



**The
Alan Turing
Institute**

AI in Financial Services

Dr Florian Ostmann
The Alan Turing Institute

Dr Cosmina Dorobantu
The Alan Turing Institute and
the Oxford Internet Institute

This report was commissioned by the Financial Conduct Authority and considers a number of topics including AI challenges, benefits, and transparency. The report will be of use to all firms using, or contemplating, AI and machine learning and will support the FCA's future work on digital markets.

The authors are members of the Public Policy Programme at The Alan Turing Institute, which was set up in May 2018 with the aim of developing research, tools, and techniques that help governments innovate with data-intensive technologies and improve the quality of people's lives. We work alongside policy makers to explore how data science and artificial intelligence can inform public policy and improve the provision of public services. We believe that governments can reap the benefits of these technologies only if they make considerations of ethics and safety a first priority.

All research undertaken by the Turing's Public Policy Programme is supported entirely by public funds.

www.turing.ac.uk/research/research-programmes/public-policy

Cite this work as:

Ostmann, F., and Dorobantu C. (2021). AI in financial services. The Alan Turing Institute.
<https://doi.org/10.5281/zenodo.4916041>

Contents

Acknowledgments	4
Executive summary	5
1 Preface	7
2 An introduction to artificial intelligence	9
2.1 What is AI?	10
2.2 Machine learning	11
2.2.1 ML approaches	11
2.2.2 ML applications	12
2.3 Non-traditional data	14
2.4 Automation	15
2.4.1 Performing business tasks	16
2.4.2 Developing AI systems and other technological tools	16
2.5 The relationship between ML, non-traditional data, and automation	16
3 AI challenges and guiding principles	17
3.1 Background considerations	18
3.1.1 Understanding and managing data quality	19
3.1.2 Novel characteristics of models	19
3.1.3 Structural changes to technology supply chains	21
3.1.4 The scale of impacts associated with AI systems	21
3.2 Specific concerns	22
3.2.1 System performance	22
3.2.2 System compliance	23
3.2.3 Competent use and human oversight	24
3.2.4 Providing explanations	25
3.2.5 Responsiveness	26
3.2.6 Social and economic impact	27
3.3 AI ethics principles	29
3.3.1 Fairness	30
3.3.2 Sustainability	30
3.3.3 Safety	31
3.3.4 Accountability	31
3.3.5 Transparency	32
4 AI's potential benefits and harms in financial services	33
4.1 Consumer protection	34
4.1.1 Financial inclusion	34
4.1.2 Unwarranted denials of service in the context of financial crime prevention	36
4.1.3 Unlawful discrimination and unfair differential treatment	37
4.1.4 Mismatches between products and customer needs	38
4.1.5 Performance of investments	38
4.1.6 Consumer empowerment	39
4.2 Financial crime	41
4.3 Competition	41
4.4 Stability of firms and markets	43
4.5 Cybersecurity	43
5 AI transparency	45
5.1 Defining transparency	46
5.1.1 Relevant information (the 'what')	46
5.1.2 Relevant stakeholders (the 'who')	47
5.1.3 Reasons for accessing information (the 'why')	48
5.2 System transparency	48
5.2.1 Relevant information (the 'what')	49
5.2.2 Purpose (the 'why')	51
5.2.3 Methodology (the 'how')	52
5.3 Process transparency	55
5.3.1 Relevant information (the 'what')	56
5.3.2 Purpose (the 'why')	59
5.3.3 Methodology (the 'how')	61
5.4 Trade-offs	62
6 Conclusion	64
Appendix: ML approaches	67
A.1 Supervised learning	68
A.2 Unsupervised learning	69
A.3 Reinforcement learning	69
Glossary	70
Bibliography	74

Acknowledgments

The authors are grateful for the invaluable input and support received throughout this project from numerous FCA colleagues, including Alison Russell, Ravi Bhalla, Nick Cook, Leo Gosland, and Henrike Mueller. We also wish to express our gratitude for insightful comments and suggestions from Helen Margetts, David Leslie, Mhairi Aitken, Christopher Burr, as well as for the background research work of Michel Haddad and Alex Harris. Last, but certainly not least, we would like to thank the industry representatives who attended our six workshops and provided crucial feedback.

Executive Summary

Artificial intelligence (AI) plays a central role in current processes of technological change in financial services. Its prominent place on innovation agendas speaks to the significant benefits that AI technologies can enable for firms, consumers, and markets. At the same time, AI systems have the potential to cause significant harm. In light of this fact, recent years have seen a growing recognition of the importance of AI adoption being guided by considerations of responsible innovation.

The aim of this report is to inform and advance the debate about responsible AI in the context of financial services. It provides an introduction to relevant technological concepts, discusses general challenges and guiding principles for the adoption of AI, maps out potential benefits and harms associated with the use of AI in financial services, and examines the fundamental role of AI transparency in pursuing responsible innovation.

Introduction to AI

The field of AI has a decades-long history and substantial links to statistical methods with long-standing applications in financial services. The adoption of AI in financial services is underpinned by three distinct elements of innovation: machine learning (ML), non-traditional data, and automation. AI systems can combine all three elements or a subset of them. When considering a particular AI use, it is useful to distinguish between these three elements of innovation and examine their respective role. Doing so is crucial for an adequate understanding of AI-related risk, as each element can give rise to distinct challenges.

General challenges and guiding principles for the responsible adoption of AI

ML, non-traditional data, and automation give rise to various challenges for responsible innovation. These challenges provide the foundation for understanding the causes of AI-related risks. They are often related to the following four background considerations:

- The performance of AI systems crucially depends on the quality of the data used, but data quality issues can be difficult to identify and address.
- Models developed with ML can have model characteristics that set them apart from more conventional models, including opaqueness, non-intuitiveness, and adaptivity.
- The adoption of AI can be accompanied by significant changes in the structure of technology supply chains, including increases in supply chain complexity and the reliance on third-party providers.
- The use of AI can be accompanied by an increased scale of impacts when compared to conventional ways of performing business tasks.

Against the background of these considerations, AI can give rise to specific concerns. These include concerns about (i) AI systems' performance, (ii) legal and regulatory compliance, (iii) competent use and adequate human oversight, (iv) firms' ability to explain decisions made with AI systems to the individuals affected by them, (v) firms' ability to be responsive to customer requests for information, assistance, or rectification, and (vi) social and economic impacts.

In light of these concerns, recent years have seen a rapidly growing literature on AI ethics principles to guide the responsible adoption of AI. The principle of transparency, in particular, plays a fundamental role. It acts as an enabler for other principles and is a logical first step for considering responsible AI innovation.

Executive Summary

Potential AI-related benefits and harms in financial services

The use of AI in financial services can have concrete impacts on consumers and markets that may be relevant from a regulatory and ethical perspective. Areas of impact include consumer protection, financial crime, competition, the stability of firms and markets, and cybersecurity. In each area, the use of AI can lead to benefits as well as harms.

AI transparency and its importance for responsible innovation

The general challenges that AI poses for responsible innovation, combined with the concrete harms that its use in financial services can cause, make it necessary to ensure and to demonstrate that AI systems are trustworthy and used responsibly. AI transparency – the availability of information about AI systems to relevant stakeholders – is crucial in relation to both of these needs.

Information about AI systems can take different forms and serve different purposes. A holistic approach to AI transparency involves giving due consideration to different types of information, different types of stakeholders, and different reasons for stakeholders' interest in information.

Relevant transparency needs include access to information about an AI system's logic (system transparency) and information about the processes surrounding the system's design, development, and deployment (process transparency). For both categories, stakeholders that need access to information can include occupants of different roles within the firm using the AI system (internal transparency) as well as external stakeholders such as customers and regulators (external transparency).

For system and process transparency alike, there are important questions about how information can be obtained, managed, and communicated in ways that are intelligible and meaningful to different types of stakeholders. Both types of transparency – in their internal as well as their external form – can be equally relevant when it comes to ensuring and demonstrating that applicable concerns are addressed effectively.

In covering these topics, the report provides a comprehensive conceptual framework for examining AI's ethical implications and defining expectations about AI transparency in the financial services sector. By doing so, it hopes to advance the debate on responsible AI innovation in this crucial domain.

1 Preface

Financial services are undergoing a period of significant technological change. Advancements in artificial intelligence, happening within a wider context of digital transformation, contribute to this change.

1 Preface

The growing popularity and adoption of AI technologies speak to the advantages that these technologies bring. AI occupies a prominent place on firms' innovation agendas¹ because of its potential to enable significant benefits for firms, customers, and markets. At the same time, vast research efforts and an expanding literature on AI ethics and governance speak to the challenges that these technologies pose. Academics and policymakers alike are devoting time and resources to the development of governance principles², organisational processes³, and technical tools for safe and ethical AI because of AI's potential to cause significant harm.

This report does not take a stance in favour or against the adoption of AI technologies. Instead, it seeks to inform a wide audience: from firms considering AI systems to members of the public interested in these technologies. The report provides a foundational understanding of what AI is, how it can lead to both benefits and harms, and why transparency is crucial in addressing the challenges it poses. It focuses on the financial services sector⁴ and the importance of trustworthy and responsible innovation.

The report starts by providing a general introduction to AI in financial services. Chapter 2 touches on AI's historical and methodological connections. It defines relevant terms and introduces three elements of AI innovation that underpin current advancements: machine learning (ML), non-traditional data, and automation. Chapter 3 discusses the challenges that can arise in the context of AI and introduces principles that serve to guide the pursuit of trustworthy and responsible innovation. In doing so, it not only covers the concerns raised by AI technologies but also acknowledges the important progress that has been made in steering the responsible design, development, and deployment of AI systems. Chapter 4 turns its attention to the financial services. It discusses potential benefits and harms associated with the use of AI in financial services from a perspective of outcomes for consumers and markets.

While Chapters 2, 3, and 4 provide a broad introduction to AI in financial services, Chapter 5 takes a deep dive into AI transparency. The successful and beneficial adoption of AI in financial services depends on the ability of firms to ensure as well as to demonstrate that AI systems are trustworthy and used responsibly. Transparency plays a crucial role in achieving these objectives. In recognition of this role, the chapter takes an in-depth look at AI transparency. It moves beyond discussions about transparency as a general principle and discusses the forms that it can take, the purposes it can serve, and relevant practical considerations.

The report's conclusion draws attention to the important work that remains to be done to ensure that firms and regulators find the fine balance between enabling AI innovation and mitigating the risks that it poses. AI holds unprecedented potential to transform financial services. It is our collective responsibility to ensure that this transformation occurs in a responsible and socially beneficial way.

1 See FCA and Bank of England 2019; Buchanan 2019; Cambridge Centre for Alternative Finance and World Economic Forum 2020; World Economic Forum 2018.

2 For example, see High-Level Expert Group on Artificial Intelligence 2019; OECD 2019; IEEE 2020.

3 For example, see ICO and The Alan Turing Institute 2020; Leslie 2019; Raji et al. 2020; Ashmore, Calinescu, and Paterson 2019; Brundage et al. 2020.

4 Other noteworthy reports related to AI in financial services include European Banking Authority 2020; Association for Financial Markets in Europe (AFME) 2018; De Nederlandsche Bank 2019; Autorité de Contrôle Prudentiel et de Résolution 2020; Hong Kong Monetary Authority 2019a; Hong Kong Monetary Authority 2019b; Monetary Authority of Singapore 2019.

2 **An introduction to artificial intelligence**

The last few years have seen an explosion of interest in AI. Yet, use of the term AI is often vague and the technological changes that underpin the adoption of AI are not always made clear. In addition, the recent surge in attention can obscure the history of AI as a research field and the fact that technological advancements often build on long-standing methods and take incremental forms.

This chapter provides a short introduction to AI. It defines relevant terms and considers three elements of technological change that drive AI innovation in financial services: machine learning, non-traditional data, and automation.

2 An introduction to artificial intelligence

2.1 What is AI?

AI is the 'science of making computers do things that require intelligence when done by humans.'⁵ While the last two decades have seen a series of research advances that have triggered an unprecedented expansion of interest in the topic, the history of AI as a field of systematic research extends back to the 1950s when the term 'artificial intelligence' was originally coined.

Within the field of AI, several high-level distinctions can be drawn. Here, we introduce two of them, namely the distinction between symbolic and statistical AI, and the distinction between general and narrow AI.

Symbolic AI

Symbolic AI relies on translating human knowledge and logical statements into explicitly programmed rules. For example, 'if a financial transaction is above £10,000, then flag it for human review.' Many of the early chess programmes fall within the category of symbolic AI. They are programmed in a top-down manner, encoding human knowledge about chess in software rules, and they identify playing strategies by searching through possible moves. Symbolic AI was the predominant approach in AI research between the 1950s and 1980s.

Statistical AI

Statistical AI, in contrast, refers to the development of bottom-up, data-driven systems. The capabilities of such systems are not the result of the rule-based application of encoded human knowledge but instead arise from the analysis of data. AlphaZero, a computer programme which can play highly complex games, is an example. Rather than relying on hard-coded rules, AlphaZero learns how to master games from the data it generates by playing itself.

Most of the recent progress in AI has been due to advancements in the field of statistical AI. These advancements have been enabled by rapid expansions in available computing power, improvements in algorithmic techniques, significant increases in available data, and growing investments in AI development.⁶

General AI

In addition to the contrast between symbolic and statistical AI, there is an important distinction between narrow AI and general AI.⁷ General AI, an ambition rather than a reality, refers to systems that have universal abilities on par with those of the human mind. These abilities include the versatility to learn and perform any intellectual task that humans are capable of. Although significant research efforts are being dedicated to the development of general AI, there is limited prospect of it being achieved in the foreseeable future. Experts disagree on whether the aim will ever be realised.

Narrow AI

The AI systems that we see in business use today take the form of narrow AI. The abilities of these systems are limited to the relatively narrow pre-defined tasks for which they were developed. They are far from replicating the generalised intelligence shown by humans. Having said that, the transformative potential of narrow AI should not be underestimated. For certain specific tasks, the performance of narrow AI systems may well exceed that of humans. Recent successes in image recognition or playing complex games such as chess show this. This report is limited to developments in narrow AI. Unless otherwise stated, this is what we mean when we use the term 'AI'.

⁵ Marvin Minsky, quoted in Leslie 2019.

⁶ Brundage et al. 2018.

⁷ Also referred to as Artificial Narrow Intelligence and Artificial General Intelligence.

2 An introduction to artificial intelligence

Distinctions like symbolic and statistical AI or general and narrow AI are helpful for a high-level understanding. What does this mean for the current debate, in particular its application in financial services? What are the main technological changes that we see as AI-enabled tools and systems become more widespread? Three elements of AI innovation, in particular, are worth emphasising: **ML**, **non-traditional data**, and **automation**. All three elements are relevant to financial services, and we discuss each of them in more detail below.

2.2 Machine learning

ML refers to the development of AI systems that are able to perform tasks as a result of a ‘learning’ process that relies on data. This contrasts with approaches and methods in the field of symbolic AI and traditional software development, which rely on embedding explicit rules and logical statements into code. ML is at the core of recent advances in the field of statistical AI and the methodology behind technological feats such as computer programmes outperforming humans in tasks ranging from medical diagnosis to complex games. The recent surge of interest in AI is, in large part, due to achievements made possible by ML.

As the term statistical AI suggests, ML draws on concepts from statistics and probability theory. Many forms of ML go beyond traditional statistical methods, which is why we often think of ML as an exciting new field. Yet, despite the hype surrounding this technological development, the boundary between ML and statistics is blurry. There are contexts in which ML is best thought of as being on a continuum with traditional statistical methods rather than representing a clearly defined separate field. And regardless of definitional boundaries, ML is often used to perform the same analytical tasks that traditional statistical methods have been used for in the past.

2.2.1 ML approaches

ML is a very active research field, encompassing a broad and evolving range of methods. At a high level, three primary approaches can be distinguished: **supervised learning**, **unsupervised learning**, and **reinforcement learning**.⁸ (the Appendix describes them for the interested reader.) A vast array of individual ML methods, such as linear regression, decision trees, support vector machines, artificial neural nets, and ensemble methods, sits under the umbrella of these three approaches. Two general points on methodological differences are worth noting.⁹

First, ML methods vary significantly in their complexity. Discussions of ML often focus on methods with a high degree of complexity. For example, neural networks, a family of methods that search for patterns and relationships in datasets using network structures similar to those we see in biological brains, receive significant attention. However, ML also comprises less complex methods such as ordinary least squares regression and logistic regression. These simpler methods have a long history of use in the fields of statistics and econometrics and predate the emergence of the field of ML in its current form. We will return to the issue of complexity and its practical implications in later chapters. For now, it should be noted that ML, as a field, includes certain highly complex methods but is not limited to them.

Second, ML methods can be used to develop **static** or **dynamic systems**. In the case of static systems, ML is used to develop models that, once deployed, do not evolve further unless they are deliberately replaced with a new model. In dynamic systems, in contrast, models, once deployed, continue to adapt based on new data that becomes available during operation.

⁸ Mixed approaches, such as semi-supervised learning, also exist.

⁹ For an overview of the most prominent supervised learning methods, see Annexe 2 in ICO and The Alan Turing Institute 2020.

2 An introduction to artificial intelligence

Such dynamic (or incremental) learning can have significant benefits in situations where the data available during development is limited or where models capture phenomena that have rapidly changing characteristics. An example of the latter is where a continuous stream of data is generated as a result of market interactions, much like we see in stock trading or foreign exchange trading. Discussions of ML often focus on dynamic systems. Their potential benefits, combined with the challenges to ensure dynamic systems' safety and trustworthiness, tend to capture more attention than static systems. Static systems, however, are much more widely used.

