportunities to expand data sharing:

- In July 2019, we hosted our Global Anti-Money Laundering and Financial Crime TechSprint, focused on how PETs can facilitate the sharing of information regarding money laundering and financial crime concerns.

- In July 2022 the UK and US governments launched a joint PETs innovation challenge to explore the potential of these technologies to combat financial crime and public health emergencies. The FCA acts as an observer in this challenge.

3.27 In addition, the FCA hosted an Open Finance Policysprint in November 2022 to explore key considerations for a regulatory framework for Open Finance.

3.28 We welcome future opportunities to collaborate with industry, academia and government to progress our work and the field more broadly. The 2022 Gartner Hype Cycle for Artificial Intelligence places synthetic data at the 'Peak of Inflated Expectations', so we believe now is the time to work together efficiently and inclusively to identify the critical use cases and best practices that will drive this technology forward.

3.29 To this effect, we are establishing a Synthetic Data Expert Group to create an effective framework for collaboration across industry, regulators, academia and wider civil society on issues related to synthetic data. This group will explore key issues in theory and in practice with the use of synthetic data in UK financial markets and identify best practices for adoption. It will also provide a sounding board on specific FCA synthetic data projects, for example our upcoming project to utilise synthetic data to test the effectiveness of transaction monitoring systems in identifying money laundering.

3.30 Applications to join the group will open in February, and we will hold the first session in the spring.

Chapter 4

Synthetic data use cases

4.1 The FCA has previously experimented with synthetic data generation to solve a variety of use cases including:

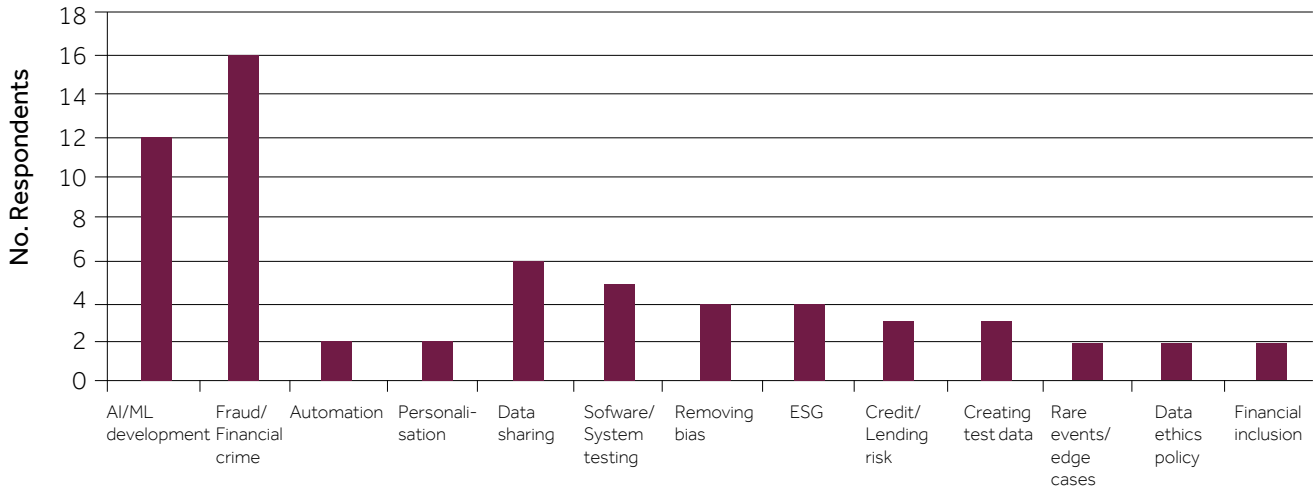
- Financial inclusion & Women's Economic Empowerment
- ESG data and disclosure
- Fraud, vulnerability and SME Finance
- Anti-money laundering
- Authorised Push Payment fraud

4.2 In the Call for Input, we were interested in exploring industry perspectives on the most valuable use cases for synthetic data. Our interest encompasses both technical use cases, for example AI and machine learning model development, and thematic use cases, including financial crime, credit risk, and financial inclusion.

4.3 We also wanted to understand the granular data requirements for realising the benefits of synthetic data. Together with the feedback on use cases, these findings could feed a future pipeline of collaborative projects in synthetic data sharing.

