FINANCIAL
CONDUCT
AUTHORITY

Generating and Using Synthetic Data for Models in Financial Services: Governance Considerations

The Synthetic Data Expert Group

August 2025

About the Synthetic Data Expert Group

The Synthetic Data Expert Group (SDEG) is a specialised sub-group of the Innovation Advisory Group (IAG) and was established in February 2023 by the FCA Innovation department. It operates under the guidance of the IAG Terms of Reference and chaired by the FCA. The SDEG helps to foster collaboration across industry, regulators, academia, and civil society to advance the responsible use of synthetic data to shape digital markets to achieve good outcomes and digital transformation at the FCA.

The SDEG was launched in March 2023 and closes with this second and final publication.

The SDEG is comprised of 21 members who were selected against a set criterion, in an open and competitive process. The group explore issues surrounding the use of synthetic data in UK financial markets by identifying relevant use cases, key theoretical challenges and sharing practical experiences of using synthetic data. Additionally, the group has provided valuable feedback on FCA projects involving synthetic data. More information on SDEG membership can be found in the Appendix.

Contents

Disclaimer

This report has been collectively authored by members of the SDEG and colleagues across the FCA. The contents of this report reflect the practical experiences members of the SDEG have encountered when generating or using synthetic data. The report is design to help regulators and industry practitioners better understand the opportunities and challenges of synthetic data.

The contents of this report do not represent the views of the FCA or any participating organisation. It does not endorse or condemn the use of synthetic data and does not imply compliance with UK data protection law.

This report and the applications, discussions and outputs of the Synthetic Data Expert Group should not be taken as an indication of recommendations, guidance, or future policy.

Executive foreword

Jessica Rusu,
CDIIO, FCA

Financial services play a key role in supporting economic growth and competitiveness. Growth relies in part on firms' ability to innovate and explore new technologies and approaches, which can benefit consumers, drive market efficiencies, and manage risks. Synthetic data is one such technology.

It offers a powerful way to unlock the value of data, enable experimentation, model development, and broader innovation across the financial system – all while maintaining strong privacy protections and public trust.

Recognising the potential of this technology, we convened the FCA's Synthetic Data Expert Group (SDEG) to bring together leaders from across financial services, academia, and the public sector. Our aim was simple: to enable open and practical conversations about how synthetic data is being used, where the challenges lie, and what's needed to move forward responsibly.

The Group's first report laid the groundwork, showcasing promising use cases and signalling the potential of synthetic data across areas such as fraud detection, customer insight, and fairness in credit. This second report builds on that foundation, focusing not only on what synthetic data could do, but on what it is doing, how it does it, and how leading firms are working through questions of governance, compliance, and trust.

At the FCA, we believe in the power of collaboration to enable innovation. This work reflects our ongoing commitment to convening expert communities, sharing lessons, and supporting safe experimentation in a fast-moving space. By working together, we can lower the barriers to adoption, build confidence in new techniques, and build a more competitive, future-ready financial system to support economic growth and UK competitiveness.

This report is likely to be of interest to synthetic data practitioners and those working with, or exploring, synthetic data across financial services. While it is not exhaustive, this paper brings together a wide range of insights and real-world experience that we hope will prove useful to those navigating the challenges and perceived barriers to using synthetic data safely, responsibly, and effectively within the sector.

I'm grateful to the members of the Synthetic Data Expert Group for their openness and ambition, and to the FCA colleagues who have supported this work. I hope the insights and actions in this report will be informative and encouraging for those exploring the use of synthetic data in financial services.

Jessica Rusu

Jessica Rusu, Chief Data, Information and Intelligence Officer, FCA

Chapter 1

Introduction

1.1 Modern financial services are powered by data. This data allows firms to understand changing consumer behaviour and market conditions, to make more informed decisions, and increasingly to integrate machine learning and artificial intelligence into their operations.

1.2 However, data use and processing carry significant risks, which can lead to harms at both consumer and industry level if left unaddressed. These risks, and the measures introduced to mitigate them, can create friction for innovation and development. Synthetic data – artificial data which replicates the statistical properties of real data while omitting private or sensitive information – is a potential solution to the challenges associated with data use.

1.3 In August 2020, the FCA hosted a 'DataSprint', with the objective of developing synthetic data assets in collaboration with financial services industry participants and experts. The 'DataSprint' provided a unique opportunity for regulator-industry collaboration, allowing participants to explore different methodologies of synthetic data generation. This collaboration, and the feedback received from 'DataSprint' participants, paved the way for further FCA exploration into synthetic data, leading the Digital Sandbox pilot running from October 2020 to February 2021. Following a successful pilot phase, the Digital Sandbox was made permanent in August 2023, clearly signalling the FCA's commitment to further use of synthetic data in an innovation context.

1.4 In March 2022, we launched a Call for Input, which asked practitioners and industry professionals to share their experiences and opinions on a wide range of topics, including perceived barriers to synthetic data adoption, priority use cases for synthetic data, and the role of regulatory bodies in supporting synthetic data adoption and innovation.

1.5 Based on the feedback to the Call for Input, the FCA began to pursue a dual-pronged approach to investigating synthetic data use. Part of our strategy involves working to develop proofs of concept and test synthetic data in different use cases. A key example is the Anti-Money Laundering and Synthetic Data Project, which is a multi-stakeholder initiative including the FCA, the Alan Turing Institute, Plenitude Consulting and Napier AI. Together, this consortium is working to explore synthetic data's role in the development, evaluation, and effectiveness of anti-money laundering systems.

1.6 In addition to testing practical applications of synthetic data, the FCA has also sought to provide thought leadership and bridge the gap between the theory and practice of synthetic data adoption. We have championed this work through the Synthetic Data Expert Group (SDEG), which was formally established in March 2023. Members of the Group were selected following a competitive application process and taken together they represent a wide range of stakeholders from across industry, academia, and other regulatory bodies. The SDEG's first report, *Using Synthetic Data in Financial Services*[1] was published in March 2024. This first report examined six use cases for synthetic data in financial services, with the aim of helping practitioners to understand the tools, techniques, opportunities, and practical challenges associated with synthetic data.

1.7 This second paper builds on the considerations outlined in the first paper and responds to key feedback contained in the 2022 Call for Input. Respondents were clear that common standards and granular guidance could play a central role in building up trust and facilitating synthetic data adoption. Although not explicitly outlined in the Call for Input, 31% of respondents indicated that the regulator should produce guidelines, standards, and/or governance frameworks to facilitate the adoption of synthetic data.

1.8 While this report does not constitute guidance from the regulator, it seeks to highlight insights and best practices identified by the SDEG members. The considerations and actions to assess, manage, and mitigate outlined are non-exhaustive and will require further iteration as synthetic data usage expands. Nevertheless, they inevitably should demonstrate how synthetic data considerations can fit within, or act as a complement to, existing governance frameworks for conventional models and data usage.

1.9 These different approaches to exploring and working with synthetic data form an important part of the FCA's five-year strategy[2]. Both the Synthetic Data Expert Group and the Money Laundering and Synthetic Data Project have pioneered important new models of collaboration between the regulator, academia, and industry practitioners, and we anticipate that such cooperation will be crucial in helping the FCA to champion growth and innovation. Similarly, our exploration of novel and emerging technology solutions represents our ambition to become a smarter, more adaptive regulator, which is ready to embrace the challenges and opportunities presented by novel and emerging technologies.

Existing Governance Frameworks

1.10 The key aim of this report is to explore potential governance considerations for organisations and practitioners planning to work with synthetic data. In the absence of specific existing governance frameworks for synthetic data, members of the SDEG started by carefully reviewing established Data and AI Ethics and Model Risk Management frameworks (MRM), and drawing out key principles which were considered to be relevant to synthetic data usage. As well as providing a strong theoretical basis for a synthetic data governance framework, principles drawn from existing Data and AI Ethics and MRM frameworks benefit from extensively stress-tested and refined in real-life conditions.

1.11 These key principles formed the basis of the SDEG's investigation into synthetic data governance, allowing us to identify both those considerations which overlap with other frameworks, and those which are unique to synthetic data governance. Understanding these areas of overlap and difference may enable practitioners to better understand where synthetic data considerations can exist within existing governance frameworks.

[1] Using Synthetic Data in Financial Services (2024)

[2] FCA Strategy 2025-2035

1.12 Data and AI Ethics frameworks seek to uphold transparency, fairness, and privacy, all of which are important when considering the generation of synthetic data to simulate real-world scenarios. This is particularly relevant when sensitive underlying datasets are involved. Ensuring that synthetic data is accurate, unbiased, and secure supports fairness in decision-making and safeguards against unintended harm. It is also vital when models are trained wholly or partially on synthetic data, to minimise the risk of biases or inaccuracies propagating through models and unfairly impacting on real-world decisions. Organisations may find that the principles developed in Data and AI Ethics frameworks can be useful for building trust in synthetic data technologies and maintaining compliance with regulatory expectations and ethical standards.

1.13 MRM frameworks, such as the one outlined in 2023 by the Prudential Regulation Authority[3], emphasise the importance of validating models, managing their limitations, and maintaining accountability for their outcomes. Many of these principles will extend directly to models trained on synthetic data or a combination of real-world and synthetic data and can thus serve as a useful starting point for organisations wishing to develop a synthetic data governance framework.

1.14 Drawing on the approach to MRM and Data & AI Ethics frameworks, as well as the Government's AI Principles, the Synthetic Data Expert Group have drawn out[4] nine key principles relevant to synthetic data projects which firms may wish to consider when developing their own approaches. These principles served as a reference point for much of our exploration of synthetic data, and readers will find them embedded throughout the report.