Figure 1 summarises our discussion about ML approaches. There are three primary ones: supervised, unsupervised, and reinforcement learning. Numerous ML methods fall underneath these three umbrella terms. These methods vary in complexity and can be used in static or dynamic ways.

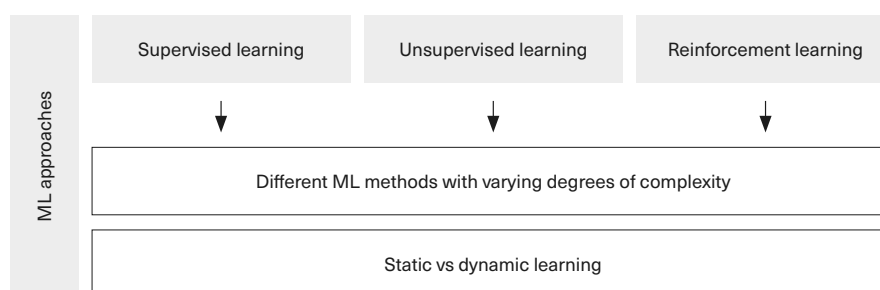


Figure 1: ML approaches

2.2.2 ML applications

We can distinguish two broad types of ML applications: **traditional data analysis and modelling tasks** and the **processing of unstructured data**.

2.2.2.1 Traditional data analysis and modelling tasks

Traditional data analysis and modelling tasks are tasks that involve **structured data**, such as stock price data, financial transaction data, or other types of data that are stored in machine-readable formats. In financial services, there are many operational areas and decision-making processes where data analysis and modelling have a long history. Examples include assessing credit eligibility, predicting insurance costs, forecasting stock prices, or modelling prudential risk. In other cases, business decisions have started to rely on data analysis and modelling more recently, as data-driven solutions become easier or cheaper to pursue. Examples include the introduction of ML-based approaches to the detection of suspicious financial transactions.

Specific types of traditional data analysis and modelling tasks include:

- **Prediction¹⁰ and forecasting tasks** focused on estimating the future value of a variable of interest. Examples include predicting loan defaults, insurance costs, purchasing decisions, or stock values and portfolio returns.

¹⁰ It is worth noting that the technical ML literature often uses the term 'prediction' in a different sense according to which any ML problem that involves a target variable is thought of as a prediction problem. Here, we use the term more narrowly, in the common language sense of statements about future events or states of the world.

2 An introduction to artificial intelligence

- **Optimisation tasks** focused on estimating the optimal value of a given variable or range of variables of interest. Examples include identifying optimal pricing or prudential risk management strategies.
- **Detection tasks** focused on identifying the occurrence of phenomena of interest, often through the detection of outliers or anomalies in data. Examples include detecting cybersecurity threats or identifying different forms of fraud, market abuse, or suspicious activities in the context of anti-money laundering.

The performance of traditional data analysis and modelling tasks does not necessarily require ML. This is reflected in the longstanding use of statistical and actuarial methods to perform many of these tasks. But ML can make a difference. First, ML can make performing analytical tasks easier, faster, or cheaper. Second, it can lead to improved results in performing a given task. For example, by enabling increases in the number of variables being processed or in the complexity of relationships between variables, ML can result in more accurate models. Finally, there are certain analytical tasks that can only be performed thanks to ML. For example, where meaningful results depend on quantities of data that cannot be processed using traditional methods, ML can make a difference for the feasibility of adopting a data-driven approach.

2.2.2.2 Processing of unstructured data

Many types of data, such as human language (commonly referred to as 'natural language') and visual data, are **unstructured data**. The processing of natural language and visual data by machines is the focus of two long-standing sub-fields of AI research: **natural language processing (NLP)** and **computer vision**. Unstructured data has historically been outside the reach of computational processing capabilities. The last decades, however, have seen rapid progress in both fields, largely due to ML. In contrast to traditional data analysis and modelling tasks, many of which can in principle be performed without ML (albeit less efficiently or less effectively), current capabilities to process unstructured data fundamentally depend on ML.

Natural language processing

NLP capabilities fall into three main categories:

- **Speech recognition** refers to the processing of spoken language. This includes tasks such as transforming speech into text or recognising sentiments and emotions from the sound of a speaker's voice (voice sentiment analysis). Examples of the use of speech recognition capabilities in financial services include identifying customers based on their voice or triaging customer service calls based on voice commands.
- **Natural language understanding** involves recognising meaning in human language, with applications such as content classification, text mining, or machine translation. Examples of use include processing human responses in customer service chatbot conversations or analysing the content of corporate annual reports for investment intelligence purposes.
- **Natural language generation** is about producing written or spoken human language. It comprises tasks such as document drafting, summarising text, and generating dialogue responses. Examples include generating draft contracts or customer service chatbot responses.

2 An introduction to artificial intelligence

Computer vision

Computer vision capabilities typically take the form of recognition tasks. These include detecting physical objects and their condition, facial recognition, and optical character recognition (ie the recognition of characters from images of handwritten or printed text). Examples of the use of computer vision in financial services include vehicle damage analysis based on photographs provided by motor insurance clients, facial recognition to confirm the identity of mobile banking customers, or the use of optical character recognition to render the content of insurance claim forms or other documents machine-readable.

Figure 2 summarises our discussion in this section. ML approaches can be used for a range of different applications, including traditional data analysis and modelling tasks and the processing of unstructured data.

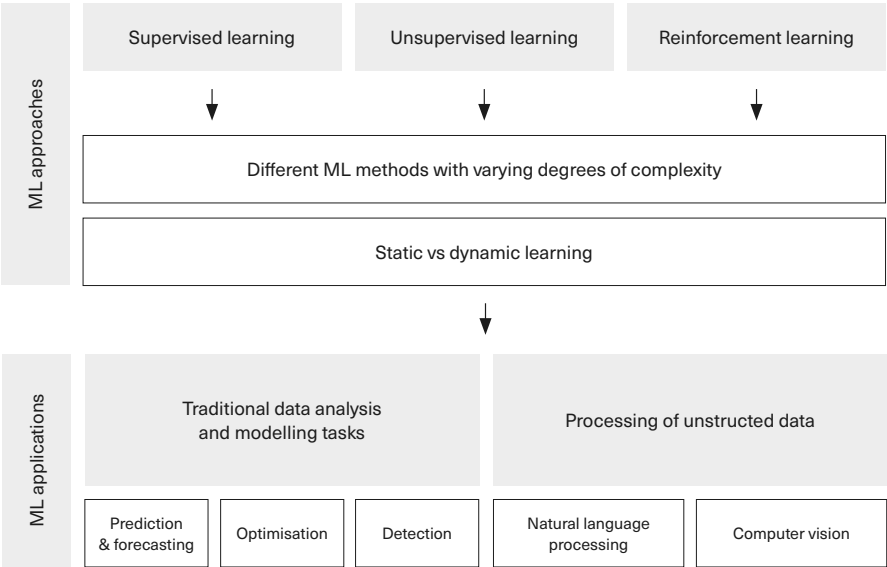


Figure 2: ML approaches and ML applications

2.3 Non-traditional data

Recent years have seen an expansion in the types of data that can inform business tasks in financial services. Firms are increasingly exploring the use of non-traditional data (ie types of data that have not been used historically to perform the task in question). Recent uses of non-traditional data include structured as well as unstructured data. In some instances, the data in question did not exist or was not previously accessible. In other cases, it was available but went unused due to a lack of technical capabilities.

Prominent examples involving non-traditional structured data include relying on financial transaction data, from the bank accounts of consumers looking for a loan, for credit risk profiling or using telematics data, from monitoring sensors installed on cars, for insurance risk profiling. Examples involving non-traditional data that is unstructured include using news media content, recordings of company earning calls, or satellite imagery¹¹ to predict stock or commodity prices or using images of damaged vehicles to assess the value of insurance claims.

11 For instance, satellite images can be used to estimate soil moisture in order to predict crop yields and prices.

2 An introduction to artificial intelligence

More generally, non-traditional data can serve two main purposes. It offers:

- **Alternative sources for established types of information:** Non-traditional data can provide alternative ways for businesses to access the information they need to perform a task. Credit lenders' use of information about loan applicants' income in credit eligibility decisions serves as an example. Traditionally, loan issuers get income information through the documents that loan applicants provide. Nowadays, they can also get income information by accessing applicants' financial transaction data. The new source of information (financial transaction data) could be used alongside the traditional source (documents provided by the applicants) to verify information. Or it could be used to replace the original source of information (eg reducing the reliance on information manually provided by loan applicants).
- **Sources of new types of information:** Non-traditional data can provide information that businesses have not been able to gather or use in the past. When used responsibly, the new types of information that non-traditional data opens up may enable significant improvements in the performance of prediction, optimisation or detection tasks. For example, using telematics data in the context of motor insurance or spending pattern data in the context of consumer lending may enable more precise forms of risk profiling compared to approaches that rely on traditional types of information.

The growing significance of non-traditional data is intertwined with the broader rise of digitisation in firm's interactions with customers as well as their internal operations. This includes trends such as the increased use of digital forms of communication and firms' reliance on end-user computing solutions.

2.4 Automation

Automation reduces or removes the role of humans in performing tasks and processes. Automation and ML are often discussed in connection with each other. This is helpful in raising public awareness and motivating research into the benefits and risks posed by the marriage of these two elements. But automation and ML do not necessarily go hand in hand. Automation can exist in contexts that do not involve ML. This is illustrated by long-standing approaches in the areas of workflow automation or robotic process automation. In turn, ML can exist in contexts that do not involve automation. Many ML models, for example, are designed to simply generate predictions without these predictions resulting in any automatic decisions or action.

It is also important to note that automation is not a binary issue of full human involvement versus a complete absence of human involvement. Instead, different forms of automation fall on a continuum of varying degrees of human involvement. This continuum includes arrangements in which the need for human input is reduced, but where humans retain certain forms of control. For example, there are **human-in-the-loop** arrangements, in which humans confirm the execution of actions or decisions, or **human-on-the-loop** arrangements, in which humans play a supervisory role and can override the execution of actions or decisions.

Businesses in financial services use automation in various ways. This includes the reliance on automation to perform business tasks and the use of automation to develop AI systems and other technological tools. We turn to these two uses below.

2 An introduction to artificial intelligence

2.4.1 Performing business tasks

When it comes to performing business tasks, three uses of automation are worth highlighting:

- **Automated decision-making** reduces or removes human involvement in decision-making processes. This includes decisions purely based on explicitly programmed rules (as in the case of conventional approaches to detect suspicious financial activity) as well as decisions based on the output of models designed to perform prediction, forecasting, optimisation, or detection tasks (eg credit eligibility, price optimisation, or stock trading decisions).
- **Automated information management** facilitates information and data management tasks. Examples include automating data procurement processes (eg different forms of web scraping), data updating, data transfer processes, or the use of auto-completion features in relation to customer application forms or other business documents.
- **Automated information verification** can compare information from different sources to identify inaccuracies or inconsistencies between them. This can be useful, for example, in assessing the validity of customer-provided information during product application processes or in detecting fraudulent insurance claims.

2.4.2 Developing AI systems and other technological tools

Automation can also be used to develop and maintain technological and analytical tools. In the context of tools that rely on ML, the term **automated machine learning** (AutoML) is commonly used to refer to the automation of processes related to developing or maintaining ML models. This type of automation can pose risks, but when used responsibly, it can make ML capabilities more widely accessible and cheaper to develop.

2.5 The relationship between ML, non-traditional data, and automation

For analytical clarity, we presented the three elements of AI innovation separately. However, AI systems can combine all three elements – ML, non-traditional data, and automation. Consider the following hypothetical examples:

- A stock trading system may make automated trading decisions based on a ML model that relies on free-text data from news reports as one of its inputs.
- A loan application processing system may rely on a credit scoring model developed using ML methods and non-traditional data from applicants' financial transaction history to make automated decisions about loan eligibility.

Each of the three elements of AI innovation can be the source of distinct benefits and concerns. Moreover, just as we have examples of important innovations that involve all three elements, there are also examples of innovations that only involve one or two of these elements. To reflect this, the discussion in the rest of this report will – where appropriate – highlight and distinguish between the roles of ML, non-traditional data, and automation.

3 AI challenges and guiding principles

The use of ML, non-traditional data, and automation holds significant promise, but can also contribute to challenges in managing the trustworthiness and responsible use of AI systems. These challenges have spurred a rapidly growing literature on trustworthy and responsible AI.

This chapter discusses the challenges that AI poses for responsible innovation, explains how they arise, and introduces AI ethics principles that serve to guide the adoption of AI.

3 AI challenges and guiding principles

In the first part of this chapter, we look at four background considerations:

- understanding and managing data quality
- novel characteristics of models
- structural changes to technology supply chains
- the scale of impacts associated with AI systems

These four considerations are at the root of many of the concerns we encounter in AI innovation and provide a foundation for understanding them.

The second part of the chapter provides an overview of specific concerns that can arise from these background considerations in six areas:

- system performance
- system compliance
- competent use and human oversight
- providing explanations
- responsiveness
- social and economic impact

The concluding section considers the relationship between these six areas of concern and the most prominent principles that have emerged in the literature on AI ethics and governance.

Before we delve deeper into the mentioned background considerations and areas of concern, we note that they are not necessarily unique to contexts that use ML, non-traditional data, or automation. Many of them apply to other contexts but can become more salient or more difficult to manage in the context of AI innovation. Furthermore, the background considerations and areas of concern discussed below are not specific to the financial services sector. Indeed, they are applicable to the use of AI technologies in any sector.

3.1 Background considerations

With these caveats, we are ready to move into a detailed discussion of the four background considerations, summarised in Figure 3.

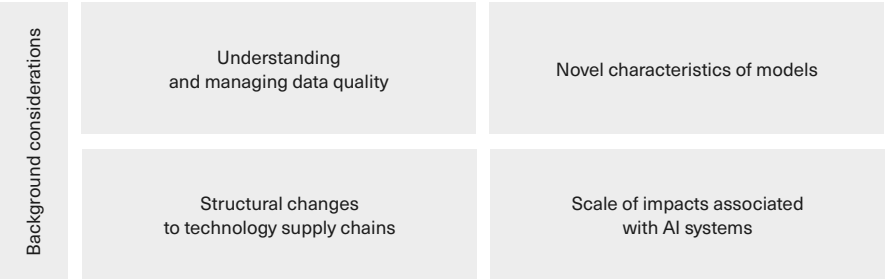


Figure 3: Background considerations

3 AI challenges and guiding principles

3.1.1 Understanding and managing data quality

Data quality is crucial to the performance of all data-reliant systems. This includes the quality of both the data that systems draw on during their operation and the data used in developing them. There are various conceptual frameworks for distinguishing between different aspects of data quality. Without being exhaustive, five aspects of data quality are worth highlighting here:

- **Accuracy:** Do the values recorded in a dataset correspond to the true values of the variables in question?
- **Recency:** Is the data up to date?
- **Conceptual validity:** Do the variables in the dataset measure what they are assumed to measure?
- **Completeness:** Does the dataset contain values for all required variables and observations?
- **Representativeness:** Is the composition of the dataset adequate to serve as a representation of the real world for the intended purpose?

The pursuit of AI is characterised by an increase not only in the types and amounts of data available but also in ways of obtaining and using data. This makes data quality issues and data quality management challenges especially prominent. Reasons why data quality limitations may be more difficult to identify and address include the following:

- The use of non-traditional data may be accompanied by a lack of awareness of and experience with addressing the quality limitations for such data. This could be due to inherent data characteristics or due to obstacles resulting from data procurement arrangements. For example, when using scraped social media data, it is difficult to assess representativeness.
- Increases in the size and number of datasets used and the reliance on automatic processes for handling increased quantities of data can make data quality issues more difficult to identify than in systems that rely on less data and a higher level of human involvement.
- New ways of drawing on dynamic, constantly updating types of data in AI systems that are deployed in real-time can limit the ability to detect and mitigate data quality issues before they affect performance.
- Using existing datasets for new purposes, for example as training data for supervised machine learning, can mean that some data quality aspects become crucial (eg data representativeness).

3.1.2 Novel characteristics of models

Models developed using ML methods and non-traditional data can have certain characteristics that set them apart from more conventional models. Three characteristics are particularly noteworthy: opacity, non-intuitiveness, and adaptivity.

3.1.2.1 Opacity

ML can be accompanied by increases in dimensionality (the number of input variables that a model relies on) and in the complexity of relationships between variables.¹² Two of the very features that make ML attractive – the ability to accommodate more data and more complex relationships in the data – can create challenges in understanding how a model's output relates to its input variables. These difficulties in discerning 'how the model works' are commonly referred to as 'black box' or model opacity problems.

¹² Examples for increases in the complexity of variable relationships include non-linear, non-monotonic, and discontinuous relationships. See Selbst and Barocas 2018.

3 AI challenges and guiding principles

Model opacity can occur in two forms:¹³

- **Opacity due to inscrutability:** Some models are so complex that determining relationships between model inputs and outputs based on a formal representation of the model is difficult to achieve as a matter of principle. Such models are opaque to anyone, including experts with high levels of specialised technical knowledge.
- **Opacity due to non-expertise:** Models that are scrutable, ie intelligible in principle, can exhibit forms of complexity that mean that understanding them requires a certain level of technical expertise. Such models can appear opaque to anyone not equipped with the required level of expertise.¹⁴

3.1.2.2 Non-intuitiveness

Models can draw on statistical relationships that are non-intuitive. Non-intuitiveness is not a new problem. It can occur in any modelling context, but the use of ML and non-traditional data can be more likely to identify non-intuitive relationships compared to conventional modelling approaches.