Key findings

Highest priority synthetic data use cases



4.4 When asked about the highest priority use cases for synthetic data, respondents indicated a range of both technical and thematic use cases. The most commonly cited use case(s) related to fraud, financial crime and anti-money laundering (48%), AI and machine learning model development (36%), and data sharing (18%).

4.5 The respondents who indicated fraud, financial crime and/or anti-money laundering indicated a range of benefits for solving challenges in this field. Synthetic data can enrich

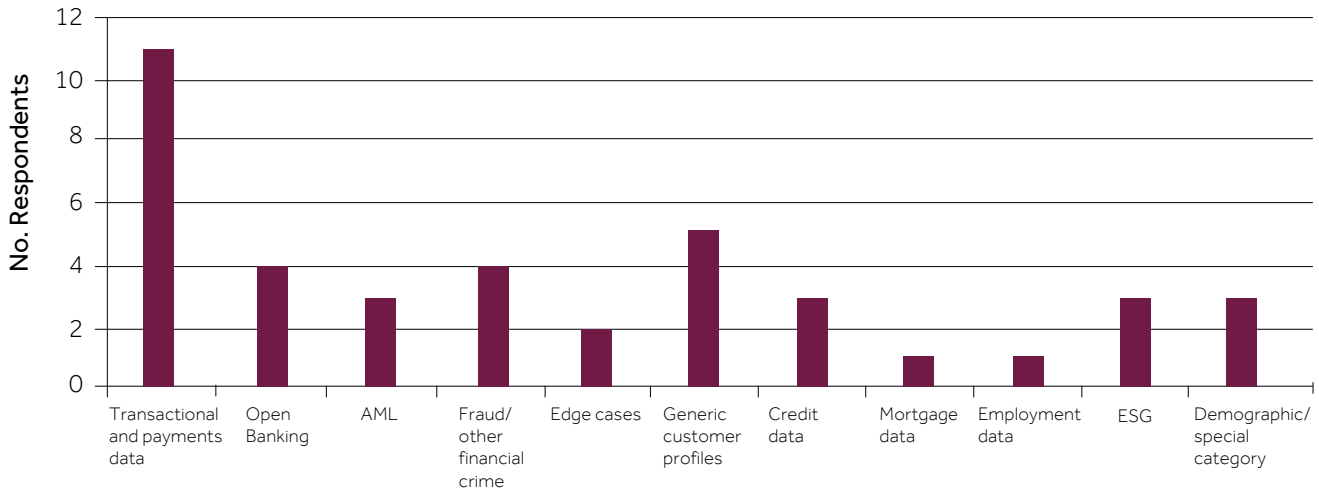
real datasets with known fraudulent typologies in order to develop fraud algorithm maturity, developers can evaluate fraud algorithm performance on synthetic datasets, and firms can automate the process of user identity verification using AI trained on synthetic data. The latter has particular benefits for Know-Your-Customer (KYC) checks. For anti-money laundering, respondents cited the benefit of synthetic data in assessing cross-border money laundering risks and for testing the effectiveness of transaction monitoring systems.

4.6 Synthetic data is also particularly useful for training, testing and validating machine learning and AI models. As mentioned, organisations can use synthetic data to augment datasets and create the large volumes of data required to test algorithms. Synthetic data can also model rare events and anomalies, and increase the frequency of these events in training data, to train algorithms to better respond to market-moving or unknown events.

4.7 In the field of data sharing, respondents cited particular benefits for cross-border data sharing and third-party collaboration, particularly with academics and researchers. Respondents also indicated a range of other use cases, including ESG, credit and risk lending, the personalisation of services, and removing bias from real datasets.

4.8 We were also interested in understanding which synthetic data assets organisations would find the most valuable. FinTechs and RegTechs were the largest response group to this question.

Most useful synthetic datasets



4.9 Respondents indicated transactional and payments data as the most useful type of synthetic data asset, specifically to train trading models and for use cases related to fraud and anti-money laundering. 22% of respondents also specified synthetic data that can be used to build customer profiles and ultimately improve the personalisation of services. This category of data includes credit data, mortgage data and employment data for example.