1. **Accountability:** Establish clear accountability structures for data, algorithmic and AI systems, defining responsibilities throughout the data and AI lifecycle. Accountability extends to technologies or models from third-party providers and managed service providers, with documented chains of responsibility.
2. **Safety:** Design systems with safety as a priority, encompassing reliability, robustness, and accuracy.
3. **Transparency:** Maximise the information available to a decision-maker validating the system and its outputs.
4. **Explainability and Interpretability:** Ensure system's internal processes are understandable to humans and provide justification for specific outputs.
5. **Security and Privacy:** Design systems to protect both data security and individual privacy rights throughout the data lifecycle.
6. **Fairness:** Systems which process or impact social or demographic data are designed to prevent discriminatory outcomes.
7. **Agency:** Model operators reviewing algorithmic outputs to have meaningful ways to understand, question, and contest these decisions.
8. **Suitability:** Use cases are justified by genuine needs, informed by an understanding of current technological constraints, and considerate of broader socio-technical context.

[3] PS6/23 - Model risk management principles for banks

[4] AI Update , Implementing the UK's AI Regulatory Principles, A pro-innovation approach to AI regulation - GOV.UK

9. **Continuous Monitoring and Improvement:** Regularly assess models and systems to ensure they remain effective, compliant, and fit for purpose.

1.15 While this report does not aim to cover all possible use cases and deployment opportunities for synthetic data in the financial services sector, it highlights some of the core challenges that practitioners may face. The report seeks to outline practical approaches to identifying these challenges within synthetic data projects, along with potential mitigating actions and ways to foster good governance strategies, based on the insights of SDEG members.

Chapter 2

Readiness for synthetic data projects

2.1 Before launching a synthetic data project, organisations and practitioners may benefit from assessing their governance and strategic readiness. To support this, the SDEG have highlighted a set of governance foundations that can support the effective and responsible use of synthetic data. These foundations may also serve as ongoing reference points to support good governance throughout the synthetic data lifecycle

2.2 Alongside these foundations, this section outlines a set of pre-project considerations that can help practitioners to explore whether a synthetic data initiative is appropriate, feasible, and aligned with relevant regulatory, legal, and ethical organisational priorities.

Governance foundations

2.3 Strong governance structures can help organisations manage the use of data, models, and AI – with or without synthetic data. While many organisations may already have elements of these structures in place, the foundations outlined here may offer a helpful baseline for managing the common governance challenges associated with synthetic data.

2.4 These governance foundations are structured into three key areas for consideration:

2.5 **Frameworks, controls and processes:** The internal forums, frameworks, guardrails, controls and escalation pathways necessary to support governance and oversight across the synthetic data lifecycle.

2.6 **Roles and responsibilities:** Assigning clear roles and accountability across all synthetic data projects to ensure that governance is upheld throughout the project's delivery.

2.7 **Documentation and continuous monitoring:** Maintaining comprehensive and transparent documentation throughout the synthetic data lifecycle. This includes capturing decisions, assumptions, and trade-offs made at every stage, as well as tracking changes and updates to datasets, models, and governance frameworks.

2.8 While these governance foundations may vary in maturity across organisations, they can offer a useful starting point for shaping synthetic data governance. In this sense, they can function as both readiness markers and reference points to support ongoing decision-making and oversight throughout the synthetic data lifecycle. Many of the considerations explored later in this paper may be shaped or informed by the presence and maturity of these foundational components.

Pre-project considerations

2.9 Pre-project assessment can help organisations and practitioners to explore whether a synthetic data project is viable, valuable, and aligned with internal priorities. This stage is an opportunity to clarify the project's purpose, surface potential regulatory or ethical implications, ensure that stakeholders are engaged and governance pathways are understood. Decisions made on these pre-project considerations will shape how synthetic data is generated, consumed, and managed – impacting both downstream applications and the overall project value.

Purpose, value proposition and risk assessment

2.10 Synthetic data offers a solution to various challenges, such as privacy constraints or access to real-world data. Increasingly, it is expected that successful leveraging of synthetic data could allow firms to benefit from performance and efficiency gains. For example, the ability to augment existing datasets using synthetic data could expand their usefulness when designing and developing applications.

2.11 However, organisations may wish to consider whether synthetic data is the optimal solution for their needs or if other methods could address the problem more effectively. Evaluating the value of using synthetic data value requires an honest assessment of its potential benefits against inherent risks. For instance, while synthetic data may enable secure data sharing and innovation, it may also present unique challenges, such as measuring and managing potential bias propagation, or privacy vulnerability.

2.12 Not every project will benefit from the use of synthetic data, and the end use will determine the level of expected benefits and associated costs. By framing synthetic data initiatives around measurable outcomes and clear priorities, organisations can maximise their return on investment.

2.13 Insights from the SDEG highlight several actions that may be helpful for practitioners when assessing and managing the purpose, value, and risk of a synthetic data project.

2.14 **Define clear objectives and end-use purpose, including limitations:** Practitioners may benefit from establishing well-defined, measurable objectives for the use of synthetic data within a project from the outset. This includes articulating the specific problem that synthetic data is intended to solve, such as enhancing data availability, addressing privacy concerns, or enriching datasets for particular downstream applications. A clear statement of intent can help determine whether synthetic data is the right approach and can support alignment between technical choices, business needs, and governance expectations throughout the project lifecycle.

2.15 **Conduct a structured value-risk assessment:** Using synthetic data will involve trade-offs, such as balancing privacy with fidelity, or prioritising utility for downstream applications. Organisations may benefit from implementing a formal process to evaluate these trade-offs before taking a decision on starting synthetic data projects. Structured risk assessments such as data protection impact assessments (DPIAs) may also help inform the decision-making process. These approaches can also be strengthened by scenario testing to understand how different approaches to the potential trade-offs affect outcomes.

Regulatory, ethical and compliance considerations

2.16 SDEG members emphasised that the generation and use of synthetic data requires careful consideration of key regulatory, ethical, and compliance obligations. While practitioners are best placed to determine the specific obligations relevant to their context, SDEG members highlighted the following as key areas to consider at the outset of a synthetic data.

2.17 **Data protection and privacy compliance:** Data protection law will need to be considered when personal data is involved in the synthetic data generation process, as well as for the building of models that will go on to process personal data at deployment.

2.18 Key considerations will include determining the legal basis for any processing of personal data, applying data minimisation principles, and conducting Data Protection Impact Assessments (DPIAs) where required. When considering their data protection obligations, practitioners may need to consider the extent to which synthetic data carries a risk of possible reidentification, including testing for vulnerabilities which may not be readily apparent. They may also need to consider how to implement a 'data protection by design and default' approach to each stage of a synthetic data project.[5]

2.19 **Fairness, bias, and non-discrimination:** A range of legal, regulatory and ethical considerations may be relevant here. In particular, practitioners may need to identify, assess, and mitigate risks in relation to both the original and the synthetic data sets. This may include evaluating how different approaches to data generation could perpetuate or amplify bias. This may be a particularly salient consideration in a financial services context where synthetic data is being used to shape materially important decisions such as credit scoring or fraud detection. In such contexts, the Equality Act 2010 and the FCA's Consumer Duty may be of particular relevance in terms of regulatory considerations.

2.20 Establishing effective controls after conducting algorithmic reviews of models can also help identify the risks of a particular use case. Furthermore, internal assurance such as algorithmic audits can improve transparency and enable trust with users or parties that may be affected by the use of synthetic data in models.[6]

2.21 **Ethical use of data:** Practitioners may also consider the ethical implications around the approach to generating and using data, taking into account the principles of transparency, accountability, and a commitment to neutral or positive outcomes. This may involve considering the broader societal impacts of synthetic data applications.

2.22 **Oversight and accountability:** Organisations may also find it helpful to consider how their approach to synthetic data governance aligns with internal data management policies, including both oversight and accountability of the synthetic data project. This includes the maintenance of audit trails and being prepared to demonstrate compliance with relevant regulatory frameworks during supervisory reviews.

[5] For more information on data protection considerations relevant to synthetic data projects see the ICO's guidance.

[6] Auditing algorithms: the existing landscape, role of regulators and future outlook, DRCF, September 2022

2.23 By addressing regulatory, ethical, and compliance considerations at the outset, organisations can build a robust foundation for their synthetic data projects. This can support legal compliance, as well as enable firms to take into account ethical considerations. It may also play an important role in building internal confidence in the use of synthetic data.

2.24 To support practitioners in assessing and managing the risks outlined above, members of the SDEG have identified the following actions that may be helpful in practice:

2.25 **Engage relevant compliance, legal, data and AI ethics and risk functions** early in the project planning phase to ensure alignment with data protection and relevant sectoral laws, and to understand how to adhere to appropriate ethical practices.

2.26 **Document the outcomes of the review including:** identified regulatory risks, mitigation plans, and any regulatory reporting obligations to establish a clear compliance baseline before the project progresses. As discussed later in this paper, this sets the foundations for a continued, unbroken chain of documentation throughout the project.