In the case of ML, models that allow for more complex relationships between variables can lead to higher degrees of non-intuitiveness compared to models that rely on the same input variables but are simpler. For example, a simple model to predict the likelihood of loan default may assume, in line with common intuition, that increases in an applicant's income can only ever reduce the applicant's likelihood of default. A more complex modelling approach, in contrast, could uncover that, under some specific conditions, increases in income are associated with a higher risk of default.

Using non-traditional data can be a separate source of non-intuitiveness. Traditional models for predicting loan default or insurance costs, for example, tend to rely on variables whose association with the outcome of interest is intuitive and easy to grasp. Non-traditional input variables such as information about shopping habits or social media behaviour, in contrast, could give rise to statistically significant relationships that are difficult to discern intuitively.¹⁵

3.1.2.3 Adaptivity

As noted in Chapter 2, ML methods can be used to develop dynamic models whose structure or parameters adapt during deployment in response to new data. When compared to static models, which remain unchanged unless they are deliberately updated, the use of dynamic models can give rise to specific challenges when it comes to ensuring system safety and trustworthiness.

In particular, monitoring and managing the performance of dynamic models can be a more demanding and difficult task compared to static models, as adaptive changes can entail unexpected and unintentionally occurring performance deteriorations. In addition, dynamic models can be uniquely vulnerable to certain forms of deliberate manipulation aimed at undermining their performance. An example for this is the risk of data poisoning attacks, whereby dynamic models are exposed to data that is intended to 'retrain' them with the aim of reducing their effectiveness.

¹³ Some authors discuss 'secrecy' as a third relevant form of opacity in the context of AI systems. We discuss secrecy below as a separate challenge. Here, we focus on opacity as a property that arises from characteristics of models rather than intentional decisions not to disclose relevant information about them. See Burrell 2016; Selbst and Barocas 2018.

¹⁴ Since the level of expertise required can differ between model types, opacity due to non-expertise is a relative concept: in some cases, a basic level of statistics training may be sufficient to avoid this form of opacity, in other cases understanding the model may require advanced forms of ML expertise.

¹⁵ Consider the example of a consumer switching from an upmarket to a discount supermarket chain. Does this switch represent a signal for responsible management of financial resources or a signal for expected financial difficulties? Intuitively, a predictive effect in either direction may seem conceivable.

3 AI challenges and guiding principles

3.1.3 Structural changes to technology supply chains

We have examined two background considerations that give rise to some of the concerns that we encounter in AI innovation: understanding and managing data quality and novel characteristics of models. We now step away from the technologies themselves and turn to a third background consideration, related to the structure of technology supply chains.

The adoption of AI systems can be accompanied by changes to the structure of technology supply chains. These changes include not only increases in the complexity of supply chains, but also increases in outsourcing and the reliance on third parties.

Three specific aspects are worth distinguishing:

- As AI systems increase in their technological sophistication and the types of data they draw on, the number of actors involved in the design and development of systems – be it directly or indirectly (eg through the provision of data) – grows. This growing number of actors can include employees within the firm that is using the AI system as well as third-party providers.
- The development of AI systems may rely on off-the-shelf tools, pre-existing models or software that were not developed specifically for the purpose at hand. Such tools may have an in-house origin or be obtained from external sources (on a commercial or open-source basis).
- Third-party providers can play an increasingly prominent role in the context of AI systems. Firms can use third-party providers to source off-the-shelf tools, procure data, or outsource to them the development and even the operation of bespoke AI systems.

These structural changes to technology supply chains can create challenges for responsible AI adoption. At a general level, increases in the number of actors involved in the design, development, and deployment of AI systems can make it more difficult to achieve effective governance arrangements and adequate flows of information. This difficulty applies to employees within the same organisation, but is particularly salient when it comes to supply chains and business relationships that cross organisational boundaries. The reliance on third-party providers can, for example, limit the commissioning company's access to information about data quality and provenance, the specification of relevant tools or software, or providers' quality assurance and risk management practices. These challenges are even greater when contractual arrangements render certain forms of information legally inaccessible (eg the code of a given model being treated as a trade secret).

3.1.4 The scale of impacts associated with AI systems

The final background consideration that we want to highlight is the scale of impacts associated with AI systems. Turning to AI for a given business task can affect the scale of potential positive or negative impacts associated with the performance of that task. Three factors, in particular, are worth mentioning:

- The introduction of AI systems to support the execution of business tasks that would otherwise be performed by humans increases the scale of potential impacts. As humans, we cannot perform the same volume of tasks that an AI system can conduct. The large volume of work that an AI system can complete means that the good or bad outcomes it produces vastly outnumber the good or bad outcomes generated by the work of any human being.

3 AI challenges and guiding principles

- The reduction of the role of human involvement through advances in automation means that individual failures in the performance of human supervision and intervention can have graver consequences. As the number of possible points of corrective intervention decreases, a human's failure to intervene when needed can lead to cascading impacts.
- Where the same tools and systems are used across different firms – for example as a result of reliance on the same third-party vendors – the potential positive or negative impacts associated with individual tools and systems are amplified.

3.2 Specific concerns

The four background considerations just described give rise to a range of possible specific concerns related to AI systems. The most prominent of these concerns can be divided into six areas. The following sub-sections will look at each of these areas in turn. Figure 4 provides a summary of the six areas of concern and the previously discussed background considerations, illustrating the relationship between them.

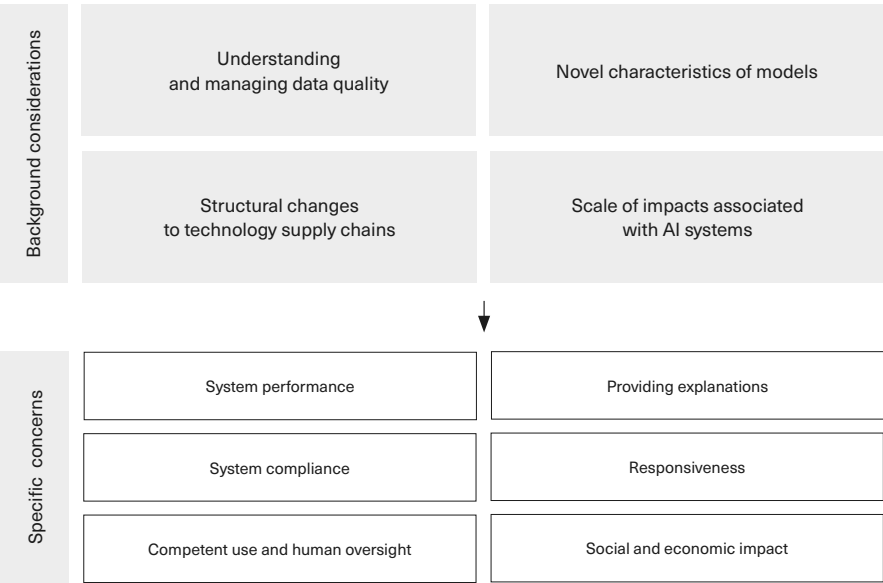


Figure 4: Background considerations and specific concerns

3.2.1 System performance

Performance is central to the trustworthiness of any AI system. This includes the system's performance at the time of its development as well as any changes in performance that may occur post-deployment. An AI system's development phase typically involves measuring different aspects of the system's effectiveness, such as accuracy and error rates. In addition, responsible AI system development practices include assessments of the system's reliability and robustness, ie the expected stability of the system's performance during deployment under normal and exceptional operating conditions. Once an AI system is in operation, managing performance involves adequate arrangements of performance monitoring and, where needed, system adjustments.

16 For a more detailed discussion of relevant issues, see, for example Leslie 2019.

3 AI challenges and guiding principles

The reliance on ML, non-traditional data, and automation can make it more difficult to obtain an adequate understanding of an AI system's performance and performance weaknesses. The four background considerations introduced in the previous section illuminate different grounds for concern about the performance of AI systems. In particular:¹⁶

- The quality of the data used during an AI system's development or operation is a crucial determinant of performance. As a result, difficulties in understanding and managing data quality can create problems for the assessment and management of system performance.
- Novel characteristics of models, such as opacity and complexity can make it more difficult to identify potential system weaknesses and ways of resolving them. Furthermore, adaptivity can introduce difficulties in assessing the performance of AI systems over time and preventing unexpected deteriorations of performance. This includes the potential vulnerability of adaptive models to adversarial attacks through data poisoning.
- Increases in the complexity of technology supply chains – and the reliance on off-the-shelf tools and outsourcing in particular – can create challenges in obtaining the information needed to understand and assess system performance.
- Increases in the scale of impacts that can be associated with the adoption of AI can mean that performance weaknesses in AI systems have graver consequences than conventional performance problems. For instance, a poorly performing AI system used to make investment decisions might cause greater financial losses across a firm's accounts when compared to poor judgment on the part of an individual portfolio manager.

3.2.2 System compliance

The second area of concern that we wish to cover in this chapter is system compliance. Firms need to ensure that their AI systems comply with any applicable legal and regulatory requirements. Existing laws and regulations with relevance to the use of AI in UK financial services include:

- the FCA Handbook¹⁷ (eg avoidance of market abuse)
- the Prudential Regulation Authority Rulebook¹⁸ (adequacy of prudential risk management practices)
- equality law (avoidance of unlawful discrimination)
- competition law (avoidance of collusion and other anti-competitive practices)
- data protection law (adherence to relevant principles for managing and processing personal data)¹⁹

Legal and regulatory requirements in relation to the use of data and AI are evolving and subject to ongoing policymaking initiatives at the national and international level. In light of these developments, the list above is not necessarily exhaustive. New relevant requirements may emerge in the future. For instance, the Council of Europe is currently exploring the possibility of a dedicated legal framework for the development and use of AI.²⁰

The four background considerations outlined earlier in the chapter serve, once more, as a framework for understanding why the use of AI might give rise to concerns about the violation of compliance requirements.

¹⁷ See www.handbook.fca.org.uk.

¹⁸ See www.prarulebook.co.uk.

¹⁹ For an overview, see ICO 2020; ICO 2019.

²⁰ Council of Europe 2019.

3 AI challenges and guiding principles

In particular:

- Difficulties in understanding and managing data quality can translate into difficulties in ensuring compliance in various domains. For example, the success of efforts to avoid unlawful discrimination depends on data representativeness and other aspects of data quality. Similarly, the adequacy and accuracy of prudential risk models depends on the quality of the data the models rely on.
- Novel characteristics of models, such as opacity or adaptivity can affect the ability to understand and predict the behaviour of models, making it more difficult to identify and fix problems that may undermine compliance. For example, substantial research efforts are dedicated to understanding how, and whether, we can ensure that highly complex AI systems do not generate outcomes that amount to unlawful discrimination.
- Technology supply chain complexities can create challenges for the commissioning companies to obtain information needed to assess system compliance. This includes information about data quality and provenance, about system design and development, as well as about relevant governance and due diligence arrangements.
- The amplified scale of impacts associated with increased reliance on AI solutions, or firms' reliance on the same AI tools, means that the effects of compliance violations can be more severe. For example, AI systems that lead to discriminatory decisions or inadequate prudential risk assessments may have more widespread impacts compared to violations that may have occurred in the absence of AI systems.

3.2.3 Competent use and human oversight

Setting aside properties of AI systems, their responsible use also depends on how they are deployed and how they interact with human decisions. This leads us to the third area of concern: using AI systems competently and with adequate human oversight.

To ensure competent use, AI system users need to be equipped with an adequate level of understanding that includes aspects concerning the system's development and deployment, such as the system's intended purpose, design, underlying data, performance, the quantification of relevant uncertainties, and deployment practices. This is a pre-condition for the informed exercise of human discretion or intervention in relation to system outputs. In the absence of a thorough understanding, human beings can end up over-relying on or placing undue trust in AI systems' outputs. At the other extreme, people can end up distrusting AI systems excessively, either as a result of their scepticism about AI technologies or their conviction that human judgement and reasoning are superior.²¹ Awareness of an AI system's intended purpose and operational limitations is also important to prevent inappropriate repurposing, ie the use of the system in ways that differ from its intended purpose and for which the system is unfit.

Human oversight is equally relevant to all AI systems. A lack of adequate oversight can mean that performance problems go unnoticed and interventions needed to avert detrimental outcomes fail to occur. Inadequate oversight can take two forms: the absence of human overseers or human overseers being present but failing to exercise their role effectively.

²¹ Leslie 2019, 21.

3 AI challenges and guiding principles

The use of ML, non-traditional data, and automation can lead to increased difficulties in ensuring competent use and human oversight. We encounter these difficulties when understanding and managing data quality, navigating system complexity, opacity, and adaptivity, or adapting to increases in the complexity of technology supply chains. Each of these factors can make it harder for system users and overseers to develop the understanding they need to ensure competent use and effective oversight.

In addition, the adoption of more sophisticated AI systems can raise concerns about system users and overseers' skills and attitudes. Over time, relying on increasingly capable and high-performing systems can lead to de-skilling and unduly passive attitudes among users and overseers. But increases in system complexity and the use of automation often require the opposite: higher levels of technical skills and a more attentive approach to the exercise of oversight.

3.2.4 Providing explanations

The fourth area of concern relates to providing explanations for decisions informed by AI systems. We introduce this area of concern here, briefly, and encourage readers who are interested in an in-depth treatment of AI explanation to consult the guidance produced by the ICO and The Alan Turing Institute on 'Explaining Decisions Made with AI'.²²

AI systems' outputs can form the basis of decisions that affect individuals or organisations. Where this is the case, being able to explain decisions to the affected parties is crucial. In financial services, prominent examples include eligibility or pricing decisions that affect the customers of lenders or insurance providers.

Explanations of decisions have two components: the information that forms the basis for the decision and the decision's logic. Explaining the first component is usually easier than explaining the second. This is because when it comes to decision logic, it is important for explanations to have certain properties, such as being:

- accurate, reflecting the actual mechanisms at work
- intelligible to decision recipients
- meaningful in relation to their objectives (eg if the objective is to enable affected individuals to obtain better decision outcomes, the explanation must make clear the changes required for improvements)
- simple and intuitive, so they are easy to remember, easy to contest if incorrect, and easy to use in order to understand how the actions of decision recipients affect present or future outcomes.²³

²² ICO and The Alan Turing Institute 2020.

²³ Intelligibility does not necessarily guarantee the expected kind of simplicity and intuitiveness.

For example, non-monotonous statistical relationships whose direction changes for different ranges of value in a given input variable can be intelligible to decision recipients when described in lay terms, but may be non-intuitive and difficult to remember.

3 AI challenges and guiding principles

There are several ways in which the adoption of AI can give rise to challenges in providing explanations to decision recipients:

- The novel characteristics of systems, in particular, can make it more difficult to provide explanations. First, system complexity can make it difficult to provide decision logic explanations that are accurate and intelligible. This can be due to inscrutable models or models whose logic can be understood by experts but cannot be easily conveyed to decision recipients. Second, the non-intuitiveness associated with more complex models or non-traditional data can come into tension with decision recipients' expectations of intuitiveness in the logic of decisions. Finally, the use of adaptive systems can affect the meaningfulness of decision logic explanations. For example, an explanation about the conditions that would lead to a different decision outcome may lose its validity as a result of dynamic system changes.
- Increased reliance on third-party providers and other aspects of technology supply chain complexity can make it more difficult for firms to obtain the information needed to give decision recipients explanations. Depending on the nature of the outsourcing arrangements, this difficulty can apply to information about the logic of systems as well as the information used as a basis for a given decision.

3.2.5 Responsiveness

The fifth area of concern is the ability to address customer queries. Like in the case of providing explanations, this area of concern is particularly relevant when it comes to the use of AI systems in customer-facing contexts. In such contexts, the need for responsiveness to customers can relate to:

- **Requests for information:** Customers may seek information about a given business process (including, where applicable, explanations for decisions that affect them, as described in the previous section).
- **Requests for assistance:** Customers may ask for help to use a service or product or seek accommodation for unexpected circumstances (such as adjustments to the repayment schedule for a loan, for example).
- **Requests for rectification:** Customers may want to contest and seek rectification for outcomes that are erroneous or otherwise inadequate, such as decision outcomes that are based on incorrect information.

Across these situations, responsiveness depends on two conditions: the existence of pathways for customers to express the relevant requests and a firm's ability to respond effectively.

The adoption of AI systems can give rise to concerns about both conditions: failures to account for the need for responsiveness when designing AI systems can lead to different kinds of 'computer says no' scenarios. Regarding the existence of pathways for customers to express requests, automated online sales processes or customer service chatbots, for example, might lack the facility for customers to submit queries. When it comes to firms' ability to respond to requests effectively, certain forms of system design may prevent employees from taking required actions, for example by failing to allow for manual corrections. In addition, the ability to respond to customer requests can be affected by system opacity or outsourcing arrangements and other forms of technology supply chain complexity.

3 AI challenges and guiding principles

These factors can pose obstacles to employees' ability to provide customers with the information they seek, to assess the validity of customer requests (eg in the case of erroneous decision outcomes), or to make requested changes to existing decisions and processes.