4.10 We also asked respondents what requirements they would need in order for the synthetic data to meet their use cases. Four key themes came through in respondent answers:

1. Large volumes: respondents indicated that they need access to high volumes of data, with one respondent indicating 100,000s of data points.
2. Content: 40% of respondents stated statistical accuracy as a key requirement, where 'accuracy' referred to matching data distributions, the reproduction of rare signals in the synthetic data, corresponding typologies and behaviours, and the similarity of model performance on a synthetic dataset. 8% of respondents stated that synthetic data would need to support different data types, including numerical, categorical, date/time and location. 20% of respondents also indicated that the synthetic data must be representative and be accompanied with assurances that no bias was introduced during the generation process.
3. Transparency: respondents desired transparency about the process and algorithms used to generate the synthetic data, for example to enable firms to assess the reliability of the dataset. One firm also stated access to high-level details of the original dataset, including data volumes, as a key requirement. In addition, several respondents stated that supporting documentation, including data model diagrams and data dictionaries that explain the relationship between features, should accompany the synthetic dataset.
4. Linking data: 32% of respondents emphasised the criticality of relational and referential integrity between multiple datasets – when datasets are linked together with relationships, any changes to the contents of one dataset should be reflected in the relational datasets. 8% also indicated that synthetic data should be able to join with other data at an aggregate level.

Our response

4.11 The use cases above are consistent with the themes we have explored throughout our synthetic data initiatives, for example financial crime and anti-money laundering. Whilst synthetic data can provide benefits for fraud use cases, we would caveat that achieving these benefits can be difficult in practice. For example, injecting typologies into a synthetic dataset first and foremost requires subject matter expertise to identify the fraudulent patterns of behaviour. To remain effective, firms would require a process of constantly identifying new typologies and fraudulent behaviour, and reflect these typologies in the synthetic dataset on a regular basis.

4.12 The prominence of AI and machine learning model development amongst respondent answers reflects our findings through our Digital Sandbox pilots to date. For example, 62% of successful applicants to the Digital Sandbox have tested products and services based on AI and machine learning models across both pilots.

4.13 It is interesting that several respondents indicated ESG as a key use case for synthetic data. For our most recent Digital Sandbox pilot, we sought to generate synthetic data to tackle challenges related to ESG data and disclosures. We encountered several challenges in creating synthetic data to tackle this use case. For example, ESG is

a nascent field where real data can be rare, or may sit with multiple vendors, large institutions, trading venues and companies providing satellite imagery data. We also found that much of the data is unstructured and therefore less suitable for synthetic data generation. We published these key findings and challenges in further detail in our Digital Sandbox Sustainability Pilot evaluation report. There are also several academic and industry studies outlining the challenges associated with accessing ESG data.

4.14 Ultimately, when generating synthetic data for a particular use case, it is important to understand the availability and quality of real data in the field, especially when using generation models that require access to real data as an input.

4.15 We expected transactional and payments data to be a common theme regarding the most useful synthetic datasets due to the level of detail and variables that transactional datasets contain. Participants in both Digital Sandbox pilots, which focused on different use cases, often indicated transactional data as a key requirement for developing their products and solutions.

4.16 Our Digital Sandbox platform hosts multiple synthetic transactional datasets, including retail transactions, synthetic credit card history data, banks' synthetic data, and several transactional datasets for synthetic individuals. Participants in our Digital Sandbox and TechSprint initiatives have access to these datasets to test and develop ideas, proof of concepts and products. We are currently exploring avenues to expand access to these datasets to a broader community in a privacy-secure manner, in order to promote innovation in financial services.

4.17 With regards to the requirements for synthetic datasets, respondents indicated a variety of different metrics to assess the 'accuracy' of synthetic data, which reflects a lack of common agreement in the broader field. Ultimately, the appropriate metrics for 'accuracy' (the fidelity and utility) will depend on the purpose of the synthetic dataset and the use case. For example, users will need to consider whether they require statistical similarity on a macro or a granular scale, and whether they require queries to the synthetic data to reveal the same results as the real data (for example the number of transactions over a certain limit). When considering the features, developers may need to consider when and where they need to model the joint distributions across multiple features. The greater the need to link features, the greater the complexity of the dataset and therefore the more compute power is needed to generate the synthetic data.