Generation methodologies and downstream implications

2.27 Synthetic data generation methodologies are the central component to any synthetic data project. The decisions made on these methodologies, ranging from Generative Adversarial Networks (GANs) to agent-based models and non-AI methods such as distribution-based algorithms, influence the data's fidelity, utility, and privacy. Decisions made on the generation process from the outset will determine how effectively the synthetic data meets the requirements of its intended use case, as well as having a potential impact on data restrictions

2.28 SDEG members have highlighted that decisions on generation methodology are not purely technical: instead, they require balancing competing objectives from the outset. Fidelity and utility may be desirable for downstream model performance, but optimising metrics in these categories can increase privacy risks or entrench existing biases. Alternatively, prioritising privacy may limit the utility and fidelity of the dataset, particularly for complex applications requiring detailed data relationships.[7]

2.29 Additionally, practitioners will also need to consider the trade-offs of their chosen methodology with respect to computational demands, scalability, and alignment with the proposed end use case.

2.30 Beyond the methodology, it is also important to anticipate how these decisions impact downstream applications. Poorly selected methodologies can propagate bias, degrade model performance, or introduce ethical challenges, especially where outcomes are materially impactful. Addressing these considerations before the start of a synthetic data project, can help ensure that the synthetic data generated is reliable and confidence is built around its intended use.

2.31 Insights from SDEG members highlight that effectively managing the risks associated with synthetic data generation methodologies require a structured approach to

[7] Jordon, J. et al (2024) Synthetic Data – what, why, and how?

decision-making. This involves, prior to initiating a synthetic data project, adopting strategies that help to evaluate the suitability of different generation techniques based on technical, regulatory, and business considerations. Such approaches could include:

2.32 **A comparison matrix of synthetic data generation methodologies:** Comparison matrices can assist practitioners to systemically compare different synthetic data generation methodologies across key factors such as fidelity, utility, privacy, computational demands, regulatory compliance, and alignment with interned use case. By mapping out these factors across different techniques practitioners can better understand and communicate the trade-offs inherent to each approach and make informed decisions about which technique to use. Although comparison matrices can provide guidelines, individual implementations may vary and the utilisation of a specific approach to generation does not guarantee an exact level of fidelity, utility or privacy.

Stakeholder alignment and governance readiness

2.33 Synthetic data projects by their nature will inherently require multidisciplinary teams, requiring input and collaboration from technical individuals, legal advisors, compliance experts, data and AI ethics specialists, and business leaders. Organisations that are new to using synthetic data or seeking to generate synthetic data for the first time may lack the structures needed to convene these diverse stakeholders to support specific synthetic data governance.

2.34 Examining the adequacy of existing governance arrangements can help organisations to ensure 'governance readiness'. This may involve identifying limitations with current governance processes, or areas where changes may be needed, to accommodate the additional regulatory and decision-making considerations required to manage the risks relating to synthetic data initiatives.

Chapter 3

Generating synthetic data

3.1 The process of generating synthetic data involves a range of design choices and judgement calls that can have significant implications for governance, risk, and downstream use. This section explores three areas that members of the SDEG identified as particularly relevant during the generation phase: auditability, privacy, and bias. While these issues are not exclusive to generation, early and deliberate consideration of them during this phase can help practitioners make more informed design decisions and better manage risks across the synthetic data lifecycle.

Auditability of the generation process and outputs

3.2 Auditability is the capacity to systematically track, verify and validate processes and decisions relating to data throughout the project's lifecycle. It enables practitioners to continuously monitor whether data and algorithms are being used within the parameters and purposes which were defined and agreed. Datasets and models used in any data process, including synthetic data generation, need to be reliable, secure, and compliant with relevant laws and regulations.

3.3 Both data and the algorithms that generate or use synthetic data can be audited to ensure better and more transparent outcomes. Practitioners can check governance documentation, test the outputs of a model, or unpack the model to understand how it works. Audits enable assurance, can improve trust, enable transparency, and facilitate traceability and compliance with regulatory expectations.[8]

3.4 Synthetic data poses a unique challenge to traditional auditability by increasing the number of judgement-based decisions, with potentially significant downstream impacts. In this context, effective auditability requires maintaining a clear, unbroken chain of evidence for every decision, transformation, and validation through the lifecycle of synthetic data, from pre-project considerations to model deployment. This includes the criteria for data generation, the assumptions underpinning synthetic data generating models, the validation techniques employed.

3.5 This sub-section explores more explicitly the unique auditability considerations for synthetic data. It encompasses the topics of clear ownership, continuous governances, and transparency to ensure that errors or oversights are not propagated for downstream applications.

Establishing ownership and use models

3.6 Clear ownership models across a synthetic data initiative can support an organisation to outline responsibilities, track data provenance, and support the auditability of the generation process and its outputs. There are different ways in which this can be

[8] Auditing algorithms: the existing landscape, role of regulators and future outlook, DRCF, September 2022

achieved. In particular, ownership structures may vary depending on the organisational context. Examples of clear ownership models include:

3.7 **Distributed Ownership:** Shared responsibilities across multiple teams or departments, with clear communication channels to prevent mismanagement and misalignment throughout the synthetic data lifecycle. Clarity around roles and clear communication around expectations could be communicated both in advance of, and during the lifecycle of the project. Documenting these factors can also support in the event of changes experienced by project teams throughout the synthetic data lifecycle.

3.8 **Vendor-Specific Ownership:** Where third-party providers manage certain aspects of synthetic data generation or validation, clear contracts can establish ownership and proper oversight mechanisms. These contractual arrangements will help support organisations to better understand the responsibilities for managing risks as synthetic data is generated.

3.9 **Usage-Based Rights:** Where ownership is tied to specific use cases, projects, or permitted purposes, practitioners and third parties may benefit from drawing up well-defined usage agreements to prevent unauthorised application and usage. It may also be relevant to consider the time-bound nature of these agreements, as synthetic data may only reflect the statistical properties of real-world data at the point of its generation.

3.10 Ownership models are likely to differ across organisations and even perhaps across different synthetic data projects. SDEG member insight suggests that practitioners can benefit from establishing a clear ownership structure at the outset of a synthetic data project. This might include role allocation for the generation, validation, and oversight of synthetic data and should marry ownership models with organisational maturity and operational models. Aligning these ownership models with existing data governance policies within the organisation may also prevent inconsistencies in synthetic data governance.

Governance stages across the lifecycle

3.11 It is the view of SDEG members that synthetic data governance and approaches to auditability should not be treated as a one-time activity but a continuous process throughout the synthetic data and model lifecycle. Considering how to structure governance across key stages in the lifecycle can ensure that governance checkpoints, risks reviews and compliance assessment are conducted at appropriate points. As a baseline, SDEG members have highlighted the following key stages as potential governance checkpoints:

3.12 **Data collection and preprocessing:**

Governance at this stage may comprise:

- Ensuring that real-world datasets used in synthetic data generation are accurately sourced, cleaned, and validated. This includes removing errors, verifying statistical representativeness, and ensuring data quality.

- Documenting, justifying, and reviewing the expert knowledge used to define behavioural rules, where synthetic data is not generated from an existing dataset.

3.13 **Synthetic data generation and validation:** Governance at this stage may include validating that the synthetic data meets predefined quality, fidelity, and privacy thresholds through benchmarking tests, privacy assessments and statistical comparison.

3.14 **Model training, testing, and deployment:** Governance at this stage may include:

- Assessing model performance when trained on synthetic data, ensuring it consistently generalises well to real-world scenarios.
- Monitoring and documenting both the model and data used in order to maintain a chain of evidence regarding their performance, suitability for the end use case, limitations of use, including advisories on appropriate and inappropriate use.

3.15 Experience from SDEG members suggests that implementing 'stage gates' at critical transitions in a synthetic data project is an effective way to ensure that generation models meet predefined thresholds for fairness, accuracy, and security. Some of these potential 'stage gates' can include:

3.16 **Design to Testing:** Ensuring that generation methodologies align with the project's objectives and that the intended privacy, fairness, and ethical measures are embedded.

3.17 **Testing to Production:** Validating that synthetic data improves model performance without introducing unintended risks, and that all relevant regulatory approvals or internal reviews have been conducted.

Using metadata and tools to support auditability

3.18 Ensuring the auditability of synthetic data projects may require more than defined ownership structures and governance frameworks. These elements may benefit from the application of appropriate tools and expertise to help create a transparent, traceable, and well-documented audit process. Without these additional factors, practitioners may risk opacity in decision-making processes, which can pose further challenges to demonstrating compliance and building confidence in the synthetic data's integrity.

3.19 Given the complexity of the synthetic data lifecycle, reliance on manual documentation can also risk oversight, misalignment, and gaps in governance. Instead, SDEG members suggest that tools such as provenance tracking, data version control, and MLflow reproducibility can enable practitioners to track changes, log quality metrics, and detect issues such as data drift or bias.

3.20 Metadata and documentation can also provide critical context to support synthetic data auditability. Without records of the dataset's origin, purpose, and generation methods, practitioners may risk making uninformed decisions that undermine the data's utility or integrity.

3.21 SDEG members identified the following as key factors to document in metadata to support auditability:

3.22 **The original purpose of the synthetic data set:** Including all assumptions associated.

3.23 **The ownership and responsibility of people including:** all persons who are responsible for the various aspects of the synthetic data generation process, and the ownership of the generation models, synthetic data, and model consuming synthetic data.

3.24 **The methods used to generate the synthetic dataset including:** the transformations and augmentations made to the dataset and the rationale for the decision.

3.25 **The known limitations including:** any excluded features or trade-offs made when generating the original dataset, and why these decisions were made.

Mitigating privacy risks in synthetic data

3.26 Synthetic data can enable organisations to securely utilise and share datasets while mitigating privacy risks associated with the original data. This capability is particularly valuable to financial services where data sensitivity and privacy concerns are critical. However, while synthetic data significantly reduces privacy risks, it may not eliminate them entirely.