3.2.6 Social and economic impact

The sixth area of concern is the broadest of all. It relates to potential social and economic impacts associated with AI systems that extend beyond the issues captured by the other five areas just described. Such impacts can include consequences for individuals as well as impacts at the societal level.

At the individual level, concerns involve the ethical acceptability of specific ways of using data, ML, and automation. Several aspects of ethical acceptability are worth highlighting. First, AI can raise concerns about the differential treatment of individuals. This includes the avoidance of unlawful discrimination and other relevant aspects of compliance with legal and regulatory requirements. But AI systems can also enable forms of differential treatment that are considered ethically problematic without violating legal or regulatory compliance requirements, for example because they are perceived to be unfair or result in a lack of access to products or services for certain individuals.

Possible reasons for concerns about ethically problematic forms of differential treatment include the following:

- The data that ML models rely on can reflect existing social, economic, and historical structures and dynamics. As a result, in the absence of mitigating measures, ML can reproduce, reinforce, or even amplify existing patterns of marginalisation, inequality, and discrimination. Moreover, if the datasets on which ML models are trained over- or under-represent certain demographic segments of a population, models can perform better for some groups and worse for others. Beyond these concerns related to baked-in bias and unrepresentative datasets, human biases can arise at any point in the design, development, and deployment of an AI system. This highlights the importance of system performance and competent use when it comes to avoiding problematic forms of differential treatment.
- Non-traditional data and ML methods can make it possible to draw novel kinds of predictive inferences based on previously ignored individual characteristics or behaviour. Some such inferences, when used as a basis for decisions, may result in problematic forms of differential treatment due to hidden correlations with legally protected characteristics. In addition, the use of certain inferences may be considered objectionable on ethical grounds. For example, using the occurrence of spelling mistakes in loan applications as an input variable for risk scoring models could be considered unfair regardless of equality law implications.
- Non-traditional data and ML can enable increases in the granularity of predictive approaches. For example, the use of AI in the context of customer risk profiling may result in an increased stratification of risk profiles, with predictions becoming more specific to different customer types. This can result in certain customers being priced out of the market.

Second, AI can give rise to privacy concerns. Such concerns include but extend beyond issues of legal and regulatory compliance. For instance, ways in which data and inferences about individuals are collected, produced, or shared may comply with data protection law but nevertheless be perceived as excessively intrusive.

3 AI challenges and guiding principles

Contexts in which concerns about privacy can be particularly salient include AI systems that rely on processing biometric and sensor data. For example, AI systems may use biometric data to identify emotions or predict a person's mental state. Apart from concerns about the accuracy of such systems, their use can contribute to feelings of intrusion or surveillance.

Third, the use of AI may raise concerns about individuals' ability to make informed and autonomous choices and exercise meaningful control over their lives. These may be related to issues with providing explanations or responsiveness (see Sections 3.2.4 and 3.2.5). For example, individuals may feel disempowered if they do not receive meaningful explanations for decisions made about them, or if they do not have a way to appeal them. More generally, consumers can find it increasingly difficult to understand how companies collect, use, and share data about them, what inferences they make based on this data, or how they curate the offers they present to them. As a result of such difficulties, making informed decisions can become more difficult or burdensome. This can include deciding whether to use a given service, as well as whether to consent to specific uses of data in the provision of services.

Finally, the use of AI can contribute to concerns about digital exclusion. Adopting AI solutions can be accompanied by an increased reliance on digital communication and service provision. Where this is the case, digital literacy, ownership of digital devices, and internet connectivity become more important, including as a precondition for service use. As a result, social inequities in terms of technology access, levels of confidence and comfort in using technology, and other aspects of 'digital divide' can contribute to a sense of disempowerment and exclusion for certain individuals. When it comes to societal-level impacts, there can be concerns about effects associated with AI systems on the wellbeing, integrity, and functioning of communities, markets and economies, as well as the environment.

Regarding effects on communities, examples include concerns about increases in inequality, eg as a result of AI-enabled differential treatment or digital exclusion as described above. When the needs of underserved and marginalized social groups are not taken into account in the design and use of AI systems, their deployment at scale could exacerbate social inequities and widen existing divides. Another example are concerns about job losses associated with the automation of business tasks previously performed by human employees.

With respect to effects on markets and economies, AI systems can give rise to concerns in a variety of areas, including competition, market integrity, the distribution of economic value, and financial stability. Some of these concerns are issues of legal and regulatory compliance. Yet, there can be concerns that go beyond compliance. For instance, AI in the context of financial trading could contribute to market volatility in ways that do not violate any rules. Similarly, the increased reliance on data-driven technologies in the context of AI adoption can have consequences for competition that go beyond competition law compliance. Relevant mechanisms that could contribute to reduced competition include data monopolies.

In terms of environmental impacts, AI systems and the digital infrastructures underpinning their development and deployment can consume large amounts of energy and other resources. As a result, there can be concerns about the environmental footprint of these technologies. The training of ML models in complex applications such as NLP or computer vision can have particularly large carbon footprints.²⁴

24 Strubell, Ganesh, and McCallum 2019; Henderson et al. 2020; Lacoste et al. 2019.

3 AI challenges and guiding principles

3.3 AI ethics principles

The areas of concern outlined above highlight the need to ensure that AI systems are trustworthy and used responsibly. A growing literature proposes general principles for trustworthy and responsible AI. Together with emerging work on guidelines and technical tools, these principles are a sign of important progress in guiding the responsible design, development, and deployment of AI systems. In this section, we provide a high-level conceptual summary of the AI ethics principles landscape. Figure 5 provides a visual representation of the relationship between background considerations, areas of specific concerns, and principles for trustworthy and responsible AI.

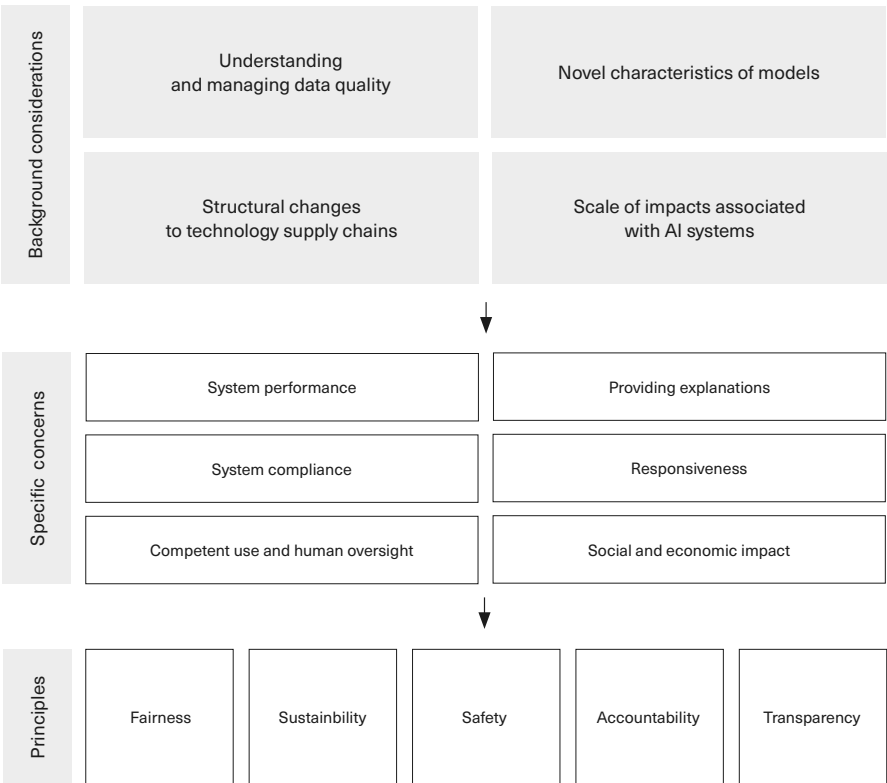


Figure 5: AI challenges and guiding principles

A wide range of organisations have published AI ethics principles.²⁵ They include government bodies, international organisations, professional associations, multi-stakeholder initiatives, and private sector organisations.²⁶ There is a lot of shared conceptual ground and statements converge around similar themes.²⁷

In this section, we follow the UK Government’s official guidelines for the responsible use of AI in the public sector²⁸ and focus on five principles. These are fairness, sustainability, safety, accountability, and transparency. These five principles resonate with most sets of AI ethics principles published to date.

25 For a repository that tracks the publication of AI Principles across different stakeholder groups, see inventory.algorithmwatch.org
26 See, for example, OECD 2019; G20 2019, 20; Select Committee on Artificial Intelligence 2019; High-Level Expert Group on Artificial Intelligence 2019; Université de Montréal 2018; IEEE 2020. Some organisations focused on financial services have also published statements of AI principles. See, for example, De Nederlandsche Bank 2019; Monetary Authority of Singapore 2019; UK Finance and KPMG 2020.
27 See Fjeld et al.
28 Leslie 2019.

3 AI challenges and guiding principles

3.3.1 Fairness

Interpretations of the principle of fairness typically include non-discrimination, the avoidance of unfair bias, and other questions of differential treatment. The principle entails prioritising bias mitigation and the exclusion of discriminatory and unfair influences at all stages of an AI system's lifecycle. It requires that AI systems do not lead to discriminatory or inequitable impacts on affected individuals and communities. To this end, AI systems should:

- be trained and tested on representative, relevant, sufficient, accurate, recent, appropriate, and generalisable datasets (Data Fairness)
- have model architectures that do not include target variables, features, processes, parameters or analytical structures (correlations, interactions, and inferences) which are unreasonable, morally objectionable, or unjustifiable (Design Fairness)
- not have discriminatory or inequitable impacts on the lives of the people they affect (Outcome Fairness)
- be deployed by users sufficiently trained to implement them responsibly and without bias (Implementation Fairness)²⁹

The scope of the principle of fairness can also include broader considerations of social justice. These may relate, for example, to the equitable distribution of the benefits, risks, and burdens associated with AI innovations. Such considerations can be relevant to choices about when, where, and how to build and deploy AI systems.

3.3.2 Sustainability

Acknowledging the diverse nature of the effects that AI systems can have on individuals, communities, economic systems, and the environment, the principle of sustainability recognises the role of a broad range of ethical values in responsible innovation.

Having a shared understanding of relevant ethical values from the outset of an AI project is essential for putting the principle of sustainability into practice. It creates a common vocabulary for informed dialogue, anticipatory reflection, and impact assessment. The UK Government's official public sector guide to safe and ethical AI has consolidated relevant ethical values into four 'support, underwrite, and motivate' (SUM) values. They are anchored in ethical concerns about human empowerment, interactive solidarity, individual and community wellbeing, and social justice. The SUM values are:

- **respect** the dignity of individuals as persons
- **connect** with each other sincerely, openly, and inclusively
- **care** for the wellbeing of each and all
- **protect** the priorities of justice, social values, and the public interest³⁰

The principle of sustainability entails giving due consideration to these values throughout an AI system's lifecycle. At the beginning of an AI project, individuals involved have a responsibility to reflect on the purposes motivating the system's design and use, and to anticipate possible impacts on affected stakeholders.

²⁹ An expanded description of each of these aspects of fairness can be found in Leslie 2019.

³⁰ Detailed explorations of each of these values and how they guide AI designers, developers, and users are available in Leslie 2019.

3 AI challenges and guiding principles

Sustainability also means demonstrating and applying continuous sensitivity to the real-world effects associated with a given AI system. Continual monitoring and re-assessment of these effects are important across the entire AI lifecycle, from project conception through to system retirement.

3.3.3 Safety

The principle of safety is closely connected with that of sustainability. It captures the need to consider the importance of safety throughout the development and implementation of an AI system. AI applications face changing and uncertain environments, unknown unknowns, and adversarial threats. Safety entails, for example, that AI systems are not successfully attacked or misused, that they do not malfunction or cause harm, and that they do not undermine public trust.

Safety is closely linked to properties like security, reliability, and robustness. It refers to the technical soundness of AI systems as well as the soundness of their use. As such, it touches on issues such as proper functioning over time, protecting against unforeseen and unintended consequences, avoiding misuse, preventing incompetent use, and defending against adversarial threats. Fail-safety and fault tolerance in the design and implementation of AI systems are important general considerations in this context.

3.3.4 Accountability

The fourth AI principle is accountability. This concept includes internal accountability, within organisations that are developing and deploying an AI system, as well as external accountability, between these organisations and outside stakeholders.

Governance arrangements and the allocation of responsibilities within an organisation play a key role for internal accountability. They ensure that specific actors within the organisation are accountable for the trustworthiness and responsible use of AI systems. External accountability concerns an organisation's accountability to its customers, regulators, or other outside actors. It can, for example, include questions such as justifying the use or outputs of AI systems to external stakeholders, managing legal liability, and providing mechanisms for appeal or redress.

Beyond these internal and external dimensions, accountability can be broken down into two subcomponents: answerability and auditability.

- **Answerability** means having a continuous chain of human responsibility across the entire AI project lifecycle. This includes clarity over the allocation of human authority in relation to a given AI system across its conceptualisation, design, development, deployment, and retirement. Answerability also means that those responsible can provide clear, understandable, and coherent explanations for AI-supported decisions and the processes behind their production.
- **Auditability** refers to maintaining records of the steps involved in the design, development, and deployment of AI systems. This includes the creation of activity monitoring protocols that enable end-to-end oversight. Auditability ensures that relevant processes and outputs can be reviewed.

Incorporating both these elements of accountability (answerability and auditability) into the AI project lifecycle can be referred to as accountability by design.

3 AI challenges and guiding principles

3.3.5 Transparency

The principle of transparency relates to disclosing information about AI systems. Transparency entails gathering and sharing information about an AI system's logic (often referred to as explainability) as well as about how the AI system was designed, developed, and deployed. External stakeholders, such as customers, regulators, or academics, have an interest in various types of information about an AI system. So do internal stakeholders, such as members of different teams involved in designing, developing, and deploying AI systems.

The principles of transparency and accountability are closely interlinked and reinforce each other. Transparency is a precondition for accountability, since accountability mechanisms depend on the availability of information about an AI system. In a complementary way, accountability operates as a precondition for transparency. Without accountability, commitments to transparent practices would remain unmotivated, unsupervised, and arbitrary.

Transparency and accountability are also overarching AI ethics principles. They act as enablers for the other three AI ethics principles. The ability to address considerations related to fairness, safety, and sustainability depends on the existence of effective accountability mechanisms and the availability of various forms of information to relevant stakeholders.

The principle of transparency plays a fundamental role in AI ethics. It is a logical first step for considering responsible AI innovation in financial services. Chapter 5 examines AI transparency in greater detail, focusing on the role it plays in ensuring and demonstrating the trustworthy and responsible use of AI systems.

4 AI's potential benefits and harms in financial services

The use of AI in financial services can lead to substantial benefits for firms, customers, and markets. Such benefits may result from AI-enabled increases in effectiveness or efficiency. At the same time, AI can also lead to significant harms.

In this chapter, we use an outcome-focused perspective and survey potential benefits and harms across the range of AI use cases in financial services. We look at five areas: consumer protection, financial crime, competition, the stability of firms and markets, and cybersecurity.

4 AI's potential benefits and harms in financial services

Two points about the perspective of the following discussion are worth noting. First, our mapping exercise does not make a case for or against the adoption of AI. Instead, it highlights the importance of ensuring and demonstrating that AI systems are trustworthy and used responsibly.

Second, we use a wide angle with respect to the likelihood of occurrence of benefits and harms. The inclusion of individual benefits or harms is not a judgment about how realistic it is for them to occur. Instead, the overview provides a broad picture of potential impacts, including those that, in the context of today's technology landscape, may seem hypothetical. The aim of this chapter is to provide a basis for thinking through potential future scenarios rather than to provide an evidence-based assessment of AI's impacts.

4.1 Consumer protection

The use of AI can have implications for a range of consumer protection issues. Here, we focus on six of them, namely:

- financial inclusion
- unwarranted denials of service in the context of financial crime prevention
- unlawful discrimination and unfair differential treatment
- mismatches between products and customer needs
- performance of investments
- consumer empowerment

For each of these issues, we will first highlight relevant potential benefits, followed by potential harms. It is worth noting at the outset that there can be trade-offs between the occurrence of some of the positive and negative impacts. For example, using non-traditional forms of data for risk profiling can bring benefits in terms of financial inclusion but also be associated with invasions of privacy. This shows that managing impacts can require value judgments to weigh up competing considerations. The SUM values described in the discussion of AI ethics principles (Section 3.3) can play an important role in facilitating deliberation about such value judgments.

4.1.1 Financial inclusion

Financial inclusion refers to customers having access to relevant financial products or services. A lack of access can take the form of customers being deemed ineligible for a given financial product or service, or the form of such products or services being unaffordable due to their price. The use of AI can positively or negatively affect financial inclusion.

Potential benefits

AI could contribute to increased financial inclusion in three notable ways. First, AI can enable reductions in operational costs that firms may pass on to consumers in the form of lower prices. This can make financial products or services affordable for customer groups for whom they would otherwise be out of reach. The use of AI in the context of robo-advice illustrates this kind of potential impact, increasing the affordability of financial advice.³¹

Second, in contexts where product eligibility or prices depend on the risk profile of customers, improved risk profiling capabilities enabled by AI may translate into favourable eligibility decisions or price reductions for customers that would otherwise lack access.

³¹ See FCA 2019c.