4.18 The need to assess fidelity and utility on a case-by-case basis creates challenges for building a synthetic dataset to serve multiple purposes. As mentioned, we will hold discussions with academia and industry on the theme of validating synthetic data and will publish our findings in the coming months.

4.19 We appreciate the desire for transparency regarding the input data and generation process to give organisations confidence in the reliability of the synthetic data. For third party synthetic data firms, there may be intellectual property considerations that prevent the sharing of details regarding the generation process and/or algorithms. Information on the real data would also have to remain at a high-level, as access to the real data could pose other legal issues and could mitigate the need for the synthetic data in the initial instance. As a potential solution to this challenge, a third party could store the real data in a secure environment, with the purchaser of the synthetic data sending

specific queries to the third party to test the fidelity of results between the synthetic and real data.

4.20 That being said, the AI Public-Private Forum final report, published in 2022, argues that transparency and communication are key elements of governance. Similarly, the joint FCA-Bank of England AI Discussion Paper published in October 2022 states that 'a lack of explainability or transparency in some AI models may mean extra care or actions are needed to ensure full accountability and sufficient oversight'. We believe it is important that further research is conducted into methods to ensure the transparency and explainability of synthetic data generation models (and input data) without revealing commercially or personally sensitive information.

Chapter 5

Role of the regulator

5.1 When used accurately, ethically and with privacy at the forefront, synthetic data has the potential to advance all three of our operational objectives:

- Market competition: democratising data access across the industry can diversify the range of firms with access to quality data. In turn, this could accelerate the development of disruptive products and services in the market.
- Market integrity: measures to promote effective data sharing could also accelerate the development of RegTech tools and initiatives and strengthen firms' compliance programmes to meet regulatory obligations.
- Protecting consumers: synthetic data, and other PETs, can enable data sharing within and between organisations in a way that reduces the risk to consumer privacy.

5.2 We are interested in exploring the role of the regulator regarding the use and adoption of synthetic data in financial services. In the Call for Input, we outlined three broad roles:

1. Data Generator: The regulator collaborates with industry experts and academia to generate synthetic data in-house, to be shared with the industry. The regulator could obtain real data from multiple entities, ensuring a cross-section of industry is sampled and the data is not biased towards a single organisation. Synthetic data could also be shared with organisations holding real data for iterative benchmarking purposes, improving the quality of the data over time.
2. Central Host: The regulator provides an independent hosting platform through which synthetic data can be stored, shared and accessed for the purposes of product development and testing.
3. Coordinator: The regulator acts as a co-ordinating body to facilitate data sharing and/or pro-competitive collaboration opportunities for synthetic data generation.

5.3 These roles are not independent, and a regulator could perform one of the above roles or a combination of the three.

5.4 Synthetic data sharing at scale would also require significant engagement and commitment from firms and public organisations. We would therefore like to assess the appetite for firms to collaborate with the FCA and other organisations, for example by providing real data or synthetic data expertise as an input in the generation process.

Key findings

5.5 Almost half of respondents indicated that the regulator should play a coordination and/or intermediary role in the provision of synthetic data. 31% of respondents stated that the regulator should play a role in the hosting and provision of synthetic data, and set clear guidelines on who is permitted to access the datasets. 18% of respondents also specified that the regulator should coordinate industry and regulatory efforts

in this space. Several respondents referenced the Digital Sandbox as an appropriate environment to a) facilitate collaboration around the creation of synthetic datasets and b) utilise these synthetic datasets to test innovative products with companies and customers.

5.6 Only 10% of respondents explicitly stated that the regulator should generate synthetic data, caveating that any generation should be done in collaboration with synthetic data experts. 13% of respondents indicated that the regulator should not produce synthetic data for a variety of reasons: it is unlikely to match the quality of experts in the industry, it could undermine third parties whose business model is to generate synthetic data for commercial purposes, or because market-led development of synthetic data should be promoted.

5.7 Although not explicitly outlined in the Call for Input, 31% of respondents indicated that the regulator should produce guidelines, standards and/or governance frameworks on the adoption of synthetic data. This insight points to a broader theme throughout responses to the Call for Input: common standards and granular guidance on case studies and use cases are needed to build trust in synthetic data. The absence of these frameworks is limiting investment into synthetic data initiatives across industry, as firms require greater confidence in the technology to move from the proof of concept stage to a broader adoption of synthetic data.