Understanding synthetic data privacy risks

3.27 SDEG members identified the following as key considerations in understanding the privacy risks associated with a synthetic data project; beginning with classifying the synthetic dataset based on how it was generated and then considering its intended use case.

3.28 **Generation methodology:** Synthetic data generated by replicating real-world patterns and relationships often carries higher privacy risks due to the potential for re-identification or correlation with real data. Conversely, synthetic data created from aggregated or anonymised statistics may present lower risks but may lack the fidelity required for certain use cases.

3.29 **End-purpose and trade-offs:** Various trade-offs will be made in light of the primary objective of the synthetic data, i.e. whether it is designed to maximise privacy, utility, or fidelity. Clear prioritisation ensures that the dataset aligns with the organisation's goals and provides an outline for the potential privacy risks associated with the dataset and its end use.

3.30 **Assessing the data environment:** The privacy risks associated with a synthetic data project are not static and will also depend on how and where the data is consumed. For example, synthetic data intended for internal testing or development may already be subject to restricted access or controlled environments. Whereas data shared externally with third-party vendors or developers may require consideration of increased privacy risks and additional safeguards. The nature of the privacy risks and adequacy of safeguards will need to be considered on a case-by-case basis; including whether organisations should restrict access, limit exports, and monitor usage in secure environments.

3.31 **Use case specific risks:** There may also be context-specific risks to consider in connection with the intended end use. In financial services for example, data may be highly sensitive and valuable for external adversaries, increasing the need for security and safeguards.

3.32 Maintaining detailed documentation and records can establish a foundation for assessing, classifying, and managing privacy risks throughout the synthetic data lifecycle, ensuring that privacy risk classification and mitigation strategies can be continuously reviewed.

3.33 A Data Protection Impact Assessment[9] can provide a structured way to identify, assess, and mitigate data protection risks, and may be a requirement under data protection law. This can cover off privacy risks in the data generation process, risks around re-identification, and how the overall approach aligns with data protection by design principles, including in consideration of the end-use of the data.

3.34 Aside from personal data, organisations may also need to consider the risk of synthetic data revealing trade or intellectual property secrets through reverse engineering techniques. This may be particularly relevant in a financial services context.

3.35 In particular, SDEG members highlighted the importance of keeping detailed documentation logs and records of:

- **The privacy risks** associated with the synthetic data generation process (including the underlying privacy risks connected to any real data used as part of the process).
- **Any privacy implications** arising from the synthetic data's purpose and end use.
- **The trade-offs** made between privacy, utility, and fidelity and why these decisions were made in relation to their end use case.
- **The consumption environment** of the data, and how the privacy measures align with the data's intended use and the privacy risks associated with that use.

Privacy testing for synthetic data

3.36 Validating synthetic datasets for privacy robustness is widely recognised as a critical yet complex process. The absence of a universal privacy metric applicable across different datasets, generation methods, and use cases further complicates this process. In response to this challenge, SDEG members have highlighted the importance of using a suite of techniques to evaluate privacy, including (but not limited to):

[9] DPIA

3.37 **Red teaming and privacy testing:** Adopting adversarial testing methods, such as red teaming[10], to identify potential vulnerabilities in the synthetic data set. These exercises are most effective when they simulate a wide range of potential attacks (eg, membership inference attacks) to assess the risk of reidentification or privacy leakage.

3.38 **Combining privacy metrics to holistically understand privacy risks:** Given there is no universal metric for measuring privacy, combining multiple metrics can help organisations to better understand the risks associated with the synthetic dataset. Below are two, non-exhaustive examples of privacy metrics:

3.39 **K-Anonymity and L-Diversity:** Measures to ensure that individual records cannot be easily distinguished from others (outlier protection testing).

3.40 **Aggregate risk analysis:** To identify whether patterns in synthetic data could reveal sensitive information through correlation with other (external or publicly available) data sources.

3.41 Thorough testing of synthetic datasets can help to provide assurances surrounding the privacy risks of the dataset and ensure that this aligns with the end use case and risk appetite of the organisation. Using a combination of privacy metrics and adversarial testing can help to provide assurances to validate the synthetic dataset's privacy. If practitioners have concerns surrounding re-identification, SDEG members suggest that they can consider introducing noise into the dataset – otherwise known as deferential privacy – to obscure sensitive details. It should be noted that this approach carries the extra burden of impacting the utility and fidelity of the dataset. Where feasible, organisations may also consider engaging with independent expert reviewers to enhance the credibility of the privacy assurances.

3.42 SDEG members also highlighted that privacy risks evolve over time. Practitioners can therefore benefit from remaining alert to how these risks may impact the assurances related to synthetic datasets. These effort can be further supported by establishing mechanisms for continuous monitoring or periodic review and re-assessment of privacy metrics and assumptions relating to any synthetic data.

Managing bias when generating synthetic data

3.43 Bias is a multifaceted and persistent challenge in both real-world data and synthetic data generation. All data contains some degree of bias, and not all forms of bias are harmful or undesirable. For example, bias can increase efficiency in tasks, by streamlining choices and relying on past experiences or learned patterns. In the context of financial services, models and algorithms that show bias towards identifying a fraudulent transaction can help flag potential fraud quickly, which may protect consumers.

[10] What is red teaming? (2024) IBM

3.44 On the other hand, bias can be harmful if it leads to unfair treatment or discrimination against certain groups, including those with protected characteristics. As such, bias in datasets used to train models may perpetuate existing biases and stereotypes. However, synthetic data offers an opportunity to mitigate or entrench biases depending on how it is generated, manipulated, and consumed.

3.45 Organisations and practitioners will have a range of legal, regulatory and ethical considerations to contemplate when considering the risks around bias, including the risk of unfavourable or harmful outcomes. Bias could have particularly harmful consequences in a financial services context.

3.46 Synthetic data introduces unique challenges because it is not limited to being a reflection of real-world data but often involves deliberate decisions regarding bias mitigation or amplification. Essentially, the decisions made when generating synthetic data must be aligned with the dataset's intended use case, its end-user context, and broader legal, regulatory and ethical considerations. Without careful consideration, synthetic data risks entrenching biases, introducing new unforeseen biases, and creating unintended downstream consequences.

3.47 To effectively address bias, it is necessary to examine the synthetic data lifecycle holistically, from the initial evaluation of the real-world data (if applicable) to the decisions made during generation, manipulation, and iterative testing. There are likely to be trade-offs between technical accuracy and ethical considerations when transforming data. In addition, robust mechanisms can be used to measure and articulate bias, track its evolution, and iteratively refine processes to address unforeseen issues.

Identifying bias in source data

3.48 A key consideration raised by members of the SDEG is the need to identify potential biases in any source data used to generate synthetic data. This includes understanding the implications of the data's provenance, its representativeness, and any pre-existing imbalances or historical inequalities. SDEG members identified the following as key considerations for understanding the risk of bias in source data.

3.49 **Data provenance:** This involves setting out to answer and document the data's provenance, including determining whether the real-world data is a sample of a larger dataset. Where possible it also includes understanding the collection methods used to gather the data, the intended purpose behind the data's collection, and whether it accurately reflects the population or the context it is assumed to represent.

3.50 **Sampling bias:** This involves interrogating the data to understand if certain behaviours, groups, or events are underrepresented or overrepresented in the real-world data set. In financial services, this can have implications for consumers or rare events such as fraudulent transactions. Understanding how these events or groups are represented in the data is essential in understanding what biases may be translated or entrenched as part of the generation process.

3.51 **Problem framing:** This involves considering the possibility that seemingly unrelated characteristics may be close proxies for demographic characteristics or characteristics of vulnerability. This may especially be the case where sophisticated predictive algorithms consume synthetic data, for example in supervised machine learning models. Examples of characteristics which may constitute, or be adjacent to, demographic characteristics, include location proxies and salary bandings.

3.52 **Propagation of bias in the collection process:** In certain circumstances, a dataset can propagate existing social biases, or biases that are apparent in the data collection process itself.

3.53 SDEG contributors have highlighted that formal bias audits on source data can help to identify and mitigate potential risks such as entrenching or amplifying biases. Such audits typically consider the following:

3.54 **Assessing data representativeness:** Evaluating whether the source dataset accurately reflects the target population or whether certain groups or events are systematically underrepresented or overrepresented.

3.55 **Identifying sampling distortions:** Determining if the data has been skewed due to the way it was collected (eg, historical exclusions, selection biases).

3.56 **Tracking demographic proxies:** Identifying whether seemingly neutral characteristics (such as postcode, occupation, or browsing behaviour) serve as indirect proxies for sensitive attributes such as race, gender, or socioeconomic status.

Bias manipulation during generation

3.57 Once the biases in the source data are identified, understood, and documented, practitioners are better placed to make informed decisions about how to address these biases during the synthetic data generation process. Crucially, decisions made to address the biases in the data set cannot remove them entirely, but decisions on end-use case and purpose can help to steer how bias should be manipulated in the generation process. Again, SDEG members have articulated a number of options:

3.58 **Observe and accept biases:** Practitioners may consider it appropriate to accept the existence of biases in the real-world data and to decline to alter the data. For example where practitioners haven't identified any significant ethical concerns or due to potential downstream impacts on a model's performance, or in cases where alternative safeguards can be used to manage risks.

3.59 **Mitigating identified biases:** Practitioners may wish to address existing biases in the dataset during the generation phase either by removing or reducing specific biases that were identified as present in the underlying real-world data. This may be particularly important, for example, when historical datasets exhibit biases based on previous decision making that has been unfavourable or undesirable towards marginalised communities.