4 AI's potential benefits and harms in financial services

In the context of loans, potential beneficiaries include so-called 'thin-file customers' who can be 'unscorable' due to a lack of documented credit history. The use of non-traditional data and ML can make it possible to establish a risk profile for such customers. Setting aside customers that would otherwise be unscorable, potential beneficiaries also include customers – in loan and insurance contexts alike – for whom more accurate or more granular risk profiling leads to a lower predicted risk when compared to more conventional risk profiling approaches. Resulting improvements in loan terms or reductions in insurance premiums may positively affect certain customers' ability to afford the financial products in question.³²

Third, in contexts where firms pursue strategies of differential pricing³³, ML and non-traditional data may enable more granular or personalised forms of price differentiation. This could potentially result in firms offering lower prices to customers for whom the relevant product or service would have been unaffordable under a less granular/personalised approach to differential pricing.

The potential benefits of the three mechanisms just described are not limited to customers who would otherwise be unable to access a given product or service. Possible beneficiaries of lower prices associated with reductions in operational costs, improved risk profiling capabilities, and more granular forms of price differentiation also include customers who are not at risk of being priced out of the market.

On a general note, it is important to recognise that, where benefits result from the processing of non-traditional forms of customer data, customers' ability to enjoy these benefits can crucially depend on the volume of data that exists about them. Customers who are comparatively 'data poor' – for example, due to a reliance on cash rather than digital payments – may see little benefit compared to similar customers with larger data footprints. This could give rise to ethical questions regarding the societal distribution of the benefits in question, especially since limited data footprints may be correlated with different forms of socio-economic disadvantage.

Potential harms

AI could also contribute to financial exclusion, either through its use for risk profiling or differential pricing. When it comes to risk profiling, poorly performing systems or problems with competent use and human oversight could result in flawed risk profile assessments. As a result, customers who are erroneously considered high-risk can find themselves locked out of the market.

Additionally, more accurate and more granular risk profiling methods, as a flip side of the benefits just mentioned, could result in ineligibility or unaffordability for customer groups for whom conventional approaches would have resulted in more favourable risk profiles. This may be considered particularly problematic in insurance markets where important insurance products (eg travel, motor or health insurance) could become inaccessible or unduly expensive for high-risk customers. The issue of 'uninsurability' is familiar from, for example, home insurance, where high risk of flood is easy to identify with traditional data sources and simple modelling approaches. The use of non-traditional data and increasingly complex models could increase the occurrence of uninsurability.

³² For an empirical analysis of the financial inclusion implications of the use of cash flow data for credit risk profiling, see FinRegLab 2020a; FinRegLab 2019; FinRegLab 2020b.

³³ Here we are referring to the practice – also known as price discrimination – of setting different prices for different customers or customer segments that are not reflective of differences in the actual or expected cost of providing the product or service in question. Contexts in which price discrimination has been observed to be a common practice in financial services include the markets for insurance, mortgages, and cash savings. See FCA 2020; FCA 2019a; FCA 2018a.

4 AI's potential benefits and harms in financial services

Like in the case of benefits, flawed risk profile assessments or less favourable risk profiles can also have significant implications for the prices paid by consumers in cases where access to financial products and services remains unaffected. Customers may be negatively impacted by higher prices without necessarily facing financial exclusion.

When used for purposes of differential/personalised pricing, AI could entail higher prices for certain types of customers. More granular assessments of willingness to pay for different customer segments could result in prices moving closer to the limit of what some customers are willing or able to pay. As noted in the discussion of benefits above, this may entail lower prices for customers whose willingness to pay is comparatively low. At the same time, it can result in higher prices – compared to a more conventional pricing strategy – for customers whose willingness to pay is comparatively high.³⁴

4.1.2 Unwarranted denials of service in the context of financial crime prevention

Screening and monitoring mechanisms aimed at preventing money laundering, fraud, and other forms of financial crime can lead to denials of service that are unwarranted. In the context of know-your-customer (KYC) procedures during the onboarding of new customers, for example, customers may be turned away due to mistaken identity or due to models with excessive false positive rates. In the context of transaction monitoring, false positives can lead to customers mistakenly being denied the execution of transactions or withdrawal of funds.

Potential benefits

AI can enable the development of screening, monitoring, and detection systems that perform better than traditional systems. This can result in fewer unwarranted denials of service. Possible benefits include a lower number of customers being affected by de-risking, ie firms deciding not to enter or to end relationships with certain customer groups in a wholesale manner to reduce their risk exposure. In addition, AI can help make the impact of false positives less severe by introducing novel and easier ways for customers to take remedial or corrective action. For example, the use of automated text message exchanges when a transaction is flagged as potentially fraudulent has already made it much easier for customers to confirm legitimacy and ensure that the transaction goes ahead.

Potential harms

Conversely, poorly performing AI systems could lead to an increased occurrence of unwarranted denials of service (including possible increases in customers affected by wholesale de-risking) compared to traditional systems and approaches. Moreover, there are ways in which changes in the technological infrastructure behind screening, monitoring, and detection systems could make the consequences of unwarranted denials for customers more severe. For example, the increased reliance on shared data and systems across firms, potentially facilitated by firms' reliance on the same third-party providers, could mean that customers affected by an erroneous KYC assessment are not just turned away by an individual firm but find themselves being locked out across the market.

³⁴ For more detailed and technical discussions of potential positive and negative implications of price discrimination from the perspective of consumer welfare, see OECD 2018; Office of Fair Trading 2013a; Office of Fair Trading 2013b.

4 AI's potential benefits and harms in financial services

4.1.3 Unlawful discrimination and unfair differential treatment

Decisions that entail differential treatment of customers can be at risk of exhibiting forms of discrimination that are unlawful under the Equality Act 2010 and related legislation. In addition, there are forms of differential treatment that can be considered unfair without necessarily violating legal non-discrimination requirements. For example, as the FCA's work has highlighted, differential treatment can be problematic if it is at odds with the interests and needs of vulnerable consumers.³⁵

Considerations of unlawful discrimination and unfair differential treatment of customers can be relevant in a diverse range of contexts, including product eligibility, product pricing (including risk-based and differential/personalised pricing) and some of the forms of denials of service related to financial crime prevention just described.³⁶

Potential benefits

AI can help to avoid the occurrence of unwanted forms of differential treatment. First, new forms of data and analytical methods can make it easier to detect such differential treatment in existing decision-making processes. Second, the use of ML and non-traditional data can contribute to finding effective ways of mitigating the occurrence of unwanted differentials in model outputs. Finally, the increased reliance on models, data, and responsible forms of automation can help to reduce the occurrence of unwanted forms of differential treatment that result from biased human judgments.

Potential harms

At the same time, the use of novel modelling approaches can unintentionally contribute to unwanted forms of differential treatment. First, there can be increased difficulties in ensuring that model outputs do not entail unlawful discrimination or forms of differential treatment that are considered unfair. As discussed in Chapter 3, such difficulties can arise, for example, as a result of model opacity, the quality of the data used, or complexities in technological supply chains that hinder commissioning firms' access to information. Where automation is involved, this can amplify the scale of any associated harmful impacts.

Second, the use of more complex models and non-traditional data can enable new forms of differential treatment that may be considered unfair. In the context of creditworthiness assessments, for example, new approaches to risk profiling for loan applicants may lead to the identification of previously ignored characteristics, such as spelling errors on an application, that are predictive of risk.³⁷ The use of some of these characteristics may be considered ethically objectionable regardless of whether it has implications that are problematic from an equality law perspective. New forms of differential treatment that are considered unfair could also arise in the context of differential/personalised pricing. The use of more complex models and non-traditional data in this context could, for example, result in (i) greater price differentials, (ii) increases in the number of consumers affected by high prices, or (iii) pricing patterns that are less transparent or based on characteristics that are more difficult for consumers to control. These are three of the six dimensions in the FCA's framework for assessing concerns about fairness in relation to price discrimination.³⁸

³⁵ FCA 2018b; FCA 2018c.

³⁶ It is worth noting that decisions that do not concern customers but, for example, a firm's employees or job applicants can also be affected by unlawful discrimination and unfair differential treatment. The potential benefits and harms discussed here therefore also apply to the use of AI systems for CV screening, hiring decisions, or other internal business decisions.

³⁷ Lee and Singh 2020.

³⁸ FCA 2018d.

4 AI's potential benefits and harms in financial services

4.1.4 Mismatches between products and customer needs

Mismatches between the needs of customers and the products they buy can occur as a result of flawed marketing practices or the design of products and services. Either case can lead to the sale of products that are not needed, poor value, misaligned with a customer's level of financial literacy, or otherwise in conflict with customers' best interest. Using AI in marketing and product design can help prevent as well as contribute to such mismatches.

Potential benefits

The use of AI in marketing can enable beneficial forms of targeting that make it easier for customers to find products that match their needs. Conversely, AI can help prevent misdirected marketing and mis-selling.³⁹

Used in the context of product and service design, AI can enable new forms of product customisation which tailor products to the needs of individual customers. Possible examples include investment products that rely on AI to design portfolios according to customers' investment goals and values, or using AI to provide tailored financial advice.

Potential harms

Poorly performing AI systems used in marketing or product customisation can contribute to mismatches between products and customer needs. For example, a poorly performing AI system used to provide financial advice may lead to customers receiving tailored advice that is less aligned with their investment goals compared to the advice that they would have received conventionally.

In addition to such unintended consequences, enhanced capabilities for targeting can also be used for deliberate ill-directed marketing to intentionally promote the sale of products that are at odds with the needs and interests of customers. From a governance perspective, challenges can arise from the fact that information and insights that can play a role in preventing misdirected marketing can also be used for the opposite purpose. For example, information that serves to identify and protect vulnerable customers can also be used to promote the sale of poor-value products to consumers in financial distress. Governance challenges can also arise from AI-enabled structural changes in marketing supply chains. New forms of affiliate marketing, for instance, may make it more difficult for firms to exercise control in preventing ill-directed marketing.

4.1.5 Performance of investments

Within asset management – including pensions, savings, and investments – the interests of customers can be harmed by conflicts of interest, excessive charges, suboptimal returns, or unexpected financial losses. Using AI for purposes of portfolio management and trade execution has the potential of having positive as well as negative impacts when it comes to preventing such outcomes.

Potential benefits

Benefits can arise in two contexts. First, increased returns could result from the use of AI to improve investment strategies. Second, benefits can arise from using AI to achieve more efficient trade execution, for example through executing trades in ways that are faster, less expensive, or have less market impact. In either context, improvements may come about, for instance, as a result of using ML and non-traditional data for modelling purposes or reducing the impact of known psychological biases from the investment process through the increased reliance on technology.

39 Money and Mental Health Policy Institute 2019; FCA 2015.

4 AI's potential benefits and harms in financial services

Potential harms

Conversely, the use of AI could be accompanied by model or system weaknesses that result in poor investment or trade execution strategies. Once again, the various challenges highlighted in the previous chapter mean that the adoption of ML, non-traditional data, and automation can come with increased difficulties to identify and protect against such weaknesses. In addition, technological innovations in areas such as algorithmic trading could increase the risk of unanticipated market dynamics caused by the interaction of AI systems in the market. Such interactions may result in financial losses and negatively impact investment returns.

4.1.6 Consumer empowerment

Some of the impacts mentioned so far have potential implications for consumer empowerment. However, there are several aspects of consumer empowerment that fall outside the categories already mentioned.

Potential benefits

Three kinds of potential positive impacts are worth highlighting. First, using AI may enhance the availability and accessibility of services. This includes the availability of and access to primary services (eg banking services) as well as improvements in the delivery of customer service. For example, AI systems can contribute to new forms of remote and around-the-clock access to customer service as a result of tools such as AI-enabled chatbots.

Second, for financial products where terms depend on customer risk profiles, such as insurance and loans, AI can introduce novel ways for customers to understand and manage their risk profiles. A prominent example for this is the use of telematics sensors or personal health devices in the context of motor and health insurance, respectively. In addition to giving customers new insights about risky behaviours, the data derived from these devices can empower customers to improve their risk profiles through the choices that they make in their daily lives.

Third, AI can enhance consumers' control over their finances by facilitating financial planning. For instance, the use of ML in the context of personal finance and budgeting apps can help anticipate expenses and suggest saving patterns suitable to individual circumstances.

Potential harms

On the flipside, there are several ways in which AI can contribute to consumer disempowerment. First, mirroring the benefits of availability and accessibility just described, the adoption of new digital modes of delivering products, services, and communication with customers could result in certain customers experiencing digital exclusion. If digital modes replace more conventional forms of delivery, customers with limited access to or limited experience using digital devices or other relevant technologies could find it difficult to access products and services or feel less confident in their dealings with financial service providers. Relatedly, the reduction in human-to-human contact that can come with AI-enabled digital solutions may lead to a perceived loss of meaningful social interaction.

In addition, there are various forms of disempowerment that could arise specifically in contexts where AI is used to inform or make decisions about customers. This includes the impact of potential AI-related obstacles to customers' ability to understand the basis of decisions. As highlighted previously, such obstacles may be due to (i) difficulties around providing explanations that are accurate, intelligible, meaningful, and sufficiently simple and intuitive or (ii) a lack of pathways for customers to request relevant information.

4 AI's potential benefits and harms in financial services

Resulting difficulties in understanding the basis of decisions could have the following consequences:

- Customers may struggle to anticipate the relationship between their behaviour and decision outcomes. This may prevent them from making informed behavioural choices and identifying legitimate ways of achieving favourable decision outcomes (eg smoothing their monthly expenditures to improve their credit scores).
- Customers may find it difficult to make informed decisions about data consent, due to a lack of understanding of the relationship between the information contained in the data they are providing and decision outcomes.
- Customers may be unable to identify instances of erroneous or otherwise unwarranted decisions, for example in the case of decisions about product eligibility, pricing, or denials of service.

In addition, where erroneous decisions have been identified, the reliance on AI could also affect customers' ability to obtain rectification. For example, automated systems might have design features that do not offer ways for customers to request corrections or prevent employees from overriding the system and taking corrective action.

Finally, there are potential harms that are specifically related to the reliance on new types of information in making decisions about customers. Where non-traditional data is used, aspects of customers' lives that were historically irrelevant can potentially become a determining factor for product eligibility and prices. Such previously 'insulated' aspects of life may include financial information (eg spending patterns or brand choices and other aspects of purchasing decisions) as well as non-financial information (eg social media behaviour). Despite the potential benefits associated with the use of such information (eg in terms of financial inclusion), this kind of expansion of financial relevance into previously insulated areas of life can raise ethical questions. In particular, it could reduce personal freedom and impose new psychological burdens on consumers. For example, customers may feel constrained in their shopping choices or in how they communicate on social media based on the inferences about their credit or insurance risk profile that may result.

As a potential added complexity, the perceived need to use products and services that rely on data that makes previously insulated aspects of life financially relevant could be unequally distributed between different groups of consumers. In particular, consumers who struggle to access loan or insurance products could perceive themselves to be under strong pressure to consent to the use of the types of non-traditional data in question. They may feel they have little choice but to accept the associated reduction in privacy in exchange for potential benefits in terms of product eligibility or affordability. Individuals that might find themselves in this situation can be thin-file loan customers but also vulnerable or economically disadvantaged consumers (regardless of their credit file). Despite the potential benefits at stake for such individuals, there could be contexts in which the perceived pressure to accept 'privacy poverty' in exchange for these benefits is considered disproportionate and problematic, especially if it is associated with broader forms of disadvantage. There could also be concerns that the societal acceptance of 'privacy poverty' as a solution to the challenges faced by disadvantaged consumer segments may mean that other, less privacy-intrusive forms of mitigating financial exclusion remain unexplored.

4 AI's potential benefits and harms in financial services

4.2 Financial crime

AI could have positive and negative consequences for the prevention of financial crime. This includes fraud, money laundering, and – in the context of securities markets – insider trading and market manipulation.

Potential benefits

AI can enable improvements in systems used to prevent financial crime. This includes detection systems, which may become more effective and efficient due to the use of ML, non-traditional data, or automation. For example, data-driven approaches to detecting fraud or money laundering that rely on regularly updated AI models may have a better detection rate than systems based on explicitly programmed rules that do not adapt over time.

Potential benefits can also result from improvements in systems that provide safeguards rather than serving to detect financial crime. In the context of fraud prevention, for instance, AI can enable new and more reliable customer identification methods, including innovative forms of sensor-enabled authentication. AI solutions can also be used to provide information to customers that have been identified as being susceptible to fraud, helping them to be aware of risks.

Potential harms

On the side of potential harms, two types of impact are worth distinguishing. First, unexpected weaknesses in the performance of AI systems used to detect or provide safeguards against financial crime could make such systems less effective or reliable compared to more conventional systems. The vulnerability of AI systems to adversarial attacks can be particularly relevant in this context. For example, model complexity may make it difficult to identify weaknesses that allow adversarial actors to develop detection evasion strategies. And where adaptive models are used, AI systems could be susceptible to data poisoning attacks.

Second, when it comes to abusive trading practices in securities markets, AI systems could contribute to the occurrence of market abuse. For instance, AI trading systems could draw on information that is material non-public information, resulting in decisions that amount to insider trading. Similarly, AI systems could pursue trading strategies that amount to prohibited forms of market manipulation. For either possibility, the use of AI could facilitate the intentional pursuit of market abuse, with a system's aims being concealed by complexity. Market abuse on the part of AI-enabled trading systems could also occur unintentionally, due to an insufficient understanding of the kinds of information that a system draws on or the strategies that it may develop – caused, for example, by model complexity, model adaptivity, or the use of non-traditional data.