5.8 We also asked whether synthetic data should be a public utility for the purposes of innovation and research, or whether it should be monetised. 22% of respondents to this question stated that synthetic data should be monetised, at a minimum to cover the resources required to generate the data. Several respondents also specified that they would be willing to pay for synthetic data that relates to their highest priority use cases. Conversely, 28% stated that synthetic data should be a public utility for the purposes of innovation, collaboration and to encourage the broader adoption of this technology in the market.

5.9 For 50% of respondents, the question of whether they would be willing to pay for synthetic data assets depended on a variety of factors. There are many available open-source synthetic datasets for the purpose of innovation and research, and therefore any attempt to monetise synthetic data would need to be accompanied by a strong value-add and quality guarantees. Several respondents also referenced the enhanced privacy risk of open-source synthetic data, suggesting that the security of the synthetic dataset should be tested in a closed environment with limited access before being made a wider public utility. The use case for the synthetic data was also an important consideration for respondents. Several firms specified that they would be willing to pay for synthetic data that met their highest priority use cases. Respondents also indicated that for use cases that provide a particular benefit to the public, for example fraud and financial crime, synthetic data should be made a public utility.

5.10 Many respondents emphasised the importance of collaboration with other national bodies and initiatives to accelerate the adoption of synthetic data and PETs in UK markets. In particular, respondents referenced the Digital Regulation Cooperation Forum (DRCF), the Government's National Data Strategy, the Centre for Finance, Innovation and Technology (CFIT) and the Open Banking initiative.

Our response

5.11 We are encouraged to see a common desire for standards and governance frameworks to ensure the consistent and fair adoption of synthetic data across industry. We agree that standards are important to deliver quality outcomes, and have outlined 'setting and testing higher standards' as a key focus area in our three-year strategy. We understand that there are several challenging questions that need to be answered – specifically around quality, privacy and transparency – to accelerate the adoption of synthetic data in public organisations and industry. We will continue to collaborate with the ICO to tackle these challenges and provide guidance where required.

5.12 Through the DRCF, we will work with other regulators to understand the cross-sectoral risks, benefits and use cases of emerging technologies, including synthetic data and PETs. We will also aim to work more closely with standard setting bodies to advance conversations around governance frameworks and standards where appropriate. We believe that by advancing and maintaining globally high standards, we can embed competitiveness throughout our regulatory approach and ensure that the UK is open to innovation.

5.13 We agree that providing a trusted and secure environment to store and share synthetic data could be a potential role for the regulator. The FCA is already performing this role through our Digital Sandbox platform. We currently host over 200 real and synthetic datasets on the platform, and we are committed to exploring ways of making the data assets on this platform more widely available in a way that complies with data protection legislation.

5.14 That being said, we need to carefully consider the risks of opening up our synthetic data assets to a broader audience as a public utility, for example the enhanced privacy risk referenced previously. In the case of synthetic data that models fraudulent typologies, there is also the risk that bad actors could use this data to train their models to become better at committing fraud. Expanding access to any of our synthetic datasets will therefore require robust governance frameworks and security measures to protect the data from malicious intent.

5.15 Whilst we acknowledge that several respondents do not believe the FCA should generate synthetic data, we do however hold a valuable position in convening the firms we regulate to help source real data for synthetic data initiatives. We recognise the importance of accuracy and quality assurances for firms utilising third party synthetic data. When generating synthetic data in previous initiatives, the FCA has collaborated with synthetic data experts across industry and academia. We will therefore continue to engage with these experts to ensure the quality of our work.

5.16 We also agree on the need to collaborate with government, regulators and industry. Further engagement will ensure a consistent approach to synthetic data and PETs, will help to identify any overlaps to drive efficiency, and will encourage greater public-private cooperation. In December, we hosted a discussion with key partners across government and regulation to share the insights from the Call for Input and hear case studies of other public sector initiatives in synthetic data and PETs. Our aim for this discussion

was to contribute to regulatory knowledge sharing initiatives in the field of emerging technologies and to open the dialogue for potential future collaboration in this field.