3.60 **Introducing controlled biases:** Practitioners may consider how they wish to intentionally add bias into a data set to improve the representation of rare events of interest. This can help to improve the synthetic dataset utility for downstream tasks like training models for fraud detection.

3.61 **Assumptions in agent-based modelling:** When synthetic data is generated through agent-based models instead of inputting real-world data, the behavioural assumptions built into the generation model have the potential to embed or amplify harmful biases. It should be emphasised that assumptions built into these models stem from explicit rule design, rather than latent data patterns. Managing bias in this context therefore requires careful evaluation of built-in behavioural assumptions.

3.62 **Trade-offs between bias, utility, and fidelity:** Any changes to biases made during the synthetic generation process may impact the fidelity or utility of the synthetic data generated. For example, reducing bias in the synthetic data generation process may make the data less representative of real-world conditions, which could later impact a model's performance if trained or validated on synthetic data.

3.63 SDEG contributors emphasised the value of documenting the various decisions taken when generating a synthetic data set, including considerations around their downstream impacts and any trade-offs made. Altering biases in one area may also inadvertently introduce bias elsewhere, and maintaining clear records of such decisions can help provide assurances regarding the dataset's intended purpose and ethical use.

3.64 Once biases in the source data have been identified and assessed, practitioners can clearly define their objectives for bias mitigation during synthetic data generation. Clear bias objectives avoid arbitrary alterations during generation or misalignment with the intended use of synthetic data.

3.65 Effective objectives should also take into consideration the intended end-use case, as well as relevant ethical considerations and regulatory requirements. Key steps may include:

3.66 **Determining the primary goal of bias manipulation:** Whether actions are taken to reduce, preserve, or balance existing biases.

3.67 **Assessing the ethical and operational impacts:** For example, determining whether historical biases should be mitigated to ensure fairer decisions, or whether they should be maintained to reflect real-world financial risk distribution

3.68 **Documenting bias-handling decisions:** Recording justifications for any changes to bias levels, including the rationale behind bias reduction, introduction, or preservation.

Measuring and testing for bias in synthetic data

3.69 Evaluating potential bias post-generation is another area of focus among practitioners. While there are no universal tests for bias, SDEG contributors have highlighted the benefit of combining quantitative methods with the qualitative experiences of subject matter experts to consider the presence of bias and how bias aligns with the intended use case.

3.70 SDEG members highlight that various fairness metrics can be used to understand the biases in data sets. The appropriate metric is likely to be context specific and depend on data availability, project objectives, and the nature of the model in development. Examples of common fairness metrics include:

3.71 **Demographic parity:** Evaluates whether different groups (eg, gender, ethnicity, or socioeconomic status) receive similar outcomes within the synthetic dataset.

3.72 **Disparate impact:** Measures whether synthetic data preserves disparities found in real-world data, often used in financial decision-making applications such as credit approvals.

3.73 However, SDEG participants have also noted that fairness metrics alone may not capture the full picture. Subject matter experts with relevant domain knowledge can provide critical qualitative insights to support the interpretation of results in context. These insights can help inform more nuanced discussions on the impacts of the existing bias and provide considerations for possible further iteration of the synthetic dataset.

Iterative bias management through generation process

3.74 In practice, it is unlikely that practitioners will generate an ideal synthetic dataset on their first attempt. SDEG contributors noted that testing outputs throughout the generation process can help practitioners to make iterative adjustments to their generation process. This in turn ensures that the biases in the synthetic dataset are desired and suitable for their end use case.

3.75 By systematically testing synthetic datasets and refining the generation process, practitioners can ensure that undesirable biases are mitigated rather than unintentionally reinforced. They can also ensure that bias trade-offs are clearly understood (eg, in relation to model performance and ethical considerations), and that synthetic datasets remain suitable over time, as market conditions or real-world data distributions evolve.

3.76 SDEG participants also emphasised the importance of periodically re-evaluating bias metrics as approaches to synthetic data generation evolve. The relevance of a given fairness metric may vary depending on the context and application, while longer-term use of synthetic datasets may require intermittent reviews to ensure that fairness benchmarks remain valid.

3.77 SDEG members view bias evaluation and correction not as a one-off process, but as a continuous cycle. The use of feedback loops at key points in the synthetic data lifecycle can support ongoing reassessment of bias and identify when an adjustment to the generation approach may be appropriate. For example, if a model trained on synthetic data demonstrates poor generalisation, this could trigger the need to review the synthetic dataset's generation process.

3.78 Additionally, SDEG contributors also highlighted the value of cross-functional reviews, ensuring that colleagues with legal, compliance, risk, and ethical expertise are consulted as part of the ongoing review process. To maintain accountability and consistency, cross-functional teams comprising of technical experts, compliance and legal professionals, and business stakeholders can opt to engage in regular reviews of bias to effectively evaluate bias in the synthetic data. These technical governance forums can then ensure that the approach to bias management aligns with legal, ethical and strategic priorities.

3.79 A structured documentation process as part of such technical governance forums can be useful to support transparency and accountability. This may include: bias specification (detailing which biases are being monitored and why), bias measurement techniques (the metrics used to quantify bias and their rationale), and decision rationale (trade-offs made between utility, fairness, and model performance). Maintaining transparent, auditable records, can support firms and practitioners to explain and justify their approaches.

Chapter 4

Using synthetic data in models

4.1 Synthetic data is primarily used as a tool within machine learning (ML) workflows to support a range of data-driven activities. Depending on the use case, it may be used as a training input to supplement or replace real-world data, as a substitute for testing or validation where access to real data is limited, or as a means of simulation and stress testing under controlled or hypothetical scenarios.

4.2 Each of these applications introduces potential benefits, such as enhanced data availability, improved privacy, or greater flexibility in experimentation. However, the use of synthetic data also requires careful validation to ensure it supports the development of fair, robust, and reliable models. Without appropriate checks, synthetic data can introduce new risks or reinforce existing ones, particularly around generalisability, bias, and performance in real-world conditions.

Evaluating synthetic data quality

4.3 Post-generation, practitioners may seek to ensure that synthetic data meets a quality standard suitable for downstream applications such as machine learning, analytics, and model validation. Depending on the context, synthetic data may be used to support model training, supplement validation datasets, or act as a test substitute where access to real data is constrained. While statistical resemblance to real data can provide a foundation, task-specific assessments are often better at determining the quality of synthetic data for its intended use. In this context, quality is not a fixed property but determined relative to its end application.

4.4 This section outlines approaches that may help assess the quality of synthetic datasets. While not exhaustive, these methods may serve as a useful starting point for practitioners seeking to understand the limitation and capabilities of a synthetic dataset for an intended use case.

High-level statistical comparison

4.5 High-level statistical comparisons can provide an initial assessment of how closely synthetic data resembles a real-world dataset. This comparison typically focuses on examining the univariate and multivariate distributions as well as the overall correlation structure of the data. Evaluating these attributes can quickly identify inconsistencies or errors that may have arisen during the synthetic data generation process.

4.6 These methods are useful for providing a computationally efficient and cost-effective starting point to identify common mistakes such as incorrect data type assumptions, or issues with variable scaling.

4.7 However, statistical similarity does not always equate to downstream performance. These comparisons may fail to capture deeper multidimensional relationships or more nuanced patterns, particularly in high-dimensional datasets, such as transactions data in financial services. In these scenarios, high-level statistical comparison may become impractical and potentially necessitate task-specific or more scalable evaluation techniques for quality assurance in a modelling context.

Model-driven evaluation techniques:

4.8 To better evaluate synthetic data quality, SDEG members have suggested practitioners explore model-based techniques such as performance benchmarking and Train-Synthetic-Test-Real (TSTR). While these methods are commonly associated with assessing a model's response to synthetic data – a topic that is explored in paragraphs 4.39 and 4.40 – they can also offer valuable insight into synthetic data quality.

4.9 If a model trained or tested on synthetic data performs similarly to one using real data, it may suggest that the synthetic dataset preserves meaningful feature relationships and supports generalisable learning. This approach to measuring synthetic data quality may be particularly relevant when synthetic data is intended for model training or validation.

Performance benchmarking for quality evaluation:

4.10 Performance benchmarking involves training and testing a model on synthetic data and comparing metrics – such as accuracy, precision, or recall – against real data baselines. While real data is not without limitations it is generally a more widely accepted starting point for an analysis task.

4.11 When performance metrics are comparable, it may indicate that the synthetic dataset captures the underlying patterns required for generalisable learning. Conversely, significant performance discrepancies may suggest synthetic data quality issues, such as missing use cases or simplified feature interactions. These divergences not only affect model outputs but signal where the synthetic data might lack fidelity or task relevance. In this sense, the performance metrics may serve as an indirect indicator of synthetic data adequacy for its intended use case.

4.12 However, practitioners may need to exercise caution when benchmarking performance to infer synthetic data quality, particularly to prevent synthetic data from inadvertently leaking into validation or test datasets. This risk is heightened by the added complexity of using both real and synthetic data, which can result in artificially inflated performance metrics. SDEG members have noted the importance of investigating unexplained improvements in performance when evaluating synthetic data, which may indicate issues such as overfitting or lower fidelity in the synthetic dataset. They have also noted that if synthetic and real datasets share overlapping records or similar seeds, data leakage can occur through memorisation, further distorting evaluations results.