4.3 Competition

When thinking about the impact of AI from a competition perspective, it is worth distinguishing two types of impacts: (i) direct effects on competition caused by AI systems and (ii) indirect effects on competition due to AI technologies bringing about changes in the structure of markets.

Potential benefits

Direct pro-competitive effects on market outcomes can arise in a range of contexts, including financial trading and retail pricing. In financial trading, AI-based trading systems could, for instance, contribute to a reduction of bid-ask spreads or otherwise enhance the efficiency of markets. In the context of retail pricing, for instance in insurance markets, the use of AI can contribute to intensified competition through improved capabilities of pricing systems.

4 AI's potential benefits and harms in financial services

This includes systems used for dynamic pricing (ie adjusting prices in response to changes in market conditions such as competitors' prices) and differential/personalised pricing.⁴⁰

Beyond such direct pro-competitive effects, the increased availability of AI systems and their capabilities can impact market dynamics in ways that have indirect pro-competitive consequences. For example, AI might enable a greater number of market actors to perform business tasks traditionally conducted by a narrower set of actors. In asset management, for instance, AI may make it feasible for buy-side firms to develop their own order execution solutions, decreasing the range of tasks that are the exclusive domain of brokers. At a more general level, ML and automation can reduce the cost for firms to develop certain technological capabilities or solutions. This reduction in costs can lower barriers of entry and lead to increased competition.

Potential harms

With respect to potential harms that take the form of direct effects on competition, the use of AI in retail pricing or securities trading systems could contribute to collusive market outcomes. It can also make it more difficult to detect and prevent such outcomes. Two kinds of collusion scenarios are worth distinguishing:

- AI systems could be used to facilitate explicit collusion by making it easier for colluding parties to monitor and enforce compliance with agreed strategies or to conceal collusive practices behind system complexity.
- AI systems could lead to collusive or anti-competitive market outcomes without there being explicit agreements or intentions. System complexity and adaptivity, for example, could make it difficult to ensure that systems used for dynamic pricing do not pursue strategies that amount to tacit collusion.⁴¹

AI-enabled pricing systems could also have effects that dampen competition in other ways. For example, improvements in capabilities for differential/personalised pricing, their potential pro-competitive effects notwithstanding, could make it easier for dominant firms to pursue forms of predatory pricing that marginalise competitors.⁴²

In terms of indirect effects on competition, there are conceivable mechanisms whereby the growing significance of AI-related technologies could also result in increased barriers to entry and higher market concentration. The development of high-performing AI solutions can require significant financial investments, know-how, and access to data. In addition, due to the importance of data, the commercial advantage associated with AI solutions can be self-reinforcing over time. For example, a market leader in customer profiling may be able to attract a greater number of new customers compared to competitors and use data about these customers to achieve an even greater advantage in profiling capabilities. As a result of these factors, there could be contexts in which firms that have an advantage in terms of proprietary data, know-how, and required financial resources or that act in a first-mover role are able to develop a market lead that places them out of reach of effective competition. These potential dynamics apply to primary financial services markets as well as to markets of third-party suppliers and intermediaries.

⁴⁰ For discussions of potential pro-competitive effects of differential pricing, see, for example, FCA 2019b; Office of Fair Trading 2013b.

⁴¹ For more detailed discussions of different scenarios of algorithmic collusion, see CMA 2021; CMA 2018; World Economic Forum 2019; OECD 2017; Autorité de la Concurrence and Bundeskartellamt 2019; Ezrachi and Stucke 2016; International Competition Network 2020.

⁴² See CMA 2021; FCA 2019b.

4 AI's potential benefits and harms in financial services

4.4 Stability of firms and markets

When it comes to the stability of firms and markets, the use of AI can have implications for ensuring that micro-prudential and macro-prudential risks are well understood and managed appropriately. In addition, certain uses of AI could actively contribute to increased market volatility and related dynamics that may be problematic from a stability or market integrity perspective.

Potential benefits

AI could enable improvements to models used to understand risk and inform risk management practices. Examples include models of investment, credit, or insurance risk, models used in stress testing, and models used for the calculation of capital requirements. Increases in the accuracy or reliability of such models can translate into more effective and more efficient approaches to risk management.

Potential harms

Relying on AI systems to understand risk and to inform risk management practices could also have drawbacks. AI can make it more difficult to identify weaknesses in risk models – for instance, due to model opacity or difficulties in understanding data quality. Such weaknesses could undermine the effective and efficient management of prudential risks, leading to worse outcomes compared to more conventional modelling techniques. As a result, firms may unwittingly find themselves exposed, for example, to excessive levels of credit or liquidity risk.

Setting aside the use of AI to measure risk, AI could actively contribute to risks and market volatility in ways that can be problematic from the perspective of financial stability or market integrity. This possibility is particularly relevant when it comes to the use of AI in financial trading, including in the context of high-frequency trading. Factors such as model opacity, data quality issues, or a broader lack of anticipatory awareness could contribute to failures to foresee and prevent problematic market impacts or dynamics caused by individual systems or the interaction between different systems in the market. Examples of dynamics that could become more likely as a result include intensified herding and flash crashes.⁴³ Factors associated with the structure of markets and technology supply chains, such as increases in market actors' shared reliance on the same sources of data, systems or strategies – for example by using the same third-party providers – could contribute to or amplify the scale of problematic impacts.

4.5 Cybersecurity

In discussing this area, we define cybersecurity to include the protection of a firm's digital infrastructure against information theft (eg intrusions aimed at stealing customers' personal data or the firm's commercially sensitive information) as well as against attacks aimed at incapacitating a firm's systems or manipulating their outcomes.⁴⁴

Potential benefits

When used in cybersecurity defence systems, AI can make them more effective. For example, the use of adaptive models can contribute to ensuring that systems' effectiveness in detecting and documenting anomalies that could be indicative of threats does not deteriorate over time. The use of AI may also enable faster and more effective responses to attacks, for instance through the automated blocking of certain IP ranges in the case of distributed denial of service (DDoS) attacks. Setting aside its use in cyber defence systems, AI can also help detect vulnerabilities in other operational systems, enabling firms to address them before they can be exploited through cyberattacks.

⁴³ For relevant discussions of these concepts, see World Economic Forum 2019; Kirilenko and Lo 2013; Government Office for Science 2011; Government Office for Science 2012.

⁴⁴ We exclude fraud prevention from the discussion of cybersecurity, as it is covered in the section on financial crime.

4 AI's potential benefits and harms in financial services

Potential harms

The use of AI in cybersecurity defence systems can also be accompanied by unanticipated system weaknesses that make them more vulnerable compared to more conventional systems. Possible reasons include difficulties in identifying and understanding vulnerabilities due to system complexity and opacity. In addition, defence systems that rely on adaptive models, for example, can be susceptible to forms of adversarial attack that do not exist in simpler non-adaptive systems.

The use of AI in cybersecurity defence systems aside, the increased adoption of AI in other areas of business can give rise to new cybersecurity risks. Such new risks can come about in two ways. First, the adoption of AI can be accompanied by an increase in connected digital infrastructures within firms, resulting in new and expanded attack surfaces. Second, the introduction of AI-based systems to replace other digital but simpler systems can be accompanied by difficulties in managing system vulnerability. Once again, complexity and opacity can make it more difficult to identify and understand system vulnerabilities; and the introduction of AI solutions can be accompanied by new types of vulnerabilities, for example to data poisoning attacks or privacy attacks in the form of model inversion or membership inference attacks.⁴⁵

-

In concluding this chapter, it is important to note that the benefits and harms outlined here do not exhaust the potential positive and negative impacts associated with AI technologies. In particular, the five thematic areas discussed above leave out possible impacts on individuals other than customers as well as some types of societal-level impact.

While we discussed potential harms to consumers, there are various AI use cases that can impact other individuals outside or within the firms employing AI technologies. For example, financial services firms might use AI systems to streamline their recruitment processes. Such systems could contribute to unlawful discrimination and unfair differential treatment among the firm's applicants. Internal-facing AI applications can also give rise to other potential harms to individuals. For instance, using AI to monitor the behaviour of employees to assess their productivity can lead to excessive surveillance and infringements of privacy.

Our discussion also has not been exhaustive of potential harms at the societal level. For example, the amplification of patterns of economic inequality or structural biases – eg in the eligibility or pricing of financial products – might be considered problematic in ways that go beyond the harm experienced by individual consumers. Similarly, technology-related impacts on privacy, for instance, could have implications for trust and the nature of social relationships with significance at a collective level that extends beyond the individual-level experience of reduced privacy.

⁴⁵ For an explanation of model inversion and membership inference attacks, see ICO 2020.

5 **AI transparency**

Transparency, one of the AI ethics principles introduced in Chapter 3, is key to responsible AI innovation. Transparency is crucial to ensuring and demonstrating that AI systems are both trustworthy and used responsibly.

In this chapter, we discuss the forms that transparency can take, the purposes it can serve, and relevant practical considerations.

5 AI transparency

The first section of this chapter introduces a general framework for distinguishing between different forms of transparency, including a high-level distinction between system transparency and process transparency. The middle sections elaborate on these two forms of transparency. The final section considers trade-offs and countervailing considerations that can conflict with transparency.

Access to information about AI systems is vital when it comes to **ensuring** that AI systems are trustworthy and used responsibly. Addressing the concerns and preventing the potential harms described in previous chapters requires information being available to relevant individuals within firms that may be involved in designing an AI system, developing it, deciding about its deployment, or using it. In addition, certain information may need to be available to customers, for example to enable them to understand decisions made about them and to prevent consumer disempowerment.

Transparency is also critical for **demonstrating** trustworthiness and responsible use, be it to corporate boards, shareholders, customers or regulators. This second role of transparency is no less important. Merely ensuring trustworthiness and responsible use may not be enough to overcome obstacles to adoption. Without reliable evidence to support claims of trustworthiness and responsibility, customer and stakeholder distrust may prevail. The ability to demonstrate trustworthiness and responsibility is therefore a separate pre-condition for successful innovation.

5.1 Defining transparency

AI transparency can be understood as relevant stakeholders having access to relevant information about a given AI system. This general definition raises two immediate questions:

- What types of information are relevant?
- Who are the relevant stakeholders?

Reflection on these two questions gives rise to a third:

- Why are stakeholders interested in information about an AI system?

These three questions can be thought of as the 'what', the 'who' and the 'why' of AI transparency.

5.1.1 Relevant information (the 'what')

Two broad categories of information can be distinguished:⁴⁶

- **System logic information:** Information that relates to the operational logic of a given AI system or, in colloquial terms, information about the system's 'inner workings.' Examples include information about the input variables that a system relies on or information about the relationship between the system's inputs and outputs.
- **Process information:** Information that relates to the processes surrounding the AI system's design, development, and deployment. Examples include information about data management practices, assessments of system performance, quality assurance (including of data) and governance arrangements, or the training of system users.

⁴⁶ While there are various conceptual approaches to describing and categorising different types of information, the literature on trustworthy and responsible AI generally agrees on the joint importance of system logic information and process information. See ICO and The Alan Turing Institute 2020; High-Level Expert Group on Artificial Intelligence 2019; European Banking Authority 2020; Brundage et al. 2020.

5 AI transparency

These two categories of information give rise to two forms of transparency:

- **System transparency:** Stakeholders having access to system logic information
- **Process transparency:** Stakeholders having access to process information

We look at these two forms of transparency in subsequent sections. Before doing so, we consider who may be interested in information about an AI system.

5.1.2 Relevant stakeholders (the 'who')

We can split those who may have an interest in system or process transparency into two categories:

- **Internal stakeholders:** Individuals within the firm that is employing the AI system. This includes people involved in the design, development or procurement of the AI system. It also includes individuals who make decisions about its deployment, operate the system, manage external communications, or perform corporate governance and oversight functions. Examples include members of development or procurement teams, risk and compliance teams, audit teams, senior management, company boards, operational teams using the AI system, and customer service teams.
- **External stakeholders:** Actors external to the firm employing the AI system that have a significant relationship with the firm deploying the system or may be affected by the AI system's use. This can include customers, shareholders, regulators, academics, and members of the public.

Based on these two categories of stakeholders, we can make a second distinction in mapping out different types of transparency:

- **Internal transparency:** Information being accessible to internal stakeholders
- **External transparency:** Information being accessible to external stakeholders

This second distinction intersects with the first one, between system and process transparency. System logic information or process information can be accessible to internal stakeholders, external stakeholders, or both. The resulting four-fold transparency typology is summarised in Figure 6.

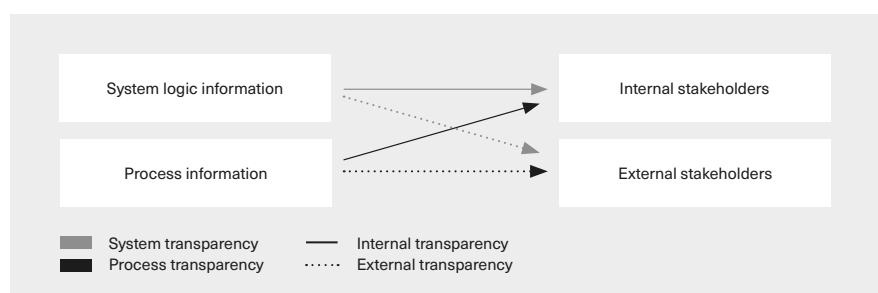


Figure 6: Transparency typology

5 AI transparency

5.1.3Reasons for accessing information (the ‘why’)

Not all types of information about an AI system will be equally important to all types of stakeholders. The reasons that underpin stakeholders’ interests in information about a given system (ie their ‘transparency interests’) are important in determining the types of information they may seek access to. When these reasons differ between stakeholders, the definition of what constitutes relevant information can change. For example, customers faced with an AI system used to make credit eligibility decisions may wish to understand the impact of, say, a 3% pay raise on their credit eligibility. The answer to this question can involve types of information that may not be relevant to the transparency interests, say, of regulators, which may be motivated by the goal of understanding different aspects of system performance and compliance.

Stakeholders’ transparency interests can differ even when their reasons for seeking information are the same. For example, a risk and compliance officer may seek information about an AI system for the same reasons and look for answers to the same questions as a different internal stakeholder (eg a customer service representative) or an external stakeholder (eg a member of the public). Each of these stakeholders, however, might expect different levels of detail.

Figure 7 summarises our discussion about AI transparency so far. Next, we look at the concepts of system transparency and process transparency in greater detail. We focus on information in the context of AI systems that rely on ML models (especially supervised machine learning).

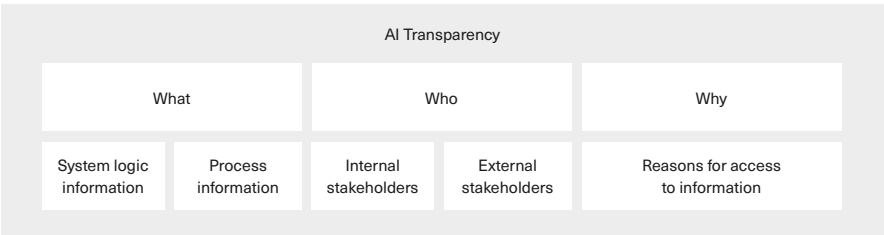


Figure 7: Summarising the three key questions of transparency

5.2System transparency

This section provides an overview of (i) what types of information fall under the category of system transparency, (ii) why different stakeholders can be interested in them, and (iii) how such information can be obtained and communicated.

In discussing ‘what’, the first subsection considers the fact that AI systems can exhibit varying degrees of interpretability. The second part, discussing ‘why’, shows that access to system logic information can be important in relation to all of the six areas of concern identified in Chapter 3.⁴⁷ The third part, in considering ‘how’, discusses different ways of obtaining system logic information and communicating it intelligibly and meaningfully.

⁴⁷ As a reminder, these are (1) system performance, (2) system compliance, (3) competent use and human oversight, (4) providing explanations, (5) responsiveness, and (6) social and economic impact.

5 AI transparency

Figure 8 shows this section's structure.

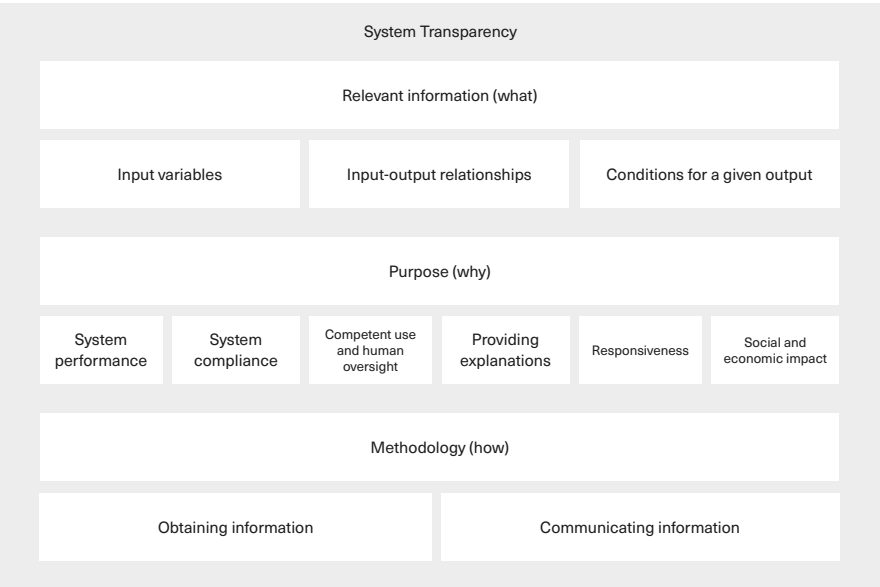


Figure 8: System transparency

5.2.1 Relevant information (the 'what')

System transparency refers to access to information about the operational logic of a system. In the case of simple systems, system logic information can be inferred purely from a system's formal representation. It encompasses:

- (1)

The input variables that a given system relies on (What are the types of information that the system uses in operation?)⁴⁸
- (2)

The way in which the system transforms inputs into outputs (What is the relationship between input variables and system results?)
- (3)

The conditions under which the system would produce a certain output (For what values of the input variables would the system return a specific value of interest?)