5.17 In January, we also hosted two spotlight sessions with the Global Financial Innovation Network, a network of over 80 financial regulators chaired by the FCA. Through these sessions, we highlighted the insights from our own work on synthetic data and PETs, and developed our knowledge of global regulatory perspectives and approaches to expanding data sharing in their respective jurisdictions.

5.18 Where relevant, we have followed up with various organisations to explore future collaboration opportunities. We will continue to engage with the DRCF, government, other public bodies and industry to ensure a consistent approach and collaborate to solve the challenges to adoption in this field.

Chapter 6

Next steps

6.1 It is our ambition to promote effective competition in the interest of consumers, and to ensure that UK markets work well for firms and consumers. To achieve this vision requires the investigation and development of key technology enablers that create incentives to innovate and invest.

6.2 Data is crucial for innovation in financial services, however there are several challenges with accessing and sharing data in this industry. Having collated and analysed the responses to our Call for Input (and based on our previous research), our current position is that synthetic data can potentially make a significant contribution to beneficial innovation in UK financial markets. The responses to this Call for Input have given us greater insight into synthetic data practices, use cases and case studies across financial services. We consider these findings an important foundation to validate further investigation of this technology (specifically use case identification and understanding its utility as a regulatory tool) and further engagement with industry.

6.3 A key next step for the FCA will be expanding access to our Digital Sandbox platform as a secure environment to host real and synthetic datasets to promote innovation in the market, specifically for firms who face challenges accessing data in current conditions.

6.4 In addition to stimulating innovation in the market, we are exploring methods to leverage synthetic data to improve our own supervisory and oversight capabilities. We believe that the potential for synthetic data in the field of SupTech is a fruitful area for further research.

6.5 In the longer term, a strong theme throughout responses to the Call for Input is the need for further guidance to build confidence in synthetic data. Setting and testing higher standards that put consumers' needs first and deliver positive change is a key commitment in our three-year strategy. Whilst we have witnessed greater experimentation with synthetic data in recent years, this remains a developing technology that will continue respond to regulatory developments in the broader field of data, and market development of use cases. Close engagement with other regulators, including the ICO and more broadly the DRCF, is therefore key. We will also look to engage more closely with standard setting bodies as and when these technologies mature.

6.6 The FCA will continue to collaborate with synthetic data experts in the market, whilst developing our in-house synthetic data generation capabilities through several proposed upcoming projects. We believe the use cases outlined by respondents – particularly fraud and anti-money laundering – provide key areas of focus for our future efforts in this field. These efforts include internal projects to mature our own capabilities as a synthetic data practitioner, valuable partnerships with industry and academia, and a potential future pipeline for our Digital Sandbox and TechSprint programme.

6.7 In addition, we are establishing a Synthetic Data Expert Group to create an effective framework for collaboration across industry, regulators, academia and wider civil

society on issues related to synthetic data. This group will explore key issues in theory and in practice with the use of synthetic data in UK financial markets and identify best practices for adoption. It will also provide a sounding board on specific FCA synthetic data projects, for example our upcoming project to utilise synthetic data to test the effectiveness of transaction monitoring systems in identifying money laundering.

6.8 Applications to join the group will open in February, and we will hold the first session in the spring.

Annex 1

List of questions in the Call for Input

Q1: How important do you think access to data is for innovation within financial services? What else do you view as significant barriers to innovation?

Q2: Do you agree that it is challenging to access high-quality financial data sets? If so, specifically what challenges do you face? (for example, understanding legal requirements around data access, commercially expensive, or technology infrastructure.)

Q3: Do you agree with the high-level benefits for synthetic data? Are there any other benefits for synthetic data for your organisation, both now and in the future?

Q4: Does your organisation currently generate, use, purchase or otherwise process synthetic data? If possible, please explain for what purpose(s).

Q5: If your organisation generates synthetic data, please describe at a high level the techniques used. Why have you chosen to use this approach?

Q6: What do you see as the difficulties and barriers for firms in creating high-utility, privacy secure synthetic data?

Q7: Does your organisation engage with privacy enhancing technologies or privacy preserving techniques other than synthetic data? How would you assess the utility and benefits of synthetic data in comparison to other techniques?

Q8: What do you see as the highest priority use cases that would benefit from synthetic data?