Train-synthetic-test-real (TSTR) evaluation

4.13 An alternative approach to testing the quality of synthetic data is Train-Synthetic-Test-Real methodology. This approach is particularly useful for scenarios where the synthetic data is intended to support predictive models or complex analysis. This approach provides practitioners with more insight into the quality of the synthetic data as part of a dynamic operational context.

4.14 By training a machine learning model on synthetic data and comparing its performance to a model trained on real data, practitioners can assess how well the synthetic data retains the critical information needed for specific use cases. This evaluation directly measures the utility of synthetic data in its intended application, making it a more reliable indicator of quality than high-level statistical comparisons alone.

4.15 However, as with benchmarking, TSTR results require careful interpretation. Practitioners may wish to consider whether real data selected for testing is itself biased or incomplete, to understand whether results may be misleading. Similarly, practitioners can consider whether the synthetic data misrepresents real-world feature interactions and therefore risks masking generalisation areas despite appearing statistically sound.

Using performance feedback to guide iterative approaches to generation

4.16 SDEG members have highlighted that evaluating synthetic data quality is not a static process, but a continuous and iterative task. Feedback from downstream model performance can support practitioners in identifying the limitations of synthetic data and take decisions on how to address these limitations in the generation process.

4.17 However, when drawing on performance metrics for iteration, practitioners may need to balance different priorities and consider trade-offs across fidelity, utility, and task-specific performance. A key consideration is the intent behind refinement. For example, practitioners may aim to improve model performance on a specific task, by tailoring synthetic data to better represent known gaps, or enhance the representational accuracy of synthetic data to better reflect the statistical characterisers of real-world data.

4.18 These goals are not always aligned. Effective iteration requires a clear understanding of both the purpose for refinement as well the resulting trade-offs between fidelity, utility, and fairness in any given use case.

4.19 Given these trade-offs, context-specific evaluations form an important part of the decision-making process. The level of fidelity, complexity, and privacy required from synthetic data will vary depending on the intended use case, and practitioners may need to tailor generation strategies accordingly.

4.20 For instance, synthetic data used in fraud detection models will likely need to accurately reflect rare events and complex correlations, while data for internal testing may require less precision but still need to align with the structural characteristics of the original dataset.

4.21 To support consistency in using performance feedback to inform generation decisions, practitioners may wish to implement structured feedback loops that include:

4.22 **Pre-defined criteria** for when performance feedback should trigger generation refinements (eg, significant divergence between synthetic-trained and real-tested models).

4.23 **Mechanisms to distinguish** between genuine performance improvements and model overfitting.

4.24 **Collaboration with domain experts** to interpret performance patterns in context and validate whether changes align with the intended use case.

4.25 Transparency and documentation also play a key role in this iterative process. Clear records of evaluation results, decision-making processes, and adjustments made during refinement provide auditability and enable effective communication among stakeholders. This documentation may also help ensure that the synthetic data generation process is both rigorous and defensible and support organisations in evidencing compliance with applicable regulatory and ethical standards.

Assessing the impact of synthetic data on models

4.26 Once synthetic data has been generated and assessed for statistical and task specific quality, its integration into machine learning workflows introduces additional questions concerning how it may influence model performance, behaviour, and fairness. Unlike real data, synthetic datasets are usually constructed to serve a specific goal, such as privacy preservation, bias mitigation, or rare event modelling. SDEG members note that these attributes can impact and shape the model learning dynamics which may require further consideration and careful evaluation by practitioners.

4.27 This section focuses on how synthetic data can impact models during training, validation, and testing. As a starting point, SDEG members have highlighted the following interrelated factors that may determine the impact of synthetic data on models:

4.28 **The model's purpose and success criteria:** This is particularly relevant in use cases like fraud detection or credit scoring, which often involve large class imbalances and require high precision.

4.29 **The specific objectives behind using synthetic data, such as:** addressing privacy concerns, compensating for censored or incomplete data, or augmenting datasets with imbalanced class distributions. While these decisions do not directly determine model quality, they shape the trade-offs practitioners will need to consider when evaluating fidelity, utility, and risk.

4.30 **The composition of the dataset:** Whether fully synthetic or a mixture of real and synthetic data, and the respective proportions of each in training, testing, and validation sets. The balance between synthetic and real data can significantly influence model fidelity and the applicability of insights derived from the model. These risks may be most pronounced when synthetic data serves as the sole or primary input for training or when synthetic data is used alongside real data and it may be difficult to determine the relative contributions of each dataset. In this latter situation, practitioners may face the additional risk of de-anonymising the synthetic dataset by combining it with additional real-world data.

4.31 The impact of synthetic data on model quality remains a key area of focus, particularly in relation to maintaining fidelity, mitigating risks, and ensuring alignment with the model's intended use. SDEG members have highlighted that adopting structured evaluation frameworks may support more consistent assessments- helping organisations to identify potential limitations and understand how synthetic data integration affects accuracy, fairness, and utility across use cases.

Dataset composition and proportions:

4.32 The composition and proportions of synthetic data relative to real data across training, validating, and testing stages plays a pivotal role in determining model quality. Depending on how synthetic data is integrated, it may impact learning dynamics, evaluation accuracy, and generalisation of models. SDEG contributors emphasised that careful evaluation of these factors can help practitioners to better understand synthetic data's impact on model performance. In particular, practitioners may wish to consider:

4.33 **Proportion of synthetic data in training data:** The proportion of synthetic data in a model's training dataset can have a substantial impact on model learning. If synthetic data dominates, there is a risk the model may learn artefacts of the synthetic generation process, such as simplified feature relationships rather than generalisable real-world patterns. This may be particularly relevant when synthetic data lacks edge cases, rare events, or reflects outdated correlations in non-stationary environments. If synthetic data does not reflect these evolving patterns, the model may appear to generalise better than it does in practice. SDEG members have suggested that monitoring such trends and adjusting generation methods accordingly can help to improve approaches to model training.

4.34 **Training dataset mix by use case:** In some contexts, involving small or imbalanced datasets, synthetic data may be used to supplement or rebalance training data. This is especially relevant in models for fraud detection, where enriching rare-event classes can improve model performance. These approaches benefit from careful consideration of use-case sensitivity, as excessive augmentation can introduce overfitting or lead to inflated performance metrics.

4.35 **Alignment across datasets:** Validation and testing phases typically require high-fidelity data that closely mirror the model's deployment environment. If synthetic data is used for testing - or if the real data used for validation is misaligned with synthetic training data – a model's performance metrics may be misleading. For example, a model may appear to perform well on synthetic data while failing to generalise to real-world distributions. In such cases, the synthetic training data may not capture key correlations and reduce a model's real-world performance.

Diagnostic tools to evaluate synthetic data model impact

4.36 Once synthetic data has been incorporated into the model lifecycle, practitioners may wish to go beyond statistical assessments or pre-training benchmarks to evaluate how it impacts model performance. SDEG members have highlighted a range of approaches to assess how synthetic data influences a model's performance, decision boundaries, and fairness. These approaches can help practitioners to distinguish between performance gains to ensure they are meaningful.

Performance benchmarking

4.37 Benchmarking performance remains relevant once a model has been trained or validated on synthetic data, especially when comparing performance against a baseline trained on real-world data. In this context, benchmarking is used to assess whether synthetic data has meaningfully improved or reduced model performance. While similar to evaluating the quality of synthetic data, this approach focuses on explaining the model behaviour.

4.38 Practitioners may benefit from using benchmarking to detect anomalies such as inflated metrics. For example, an increase in accuracy or AUC may mask an over-reliance on artefacts introduced by synthetic data generation. SDEG members emphasis that careful interpretation is required to determine whether performance improvements represent model gains or reflect superficial alignment with the synthetic data used in training.

Train-synthetic-test-real

4.39 The Train-synthetic-test-real (TSTR) method can also provide insights into how synthetic training data affects a model's ability to generalise to real-world conditions. If a model trained on synthetic data underperforms when tested on real data, it may suggest that the synthetic dataset omits critical edge cases or overrepresents simplified relationships.

4.40 This approach is particularly helpful in uncovering mismatches between synthetic training data and real-world validation distributions. While similar results may be used to inform improvements to data generation (as discussed in paragraph 4.13-4.15), here the primary concern is understanding whether synthetic data leads to misleading performance signals or behavioural drift when deployed in live environments.

Explainability and feature sensitivity

4.41 Given the inherent uncertainties associated with synthetic data, another key area of focus in understanding synthetic data's impact on models is examining feature importance and sensitivity across real and synthetic datasets. When synthetic data is primarily used to compensate for a lack of data volume, there is an expectation that the model will identify and weigh features consistently between datasets. However, disparities in feature importance or sensitivity may signal issues with data fidelity, which could compromise model performance and fairness.

4.42 SDEG contributors have noted that in some cases, models trained on synthetic data appear to rely more heavily on a narrow subset of features with disproportionately high importance. Such reliance may indicate that the synthetic data fails to capture the diversity or complexity of the original dataset. In turn, this reliance can distort model logic, introduce brittleness, or amplify unintended biases.

4.43 In these cases, engaging domain experts to review the model's feature importance and sensitivity can help determine whether these differences present concerns or are consistent with expectations for the specific use case.

4.44 In scenarios where synthetic data is used to correct bias in real-world data or improves fairness of model outputs, explainability may become increasingly important. For example, in credit scoring applications, biases in historical loan approvals might be mitigated by adjusting synthetic data distributions to better reflect broader applicant demographics. However, if these adjustments distort key relationships between creditworthiness indicators, the model's predictive accuracy and fairness could be affected. Moreover, adjusting synthetic distributions to improve fairness could alter causal dependencies, potentially degrading model utility in production. In this case and similar instances, practitioners may need to evaluate whether synthetic data undermines the trust in and operational utility of the model in question. Ensuring that synthetic data maintains meaningful, explainable relationships between features is critical for its responsible use.