To illustrate how these three types of information can be inferred from the formal expression of a simple system, let us assume that the following linear model calculates a person's credit score (Y) as a function of their weekly income (X):⁴⁹

$$Y=200+0.5X$$

The variable on the left side of the equation, Y, is the model's target or output variable. In our hypothetical example, it represents a person's credit score. The right side of the equation contains the determinants of the value of Y: the input variable X (weekly income, in our example), the coefficient 0.5, and the constant term 200.

48

This is different to the values for these variables (ie the data) that the system relies on in operation. The latter falls within the category of process information discussed in Section 5.3.

49

In reality, credit scoring models are much more complex.

5 AI transparency

This simple equation provides answers to all three questions outlined above. More specifically, it shows that:

- the model relies on a single input variable, namely X (weekly income)
- the model transforms inputs into outputs by multiplying the value of the input variable by 0.5 and then adding 200
- in order for the model to yield an output value (credit score) of 600 (for example), the value of X (weekly income) would need to be £800.

Given that it is possible to infer these three types of information from its formal expression, this simple model is fully interpretable.⁵⁰

Our example is significantly simpler than the models typically seen in financial services. Yet, many of the models that financial services firms use – and have traditionally used – meet the definition of interpretable models. Their interpretation may require a higher level of mathematical knowledge, but their structure makes it possible to infer answers to the three questions above based on a formal model expression.

The adoption of AI solutions often involves the reliance on models that are much more complex than the kinds of models that have traditionally been used in financial services, let alone the simple example just described. The increases in model complexity enabled by ML methods can entail a decrease in or loss of model interpretability.⁵¹ It will generally be possible to identify the input variables that ML models rely on (ie information in category (1) above). Yet, model complexity can make it difficult to understand – from a formal expression of the model – how inputs are transformed into outputs (information in category (2)) or the conditions under which the model yields a specific output (information in category (3)).

Decreases in interpretability can take two forms, corresponding to the two types of opacity introduced in Chapter 3. First, as model complexity increases, interpreting models requires greater technical skills. This possibility of opacity due to non-expertise shows that interpretability is a relative concept. Whether an AI system is considered interpretable can depend on the level of technical expertise of those trying to understand it.

Second, model complexity can take forms that make AI systems inscrutable, affecting their interpretability regardless of expertise. In such cases, experts may still be able to give partial answers to the question of how the model transforms inputs into outputs from a formal representation of it – for example, by providing a high-level description of the model's structure. Yet, these partial answers fall far short of the complete understanding that can be gained from the formal expression of the simple linear model above.⁵²

The lack of interpretability of certain types of models does not necessarily mean that adequate forms of system logic information are unobtainable for these AI systems. Instead of obtaining information from the formal expression of models, system logic information can also be obtained indirectly, by using auxiliary strategies and tools. The AI literature refers to these strategies and tools as **explainability methods**.

50 Note that the X in our example represents a variable that corresponds to an easily understandable real-world property (weekly income). In cases where X takes the form of an engineered variable that is difficult to make sense of, the model might be considered transparent without being fully interpretable.

51 As mentioned earlier, increases in complexity include increases in dimensionality (ie the number of input variables that a model relies on) as well as increased complexity in variable relationships, such as non-linearity, non-monotonicity, and discontinuity.

52 Neural networks used for computer vision tasks often exhibit inscrutability. For an overview of model types that may be considered interpretable or uninterpretable, see ICO and The Alan Turing Institute 2020.

5 AI transparency

The explainability methods that exist today play an important role in shedding light on inscrutable models' inner workings – and indeed, in helping us understand any kind of model. However, they cannot fully compensate for the information that can be obtained from interpretable systems. Depending on context, the inability to fully scrutinise uninterpretable models can speak against their adoption and use in financial services.

Section 5.2.3 below covers explainability methods and different ways of obtaining and managing system logic information in more detail. Before we dive deeper into these topics, however, it is helpful to consider the different purposes that system logic information can serve.

5.2.2 Purpose (the 'why')

Access to system logic information can serve to address relevant concerns (ie ensuring trustworthiness and responsible use) as well as to provide assurance about possible concerns (ie demonstrating trustworthiness and responsible use). We illustrate, using a few examples, the importance of system logic information in relation to each of the six areas of concern we identified in Chapter 3. Figure 9 provides a reminder of these areas.



Figure 9: Areas of concern

System performance: System logic information can be vital to understanding and improving the effectiveness, reliability, and robustness of AI systems. Where testing during system development reveals shortcomings, the analysis of input-output relationships can help identify possible improvements. Knowledge of input-output relationships can also be crucial when assessing the extent of possible performance issues that may arise during deployment. Stakeholders that may be interested in system logic information for these reasons include those involved in or making decisions about the development and use of AI systems as well as those seeking assurance about an AI system's performance (including evaluation).

System compliance: Knowledge of the input variables that a system relies on and other aspects of system logic can be crucial to ensuring compliance with legal and regulatory standards and rules. For example, an understanding of system logic can be critical to avoiding unlawful discrimination; ensuring the adequacy of systems used in prudential risk management; assessing the extent to which trading systems may entail risks of insider trading or market manipulation; determining the potential of anti-competitive outcomes in systems used for pricing; or avoiding the unlawful processing of personal data. As in the case of system performance, stakeholders that may be interested in system logic information for these reasons include those involved in or making decisions about the development and use of AI systems as well as those seeking assurance about system compliance.

5 AI transparency

Competent use and human oversight: System users may need access to system logic information to ensure competent use. For example, knowledge of the input variables that a system relies on can be necessary to ensure that factors already accounted for in system outputs are not accounted for more than once (and therefore distort results) within a given decision process as a whole. Similarly, internal stakeholders in charge of oversight arrangements may need an understanding of system logic to determine what kind of oversight is required and to anticipate situations that call for intervention.

Providing explanations: System logic information can be at the core of explanations sought by decision recipients. For instance, it can provide assurance that decisions are taken in non-arbitrary and methodologically sound ways. In contexts such as credit or insurance underwriting, for example, access to system logic information can also be important in order for decision recipients to understand the effect that their behaviour may have on the decisions they receive.

Responsiveness: Customer service representatives, for example, may need to understand which input variables a system relies on, how the system transforms inputs into outputs, or under what conditions a system would yield certain results to be able to respond to customer queries.

Social and economic impact: System logic information can be essential to assessing potential social and economic impacts or providing assurance in relation to concerns about such impacts. For example, knowledge of the input variables used and the relationship between inputs and outputs can be relevant to understanding whether the system relies on inferences whose use may be considered ethically objectionable. Regulators, academics, or indeed wider civil society stakeholders may have an interest in system logic information in order to assess social and economic implications.

5.2.3 Methodology (the 'how')

We now discuss how to obtain system logic information and how to communicate it to relevant stakeholders.

5.2.3.1 Obtaining system logic information

There are two methodological paths to obtaining information about an AI system's input-output relationships and conditions under which it produces certain outputs:

- **Direct interpretation:** Where complexity allows, relevant information can be obtained by analysing a formal representation of the system (as illustrated by the example of the simple linear model discussed in Section 5.2.1).
- **Indirect analysis using explainability methods:** Various auxiliary methods can help shed light on system logic. Many of these methods are perturbation-based – relying on the analysis of changes in system outputs in response to changes in input values – and can be used without access to a formal representation of the system.

Explainability methods can be used to analyse models with low and high levels of complexity alike. In cases where direct interpretation is possible, both paths will be available: system logic information may be obtained through direct interpretation, the use explainability methods, or both.

5 AI transparency

In cases where direct interpretation is not a feasible option, explainability methods are the only path. Cases of the latter kind include the following three possibilities:

- Model inscrutability limits the extent to which system logic information is obtainable through direct interpretation, regardless of technical expertise. As a result, explainability methods are the only way to shed light on certain aspects of system logic.
- A system is interpretable in principle but obtaining information through direct interpretation requires levels of technical expertise that the stakeholders seeking that information do not have. The use of explainability methods makes it possible to obtain useful information without having the relevant level of technical expertise.
- Those interested in system logic information lack the kind of direct access to the relevant AI system that is needed for direct interpretation. For example, the system may be controlled by a third-party provider, with its formal representation being treated as a commercial secret. Explainability methods can shed light on system logic without requiring access to a formal representation of the AI system.

Explainability methods can play a useful role in all three situations described above. But they are not a perfect substitute for information obtained through direct interpretation. More specifically, explainability methods can provide reliable information on conditions under which a system produces certain outputs – also known as **counterfactual explanations** – but will often only yield an uncertain or incomplete understanding of how a system transforms inputs into outputs.

Many explainability methods – including LIME, SHAP, and other prominent approaches – use input perturbations to develop a surrogate model that approximates the model being examined or to examine the relative importance of individual features in determining system outputs.⁵³ The resulting insights are approximative and probabilistic, lacking the certainty and completeness in understanding input-output relationships that can be obtained where direct interpretation is possible.

A technical discussion of explainability methods and their limitations is beyond the scope of this report. AI explainability is a rapidly evolving area of research; detailed introductions to the topic⁵⁴ and relevant technical discussions⁵⁵ are available elsewhere. Having said that, the limitations of explainability methods have two general implications that are worth highlighting here.

First, the suitability of explainability methods is context dependent. Methods differ in the kinds of insights they provide. As a result, the suitability of a particular method depends on the questions about a system's logic that one is trying answer. In examining a credit scoring model, for example, a given method may provide reliable information about the effect that a certain change in a loan applicant's income has on the model's output. Yet, other explainability methods may be needed to obtain aspects of system logic information that are required for assessing the model's reliability or robustness. In addition, the suitability of a particular explainability method can vary across model types and modelling domains. As a result, there is an active and growing research area dedicated to assessing how suitable various explainability methods are for specific financial services use cases such as credit scoring or credit risk management.⁵⁶

⁵³ Detailed descriptions of approaches like LIME and SHAP can be found in Leslie 2019, 52–54.

⁵⁴ For example, see ICO and The Alan Turing Institute 2020; Parliamentary Office of Science and Technology 2020; The Royal Society 2019; Autorité de Contrôle Prudentiel et de Résolution 2020; World Economic Forum 2019.

⁵⁵ For example, see Du, Liu, and Hu 2019; Gilpin et al. 2019; Doshi-Velez and Kim 2017; Bhatt et al. 2020; Mittelstadt, Russell, and Wachter 2019; Asher, Paul, and Russell 2020; Wachter, Mittelstadt, and Russell 2018; Sokol and Flach 2020a; Sokol and Flach 2020b.

⁵⁶ The FCA and The Alan Turing Institute are currently collaborating to examine the implications of different explainability methods in the context of mortgage default prediction. For relevant published work, see Bracke et al. 2019; Bussmann et al. 2020.

5 AI transparency

Second, the existence of explainability methods does not necessarily reduce the need to ensure that systems can be interpreted directly. For AI systems that are interpretable, this can mean ensuring access to a formal representation of systems in cases where such access does not exist (eg due to outsourcing arrangements or off-the-shelf tools whose source code is protected as a trade secret) or opting against the use of systems for which such access cannot be obtained.

For AI systems that are uninterpretable, be it due to inscrutability or limited technical expertise, it can mean choosing not to rely on such systems. The information that can be obtained through explainability methods may be insufficient to address concerns related to a given AI system, including – but not limited to – concerns about the system’s performance and compliance. Governance arrangements are therefore pivotal for guiding decisions about whether or not to use uninterpretable systems. The inability to ensure that the concerns related to an AI system have been identified and addressed can speak against the use of models whose complexity results in uninterpretability.

The decision to limit model complexity for the sake of interpretability is often portrayed as a trade-off with model accuracy. The basis for this argument is the assumption that more complex models have higher accuracy than simpler ones. Yet, this assumption is not always true. In many modelling contexts, interpretable models can be designed to achieve the same or comparable levels of accuracy as models that would be considered uninterpretable.⁵⁷ Significant research efforts are underway to advance the field of interpretable machine learning. Over time, these research efforts can be expected to further reduce the range of contexts in which interpretability-accuracy trade-offs are perceived to exist.⁵⁸

In summary, decisions in favour of interpretability do not necessarily come at the expense of accuracy. Where trade-offs between interpretability and accuracy do exist, it may be preferable to accept a lower level of accuracy in the interest of enabling direct interpretation by system developers and other relevant actors. Conversely, where uninterpretable models are being used, it is important to be mindful of the limitations of explainability methods. Ignoring these limitations risks having a false sense of understanding, potentially resulting in misplaced trust in AI systems and unexpected harmful outcomes. Governance arrangements play a key role when it comes to choosing appropriate types of models.

5.2.3.2 Communicating system logic information

System logic information is only useful if it is communicated to stakeholders in ways that are intelligible and meaningful.

Stakeholders differ in their familiarity with technical concepts. Depending on the audience, system logic information may need to be translated from technical into plain language to make it intelligible. The form and degree of translation required can vary between audiences. For example, while customers may seek information that is presented in non-technical language, senior managers may be more comfortable with technical terms. Non-textual forms of presenting system logic information, including visuals or interactive dashboards, can also enhance intelligibility.⁵⁹

Whether information is meaningful depends on the questions that stakeholders seek to answer. Questions can differ significantly between stakeholders, as can the level of detail expected in the answer to each question.

⁵⁷ Rudin 2019; Chen et al. 2018.

⁵⁸ For examples of recent advances in the area of interpretable machine learning, see Chen and Rudin 2018; Hu, Rudin, and Seltzer 2020; Sokolovska, Chevalere, and Zucker 2018; Li et al. 2018; Chen et al. 2019.

⁵⁹ For a more detailed discussion of relevant considerations and options for presenting system logic information into easily digestible ways, see the section on ‘Task 4’ in ICO and The Alan Turing Institute 2020.

5 AI transparency

This can be particularly relevant when comparing the transparency interests of customers with those involved in managing or monitoring the performance of AI systems. Three considerations are worth highlighting:⁶⁰

- **The role of counterfactuals:** The interest of customers in accessing system logic information can often be driven by questions about the conditions under which a system would yield a certain output (eg a favourable decision outcome). Such counterfactual explanations differ from the types of information that are of interest to other stakeholders, eg those who want to understand system performance.⁶¹
- **Relevance:** Excessively detailed information or information that is irrelevant to customers' queries can cause confusion and generate distrust.
- **Intuitiveness and simplicity:** Customers may expect the logic of systems to be sufficiently intuitive and simple, so that they are able to remember it in day-to-day life and make informed choices about aspects of their behaviour that may affect decision outcomes. Intelligibility alone does not guarantee that these expectations are met.

5.3 Process transparency

In this section, we consider (i) what information falls under the category of process transparency, (ii) why stakeholders can be interested in such information, and (iii) how such information can be managed and communicated. In discussing the 'what', we introduce a conceptual framework that maps different kinds of process information along two dimensions: an AI system's **lifecycle phases** and different **levels of information** that can be of interest to stakeholders. The second part, discussing the 'why', sets out how process information, much like system logic information, can be relevant to the six areas of concern identified in Chapter 3.⁶² The third part addresses the 'how' by highlighting current areas of research and debate about ways to record and present process information, relevant norms and standards, and mechanisms for verifying process information. Figure 10 illustrates this section's structure.

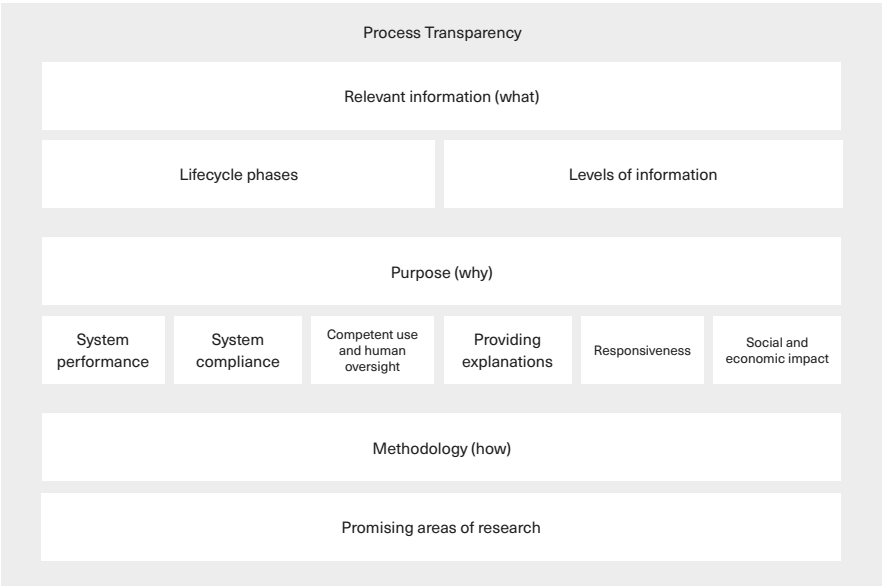


Figure 10: Process transparency

60 For more detailed discussions, see ICO and The Alan Turing Institute 2020.

61 As an existing illustration of counterfactual explanations in financial services, there are a number of commercial examples of 'credit score simulators' that allow customers to test how changes in the values of different input variables would affect their credit score.