Q9: Are the synthetic data use cases you have mentioned significant for early business phases or mature operations/ processes within your organisation?

Q10: How would your organisation make use of synthetic data if it was available (if at all)? 18 Annex 1 Financial Conduct Authority Synthetic data to support financial services innovation

Q11: What synthetic data sets would you find most valuable to have access to? For example, Open Banking, Customer profiles, account to account payments, Credit card transactions, trading data, etc. What challenges would these data sets help your organisation to solve? E.g. AML and fraud detection, ESG, etc. Please be specific.

Q12: What requirements would you need for the synthetic data to feasibly meet your use cases? Please be as specific as possible (for example, details on volume, accuracy, referential integrity between sets).

Q13: Do you agree with our assessment of the potential limitations and drawbacks of synthetic data? Are there any others?

Q14: Do you believe that regulators should play a role in the provision of synthetic data? If so, what do you think the extent of that role should be? (e.g. co-ordination, generation, hosting, etc)

Q15: To what extent would you be willing to collaborate with regulators and/or other organisations to generate synthetic data? For example, would you provide real data samples, or benchmark synthetic data against real data sets?

Q16: Do you think access to synthetic data should be a public utility for the purposes of innovation and research? Would you pay for access if it was delivered at-cost, or monetised?

Annex 2

Glossary of terms used in this document

1. This glossary should not be considered an indication of regulatory definitions. The definitions and explanations contained herein are only to clarify references to the associated concepts in the Feedback Statement.

Term	Description
Innovation	Innovation is the creation of new knowledge and ideas to facilitate new business outcomes, aimed at improving internal business processes and structures and to create market driven products and services. Innovation encompasses both radical and incremental innovation.
De-anonymise	Sometimes referred to as re-identification, deanonymisation is a data mining technique that attempts to re-identify encrypted or obscured information.
Artificial Intelligence (AI)	Artificial intelligence is a computerised system that exhibits behaviour that is commonly thought of as requiring intelligence
Machine Learning (ML)	Machine learning is the process in which a computer distils regularities from training data.
Personal data	Information that relates to a natural person that can be used, either directly or indirectly, to identify an individual.
Regulated firm	Any firm that is registered on the Financial Services Register and is regulated by the PRA and/or FCA.
Synthetic data	Microdata records created to improve data utility while preventing disclosure of confidential respondent information. Synthetic data is created by statistically modelling original data and then using those models to generate new data values that reproduce the original data's statistical properties. (Office for National Statistics)
Differential privacy	Differential privacy is a formal mathematical framework for quantifying and managing privacy risks when analysing or releasing statistical data.
Privacy Enhancing Technologies (PETs)	Technologies that can help organisations share and use people's data responsibly, lawfully, and securely, including by minimising the amount of data used and by encrypting or anonymising personal information (Information Commissioner's Office)
GDPR	The General Data Protection Regulation (GDPR) is an EU regulation that controls the processing of personal data and the free movement of such data in the European Union and the European Economic Area and was onshored post-Brexit (now known as the UK GDPR).

Term	Description
Data Protection Act	The Data Protection Act 2018 is the UK's implementation of the General Data Protection Regulation (UK GDPR). It controls how personal data is used by organisations, businesses or the government.
Digital Sandbox cohort	The Digital Sandbox cohort is an 11-week initiative hosted by the FCA and the City of London Corporation, designed to stimulate and foster the development of innovative products and solutions within financial services. Participants are given access to data, mentors and collaboration platforms to prototype and test their proof of concepts, with the aim of reducing time to market.
TechSprint	The FCA TechSprints are events that bring together participants from across and outside financial services to develop technology-based ideas or proof of concepts to address specific industry challenges. The events usually last between 2-5 days, and help us to shine a light on issues and expand the discussion and awareness of potential solutions.

Annex 3

List of non-confidential responses to the Call for Input

Beyond Encryption

Prime Dash

The University of Manchester

Hazy

Newcastle University

Synthesized

Lucinity

Nth Exception

Elucidate GmbH

INEVITABLE

YData

Pub ref: 007999

12 Endeavour Square London E20 1JN
Telephone: +44 (0)20 7066 1000
Website: www.fca.org.uk