4.45 By combining these diagnostic tools – performance benchmarking, TSTR, and feature explainability – practitioners can build a more complete picture of how synthetic data affects model trustworthiness and performance. These insights can help identify whether performance improvements are reliable, fair, and aligned with the model's intended use. SDEG members emphasise that such evaluations support more responsible integration of synthetic data into machine learning pipelines and help ensure that synthetic data interventions do not introduce unseen risks.

Chapter 5

Conclusion: building confidence in synthetic data

5.1 The adoption of synthetic data across financial services offers a potentially powerful means of enabling innovation, expanding access to data, and strengthening privacy and security. Yet these benefits can only be realised if organisations and practitioners have confidence in the quality, safety, and suitability of the synthetic data they generate and deploy. This confidence cannot be assumed, but it can be fostered by technical rigour and safeguarded by supportive governance.

5.2 Throughout this paper, we have outlined some of the governance considerations and technical foundations that can benefit practitioners to responsibly deploy synthetic data. From early-stage project scoping to model integration, we used SDEG member experiences to explore governance considerations and the associated technical practices across the synthetic data lifecycle.

5.3 In doing so, we have shown that synthetic data quality is not an abstract or standalone property. It is inherently relational: its usefulness can be assessed in relation to the use case it is expected to serve.

5.4 Building confidence in the use of synthetic data is not solely a technical challenge. It is a governance challenge- one that requires ownership, documentation, interdisciplinary collaboration, and robust validation procedures. Confidence is not achieved through technical perfection. Instead, it is built through transparency, consistency, and a shared understanding of how synthetic data interacts with the ethical, operational, and analytical goals of the project.

5.5 This approach is reflected in the FCA's ambition to develop proofs of concept that explore real-world applications, such as our collaborative Anti-Money Laundering and Synthetic Data Project. Projects like this help illuminate the practical challenges and context-specific nuances to further inform approaches to technical foundations and context-specific governance.

5.6 Confidence also depends on the ability to communicate synthetic data's role clearly and consistently. SDEG members highlighted the importance of transparency - not only as a means of encouraging internal alignment, but also as a prerequisite for external trust. Explaining why synthetic data has been used, how it was generated, and what safeguards are in place can help address misunderstandings, reduce scepticism, and support confidence in addressing regulatory and ethical standards.

5.7 This paper is not an end point in this confidence-building journey, but a foundation. Practitioners, developers, and regulators will need to continue working together- sharing best practices, testing assumptions, and developing technical approaches- to ensure synthetic data becomes a trusted component of the financial services data ecosystem.

5.8 The FCA has also championed this endeavour though its work with the Synthetic Data Expert Group (SDEG), bringing together 20 members across industry, academia, the public sector, and consumer groups. Through the Group's two publications, it has played a vital role in bridging theory and practice, offering practical insights into generation methods, evaluation metrics, and internal governance.

5.9 Ultimately, building confidence in synthetic data is not only about mitigating risk. It is about enabling purposeful, ethical, and effective innovation. With the right governance in place, synthetic data can support responsible innovation, help the FCA to become a more effective regulator, and provide opportunities for growth across the UK financial services sector.

Annex 1

Overview of considerations

The table below represents an overview of the key considerations discussed throughout the report. It outlines how practitioners can assess these considerations and suggests potential actions to manage / mitigate the challenges that emerge in relation to the use of synthetic data. This table is non-exhaustive, providing a limited overview of the nuances discussed throughout the rest of the paper.

Governance foundations

Section	Key considerations	Assess	Manage / Mitigate
Pre-project considerations Assessments to weigh a project's value against risk, by clarifying the end use case, regulatory and ethical implications, generation method, and alignment of appropriate governance processes. Decisions at this stage will shape how synthetic data is used as well as its impacts.	**Purpose, value proposition and risk assessment** Consider whether synthetic data is the optimal solution by evaluating risk vs. value.	Identifying the end use case will help practitioners to determine the expected benefits and associated risks of using synthetic data.	Practitioners may wish to establish well-defined measurable objectives to identify whether synthetic data is the best tool to deliver business goals. This may include a formal process to evaluate the trade-offs inherent to utilising synthetic data.
	Regulatory, ethical and compliance considerations Understand the key regulatory, ethical, and compliance obligations associated with using synthetic data.	Ensure careful consideration of regulatory, ethical, and compliance obligations and determine which obligations apply.	Engage with relevant compliance, legal, data and AI ethics, and risks functions in the project planning stage. Document the outcomes of cross-disciplinary reviews to establish a clear compliance baseline before the project progresses.
	Generation methodologies and downstream implications Consider the implications of different generation methodologies to determine effectiveness of synthetic data in meeting the requirements of an intended use case	Decisions on generation methodology are not purely technical and require balancing wide ranging and often competing objectives from the outset.	Effective management of the risks associated with different synthetic data methodologies requires a structured approach to evaluation based on broad technical, regulatory, and business considerations. Comparison matrices can assist in supporting practitioners to systematically compare different generation methodologies across key factors.
	Stakeholder alignment and governance readiness Consider whether your organisation has the structures to convene the multidisciplinary stakeholders required to support synthetic data projects.	Examining the adequacy of existing governance arrangements can help organisations to ensure readiness. This may involve identifying where changes are required to manage risk related to synthetic data projects	

Generating synthetic data

Section	Key considerations	Assess	Manage / Mitigate
Auditability of the generation process and outputs Auditability is the capacity to track, verify, and validate processes and decisions relating to data throughout a project lifecycle. Synthetic data poses unique challenges to auditability due to the increased number of judgment-based decisions with downstream impacts. Effective audibility requires maintaining a clear, unbroken chain of evidence through the lifecycle of synthetic data.	**Establishing ownership and use models** Consider the interplay between your organisational structure and ownership of the end-to-end components of your synthetic data project.	Assess a range of ownership models and determine which option best supports your organisational capacity and intended end use case.	Consider the use of role allocation for the generation, validation, and oversight of synthetic data and align your ownership model with existing data governance policies to prevent inconsistencies in synthetic data governance.
	Governance stages across the lifecycle Synthetic data governance and auditability can benefit from a continuous approach. Consider how to structure governance across key stages in the synthetic data lifecycle.	Assess the varying governance considerations that emerge at different stages in your synthetic data initiative.	Implementing stage gates at critical transitions in a synthetic data project can support governance and ensure that models meet predetermined thresholds.
	Using tools and metadata to support auditability Given the complexity of the synthetic data lifecycle, the use of tools and expertise can help to create transparent, traceable and well-documented audit process.	Consider where tools can help support ownership structures and governance framework to better demonstrate compliance and build confidence in the integrity of the synthetic data generation.	Practitioners can leverage tools to log changes, seek to detect issues, and support governance throughout a synthetic data project. Effectively using metadata data can also provide important context to support auditability.

Generating synthetic data

Section	Key considerations	Assess	Manage / Mitigate
Mitigating privacy risks in synthetic data Synthetic data can enable organisations to securely utilise datasets while mitigating privacy risks. While this capability is particularly valuable in financial services as a catalyst for innovation, synthetic data may not eliminate privacy risks. A deep consideration of these privacy risks can help practitioners to understand the implications and limitations of effectively utilising synthetic data.	**Understanding synthetic data privacy risks** Privacy risks can emerge across the synthetic data lifecycle. A thorough consideration of where these risks originate can help practitioners to assess and mitigate these risks.	Practitioners may benefit from assessing and classifying the privacy risks associated with a synthetic dataset based on generation methodology, intended use case, the consumption environment, and specific risks linked either to the use cases or intellectual property.	Detailed documentation and maintained records can support a continuous approach to classifying privacy risks and support effective mitigation strategies. Data Impact Protection Assessments can also provide a structured method for identifying, assessing, and mitigating data protection risks.
	Privacy testing for synthetic data Validating a synthetic dataset's privacy is complicated by the absence of a universal privacy metric. Practitioners may wish to consider a suite of techniques to evaluate privacy.	Thorough testing via a combination of privacy metrics and adversarial testing can help to provide assurances that validate a synthetic dataset's privacy.	These risks are not static so practitioners can benefit from remaining alert to these risks and revalidating their assessment of privacy metrics over time. Where feasible, organisations may also consider engaging with independent, expert reviewers or utilising other privacy enhancing technologies to enhance privacy assurances.

Generating synthetic data

Section	Key considerations	Assess	Manage / Mitigate
Managing bias when generating synthetic data Bias is a challenge in both real-world data and synthetic data generation. All data contains some degree of bias, and not all forms of bias are harmful or undesirable. Synthetic data offer an opportunity to mitigate or entrench biases depending on how it is generated, manipulated, and consumed. This introduces unique challenges as synthetic data is not limited to reflecting real-world data but often involves deliberate decisions regarding bias mitigation or amplification. In this context, the effective assessment of bias in synthetic data may require a holistic approach from initial evaluation, through generation, manipulation, and iterative testing.	**Identifying bias in source data** Where source data is used to generate a synthetic data set practitioners can interrogate the source data to understand how bias manifests.	To identify the bias in source data practitioners may wish to assess the source data's provenance, any bias in the sampling methods or propagation of bias in the collection process, and the interrelation between seemingly disparate characteristics.	Practitioners may wish to consider conducting formal bias audits on source data to support their identification and evaluation of biases. While there are a number of different approaches, SDEG members highlight the importance of developing clear and measurable criteria for evaluating biases across these audits.
	Bias manipulation during generation Once the biases in the source data are identified, understood, and documented, practitioners are better placed to make decisions on addressing these biases in their generation process. While decisions made at this stage cannot remove biases entirely, a focus on the end use case and purpose can inform decisions on how to manipulate biases in the generation process.	There are several approaches for practitioners to take when manipulating bias in the generation process, these range from: mitigating the identified bias, evaluating the assumptions in agent-based modelling, introducing controlled biases, and observing and accepting biases. It is important for practitioners to consider the trade-offs between bias manipulation, fidelity, and utility.	When manipulating biases in the generation process, practitioners may wish to consider creating clear bias objectives to avoid arbitrary changes or misalignment with the synthetic data's use case. Practitioners may also benefit from documenting the decisions taken in the generation process including the primary goals of bias mitigation, the ethical and operational impacts, and justifications for these decisions. This can help to ensure that modifications serve their intended purpose and do not inadvertently create new risks.

Generating synthetic data

Section	Key considerations	Assess	Manage / Mitigate
	Measuring and testing for bias in synthetic data Evaluating for potential biases post-generation is an important step in determining whether the synthetic dataset aligns with its intended use case and appropriate ethical standards.	While there are no universal tests for bias, practitioners may benefit from combining quantitative methods with the qualitative experiences of subject matter expertise to consider the bias present in the dataset and the implications concerning downstream synthetic data generation usage.	Practitioners may wish to leverage insights from bias evaluations to inform a more nuanced discussion on the impact of bias and considerations for future iterations.
	Iterative bias management through the generation process Testing outputs from the generation process can help practitioners to make iterative adjustments to ensure that the biases in the synthetic dataset are desired and suitable for the intended end use case.	A continuous cycle of bias evaluation, periodic re-evaluation of bias metrics, and feedback into the generation process can help to ensure that a synthetic dataset remains suitable over time, as market conditions or real-world data distributions evolve.	Cross-function reviews during technical governance forums can align bias management with broader strategic priorities. The addition of structured documentation can also support transparency and accountability in bias-related decision making.

Using synthetic data in models

Section	Key considerations	Assess	Manage / Mitigate
Evaluating synthetic data quality Post-generation, practitioners may seek to ensure that synthetic data meets quality standards suitable for downstream applications. Practitioners can leverage various methods to evaluate a synthetic dataset to understand its fitness-for-purpose, emphasising accuracy, fidelity, and utility	**High-level statistical comparison** High-level statistical comparisons can provide a first-step basis for the assessment of synthetic data quality.	Examining univariate and multivariate distributions as well as the overall correlation structure of the data can help practitioners to quickly identify inconsistencies and errors that may have occurred during the generation process.	While these methods are useful as a computationally efficient and cost-effective starting point, they cannot capture deeper multidimensional relationships. For high-dimensional datasets, these methods can become impractical and may necessitate more scalable evaluation techniques for quality assurance.
	Performance benchmarking for quality evaluation Performance benchmarking involves training and testing a model on synthetic data and comparing metrics against real data baselines.	Practitioners may wish to compare against key metrics such as accuracy, precision, or recall. If the performance metrics are comparable it can indicate that the synthetic data captures the underlying patterns required for generalisable learning.	Practitioners may wish to exercise caution when inferring synthetic data quality from performance benchmarking alone. Practitioners may wish to interrogate unexplained improvements in performance when evaluating synthetic data for issues such as data leakage of model over fitting.
	Train-Synthetic-Test-Real (TSTR) evaluation By training a model on synthetic data and comparing performance with a model trained on real data, practitioners can gather insights on the utility of synthetic data for its intended end use case.	When making these comparisons practitioners may wish to consider whether the real data selected for testing a comparable match or whether it is biased or incomplete. Incongruence between the synthetic data and real dataset may lead to misleading evaluations. Practitioners may also wish to consider whether the synthetic data misrepresents real-world feature interactions.	

Using synthetic data in models

Section	Key considerations	Assess	Manage / Mitigate
	Using performance feedback to guide iterative approaches to generation Evaluating synthetic data quality is not a static process, but a continuous and iterative task. Feedback from downstream performance can support practitioners to alter the generation process.	When drawing on performance metrics to inform the generation process, practitioners may need to balance different priorities and consider trade-offs across fidelity, utility, and task-specific performance. Practitioners may wish to consider the intent behind their approach to refining generation.	Given the trade-offs, context-specific evaluations form an important part of the decision-making process. Practitioners may also wish to implement structured feedback loops that support consistency and maintains transparent documentation on the decisions taken.
Assessing the impact of synthetic data on models The introduction of synthetic data into training processes can amplify the complexity of assessing the quality and performance of machine learning models. The impact of synthetic data on model quality depends on several factors including: the composition of the dataset, the objectives of using synthetic data, and the model's purpose.	**Dataset composition and proportions** Evaluation of the composition and proportions of synthetic data relative to real data across training, validating, and testing stages can help practitioners to understand synthetic data's impact on model performance.	When assessing the composition and proportion of synthetic dataset practitioners may wish to consider the proportion of synthetic data in training data, the mix training data by use case, and alignment across datasets for validation and testing.	
	Performance benchmarking In this context performance benchmarking can indicate whether synthetic data has improved or reduced model performance.	Benchmarking can be utilised to detect anomalies such as inflated metrics which may mask an over-reliance on artefacts introduced in the synthetic data generation.	Careful interpretation of results is required to ascertain whether performance improvements represent model gains or superficial alignment with synthetic data used in training.
	Train-Synthetic-Test-Real TSTR can provide insights into how synthetic training data affects a model's ability to generalise to real-world conditions.	If a model trained on real world data underperform when tests, it may suggest that the synthetic dataset omits critical edge cases or overrepresented simplified relationships. The primary focus is understanding whether synthetic data leads to misleading performance when deployed in live environments.	

Using synthetic data in models

Section	Key considerations	Assess	Manage / Mitigate
	Explainability and Feature Sensitivity Disparities in feature importance or sensitivity in synthetic data may signal issues with data fidelity, which could compromise model performance.	Models trained on synthetic data may rely more heavily on a narrow subset of features with a disproportionately high impact. This reliance may signal that the synthetic data fails to capture complexity and may distort model logic.	Domain expertise is important in reviewing model feature importance and sensitivity to understand whether differences are consistent with expectations for the intended use case.

Appendix 1
Group members and acknowledgements

1. The contents of this report were authored in collaboration with the SDEG members. We thank them for their contributions, insights and expertise.

2. The FCA launched an expression of interest for the Synthetic Data Expert Group in February 2023 and members were appointed in March. The SDEG is a sub-group of the IAG and operates within the IAG Terms of Reference.

Members of the Synthetic Data Expert Group

- Alexandra Ebert, MostlyAI
- Caroline Louveaux, Mastercard
- Carsten Maple, University of Warwick
- David Tracey, bigspark
- Elena Strbac, Standard Chartered
- Guilia Fanti, Carnegie Mellon
- Ismini Psychoula, Ofcom
- Janet Bastiman, Napier
- June Brawner, The Royal Society
- Lee Gregory, Barclays
- Luk Arbuckle, Privacy Analytics
- Lukasz Szpruch, Alan Turing Institute
- Marilena Karanika, Experian
- Michael Meehan, Howso
- Nick Clark, Cambridge Regulatory Innovation Hub
- Oxana Samko, HSBC
- Paul Comerford, Information Commissioner's Office (ICO)
- Robin Glover, Swift
- Tom Fiddian, Innovate UK

FCA contributions

Fatima Abukar, Lucasta Bath, Dan Gibbons, Maria Jomy, Matt Lowe, Charlie Markham, Dan Treacher, Fern Watson

Acknowledgements

With thanks to Richard Boorman in his former capacity as an FCA employee who provided feedback and input into the report.

Appendix 2

References

Bank for International Settlements (2023) Project Aurora

Jordan, J. et al (2022) Synthetic Data – What, Why, and How?

Assefa, S. et al (2020) Generating synthetic data in finance: opportunities, challenges and pitfalls

Balch, T. et al (2024) Six Levels of Privacy: A Framework for Synthetic Data

Potluru, V.K. et al (2024) Synthetic Data Applications in Finance

Houssiau, F. et al (2022) A Framework for Auditable Synthetic Data Generation

Jävergård, N. et al (2024) Tunable correlation retention: a statistical method for generating synthetic data

Jordon, J. et al (2020) Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods

Bellovin, S. M. et al (2019) Privacy and Synthetic Datasets

Blog: New data science project uses synthetic data to address the main barriers to innovation in the field of money laundering detection (2025) The Alan Turing Institute

Call for Input: Synthetic data to support financial services innovation (2022) Financial Conduct Authority

Report: Using Synthetic Data in Financial Services (2024) Financial Conduct Authority

'What is red teaming?' (2024) IBM

'Synthetic Data' (2023) Information Commissioner's Office

Auditing algorithms: the existing landscape, role of regulators and future outlook (2022) Digital Regulation Cooperation Forum

PS6/23 – Model risk management principles for banks (2023) Bank of England

Policy paper: A pro-innovation approach to AI regulation (2023) DSIT/ Office for Artificial Intelligence

Pub ref: 1-003252

FINANCIAL
CONDUCT
AUTHORITY

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