62 As a reminder, these are (1) system performance, (2) system compliance, (3) competent use and human oversight, (4) providing explanations, (5) responsiveness, and (6) social and economic impact.

5 AI transparency

5.3.1 Relevant information (the ‘what’)

Process transparency concerns access to any information about an AI system’s design, development, and deployment apart from the system’s logic. As with system logic information, such process information is important for addressing and providing assurance about concerns raised by AI systems. Correspondingly, there is a growing amount of work on how process information can be recorded, managed, and made accessible in practice.⁶³

We can categorise process information regarding AI systems along two dimensions:

- **Different lifecycle phases:** Process information can relate to (i) the design and development or (ii) the deployment of an AI system. In both areas, more specific lifecycle phases can be distinguished, each of them associated with unique aspects of information.
- **Different levels of information:** In considering a given lifecycle phase, different levels of process information can be distinguished, corresponding to the kinds of questions that the information serves to answer.

These two dimensions lead to a typology for process information whose general structure can be represented in the form of a matrix, as illustrated in Figure 11. We will consider each of the two dimensions in turn.

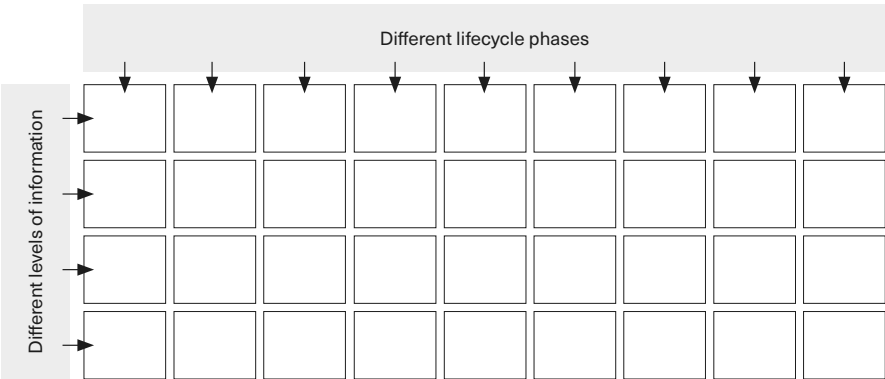


Figure 11: Process transparency matrix

5.3.1.1 Different AI lifecycle phases

System design and development and system deployment each comprise a number of analytically distinct activities that can be thought of – collectively – as the phases of an AI system’s lifecycle. There is no universally agreed breakdown of lifecycle phases for AI systems. However, the following illustrative typology is suitable for a range of contexts and intersects with prominent lifecycle frameworks.⁶⁴

63 For example, see ICO and The Alan Turing Institute 2020; Raji et al. 2020; Raji and Yang 2020; Brundage et al. 2020; Ashmore, Calinescu, and Paterson 2019; Stiftung Neue Verantwortung 2020; Cihon 2019; OCEANIS 2021.
64 Some of the activities included in this breakdown only apply to certain types of AI systems.

5 AI transparency

Lifecycle phases for AI system design and development:

- **Business case and problem definition:** Establishing the need for the AI system and the tasks it is meant to perform.
- **System requirements specification:** Translating the problem definition into technical design and performance requirements.
- **Data acquisition and preparation:** Where relevant, acquiring any data that may be needed to build the system, checking its suitability, and preparing it for use.⁶⁵
- **Building:** Creating a system that meets the design requirements previously specified. In the case of ML projects, this involves choosing between ML methods, developing and evaluating candidate models, and selecting the best performing model.
- **Validation and verification:** Verifying, on an on-going basis, that the system meets the relevant design and performance requirements. Depending on the nature of the system, assessment can rely on empirical testing or formal verification.⁶⁶

Lifecycle phases for system deployment:

- **Integration:** Preparing the AI system for operation by integrating it into the relevant business environment. This can involve technical aspects of integration with other systems or technology infrastructure. It also includes the introduction of users to the operation of the system, the delivery of user training, and other relevant aspects of organisational change management.
- **Operation:** Using the AI system to perform the business tasks for which it was intended.
- **Monitoring and evaluation:** Observing and recording system behaviour in order to assess system performance and compliance during operation, including any procedures of periodic re-validation.
- **Updating/system retirement:** Making changes to the AI system as needed, for example to improve performance or prevent performance deterioration. In the context of supervised ML models, such changes take the form of retraining the model based on new training data. Successful updating is followed by another iteration of lifecycle steps outlined above.

The lifecycle phases capture activities that are conceptually distinct, but do not necessarily occur in succession. During an AI system's design and development, for example, agile processes can involve iterative cycles and adjustments across the different phases outlined above. When it comes to deployment, operation and monitoring/evaluation typically occur in parallel. Moreover, in the case of adaptive systems, updating can occur continually during operation.

The volumes of data needed for AI systems and the complexity of technology supply chains mean that different activities across lifecycle phases are not always performed by actors within the same organisation. In contexts that involve third-party data providers, outsourcing different aspects of system design and development, or reliance on off-the-shelf tools, certain activities will be carried out by actors outside of the firm using the system.

⁶⁵ This may include relevant forms of data pre-processing and data augmentation. As regards the role of data in AI development, see the minutes of the second meeting of the AI Public Private Forum, jointly hosted by the FCA and Bank of England, dated 26 February 2021.

⁶⁶ For overviews of different types of verification, see Ashmore, Calinescu, and Paterson 2019; Brundage et al. 2020.

5 AI transparency

Indeed, some of these activities might not be carried out by human actors, but by AI systems: recent innovations make it possible to automate large sections of an AI system's development. However, in any of these cases, the structure of lifecycle phases remains unaffected by this, as the fundamental steps in designing, developing, and deploying an AI system stay the same.

5.3.1.2 Different levels of information

For each lifecycle phase of an AI system, there are various aspects of information that can be of interest to internal or external stakeholders. These aspects of information can answer questions at different levels of abstraction. The following four levels of information can be distinguished, moving from more concrete to more abstract questions:

- **Substantive information** relates to questions about substantive aspects of activities within a given lifecycle phase for an AI system. Examples of such information include: the content of problem definition or system requirement statements; the content of or summary statistics for datasets used during the AI system's development and operation; source code or other formal representations of the AI system; and the results of tests conducted to assess system performance or compliance.
- **Procedural information** answers questions about the procedures that were followed in performing the activities within a given lifecycle phase. Examples include descriptions of: the process that led to the agreed problem definition or system requirements (eg the actors and the steps involved); the procedures employed to collect and assemble the data used during the AI system's development or operation; the nature of data quality checks or processing steps carried out; the process followed to select the type of ML method used for developing models; or the procedures used to conduct system tests.
- **Governance information** answers questions about governance arrangements for activities conducted within a lifecycle phase. This information may take the form of statements of accountability and liability, or descriptions of the structure of relevant oversight mechanisms (including, where relevant, the role of risk and compliance teams, ethics review boards, audit teams, senior managers, or board members).
- **Information on adherence to norms and standards** refers to compliance with norms or standards in the design, development, and deployment of an AI system. Such norms or standards may touch on substantive, procedural, or governance questions.

The distinction between these four levels of information, like the lifecycle typology, remains unaffected by the complexities of sourcing data for AI systems or technology supply chains. Where firms rely on third-party providers, all four levels of information can be applied to activities carried out within and outside the firm. Additionally, relevant governance information in such cases can include information about accountability structures and mechanisms that govern the relationship between third-party providers and the firm employing the AI system in question.

These four levels of information, combined with the typology of lifecycle phases, lead to a more concrete version of the matrix we introduced at the beginning of this section to map out different types of process information.

5 AI transparency

Figure 12 incorporates the categories introduced in the last few pages.

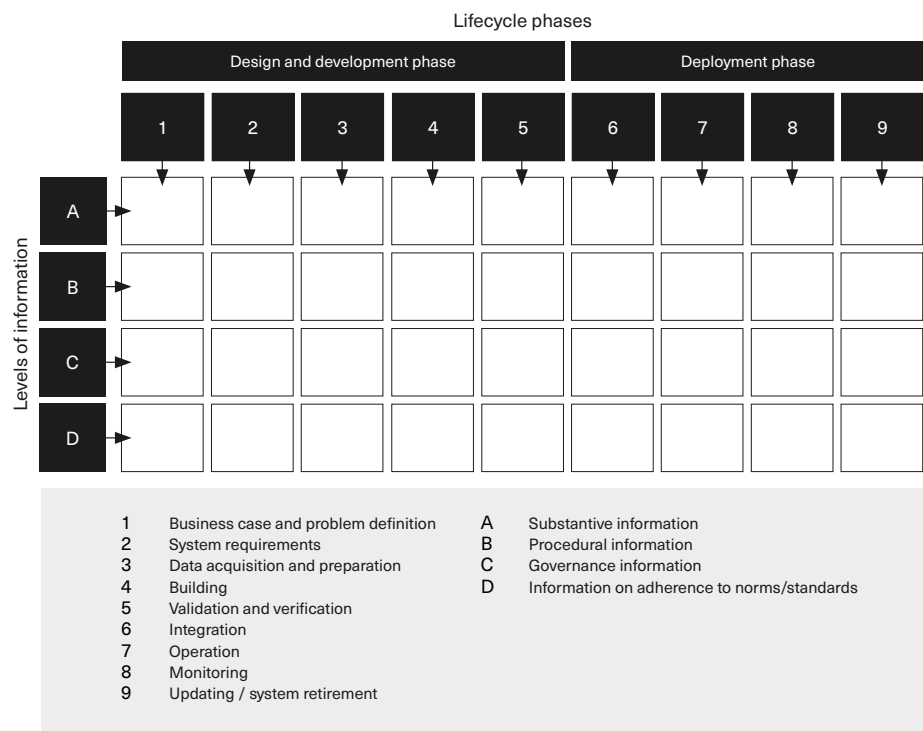


Figure 12: Detailed process transparency matrix

5.3.2 Purpose (the ‘why’)

Process information, like system logic information, can help to address concerns related to AI systems (ensuring the trustworthiness and responsible use of these systems) as well as to provide assurance that concerns have been addressed adequately (demonstrating trustworthiness and responsible use). In the following paragraphs, we illustrate the importance of process transparency in addressing each of the six areas of concern we identified in Chapter 3.

System performance: Information about the content of system requirement specifications, about the quality and origin of data used during an AI system’s development or operation, or about validation procedures is crucial for understanding the effectiveness, reliability, and robustness of AI systems. This information can be of interest to those involved in or making decisions about the development and use of an AI system, as well as those seeking assurance about the system’s performance (eg members of audit teams, board members, regulators or customers).

System compliance: Process information is crucial to assessing AI systems’ adherence to compliance requirements. For example, information about the quality of data used and system tests conducted is essential for a holistic understanding of potential risks of unlawful discrimination. Similarly, where AI systems use personal data, information about the provenance, content, and quality of this data is important for data protection assessments. Process information can be of interest to those ensuring system compliance or can demonstrate system compliance to stakeholders.

5 AI transparency

Competent use and human oversight: Information about an AI system's intended purpose, system requirements specifications, or system performance measurements can be essential to ensuring competent use and preventing the inappropriate repurposing of AI systems. This information can also be crucial to determine what forms of human oversight are needed and to enable overseers to exercise their role effectively.

Providing explanations: Explanations of an AI system's outputs can involve system logic information as well as process information. Indeed, a complete understanding of a particular decision requires both. Figure 13 illustrates this using the example of a loan eligibility decision. In terms of process information, decision recipients seeking to understand an outcome may want to know the content of the input data about them that an AI system used. This knowledge is a precondition, for example, for being able to identify erroneous decisions.

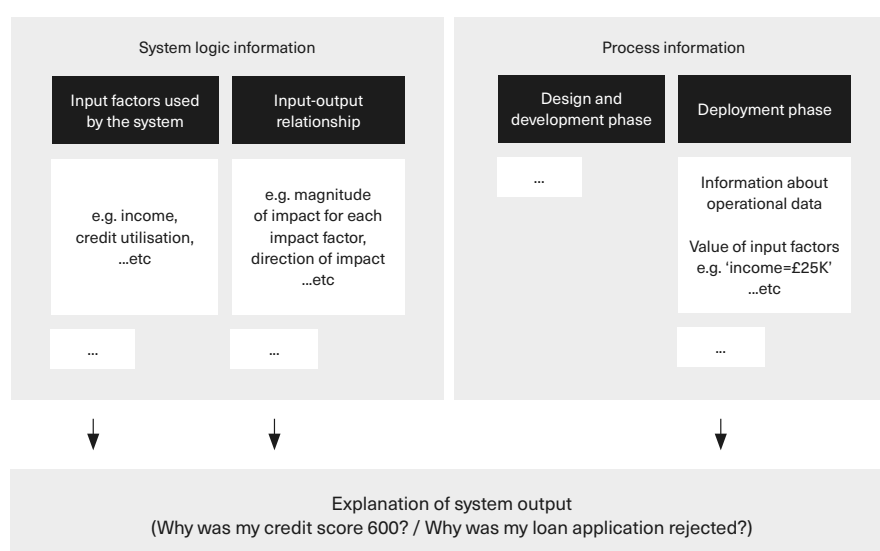


Figure 13: The combined relevance of process and system logic information in explaining system outputs

Responsiveness: Telling customers about ways in which they can ask for information, help, or redress is important to reassure them of the existence of pathways for expressing such requests. In addition, internal stakeholders may need access to different forms of process information, such as the data used during an AI system's operation, to be able to respond to customer requests. Finally, stakeholders seeking assurance about the responsible use of AI systems may be interested in information about how issues of responsiveness are managed.

Social and economic impact: Various types of process information may be needed to manage and provide assurance regarding the social and economic impacts of an AI system. For example, information about system test results can be important for understanding an AI system's potential financial exclusion implications. Similarly, information about how firms communicate personal data use to customers can be of interest to stakeholders seeking assurance in relation to concerns regarding consumer empowerment.

5 AI transparency

5.3.3 Methodology (the 'how')

The appropriate level of detail and technical sophistication in providing process information will depend on the purpose that the information is meant to serve. For example, actors involved directly in the validation of an AI system are likely to need a system requirements statement. Customers or other stakeholders interested in ensuring that the right procedures have been followed in validating an AI system will likely need less detailed information, expressed in easy-to-understand language.

Existing best practices within firms – even if they are not specifically designed for AI systems – can guide the process of identifying suitable ways of recording and presenting process information. For example:

- In financial services, there are established technology management and governance frameworks whose applicability extends to certain AI use cases (eg frameworks for model risk management or algorithmic trading).
- There are prominent best practices and frameworks in the areas of data management and data protection that intersect with aspects of managing process information in the context of AI systems and can provide valuable guidance (eg Data Protection Impact Assessments⁶⁷).

In addition, there are several areas of AI-specific debate and research related to managing and communicating process information. We highlight three of these areas below.

Recording and presenting process information for AI systems: Recent years have seen a rapidly growing literature on topics such as documentation, assurance, traceability, and audit trails for AI systems. Contributions to this literature often give examples of how different aspects of process information can be recorded and made accessible to different stakeholders. In many cases, these examples involve proposals for different 'documentation artefacts' and templates that can be used to structure process information in practice.

Some contributions to this debate are focused on subsets of process information or the information needs of particular stakeholders.⁶⁸ Increasingly, however, contributions adopt a holistic perspective on documentation needs, covering all phases of an AI system's lifecycle as well as the information needs of all relevant stakeholders.⁶⁹ An approach that is growing in popularity – especially in the context of high-stakes applications of AI – is the use of 'argument-based assurance cases', often following a specified template, in support of claims about an AI system's properties.⁷⁰

Recent years have also seen an increase in the number of open-source tools for testing AI systems and examining their properties. These tools can be useful for generating some of the process information that is of interest to stakeholders.⁷¹

Emerging norms and standards: A second evolving area with relevance to managing and communicating process information consists of work on standards for AI systems and on professional standards.

⁶⁷ ICO 2016.

⁶⁸ Gebru et al. 2018; Holland et al. 2018; Kelley et al. 2009; Mitchell et al. 2019; Bender and Friedman 2018; Arnold et al. 2019; Huynh, Stalla-Bourdillon, and Moreau 2019; VDE and Bertelsmann Stiftung 2020.

⁶⁹ ICO and The Alan Turing Institute 2020; Leslie 2019; Raji and Yang 2020; Raji et al. 2020; Ashmore, Calinescu, and Paterson 2019.

⁷⁰ Prominent frameworks for argument-based assurance cases include Claims, Arguments, Evidence (CAE) and Goal Structuring Notation (GSN). For introductory discussions, see ICO and The Alan Turing Institute 2020; Brundage et al. 2020.

⁷¹ For discussions of some such tools, see Lee and Singh 2021; Centre for Data Ethics and Innovation 2020.

5 AI transparency

Several national and international bodies are currently working to develop standards for AI systems.⁷² While the applicability of these standards to AI use cases in financial services will depend on context, they can be a useful point of reference.

In addition, recent years have seen growing support of initiatives to professionalise the field of data science. Efforts in this space are aimed at establishing commonly agreed curricula for data science courses and possible forms of professional accreditation for data scientists. For example, a group of professional bodies led by the Royal Statistical Society (RSS) is currently working to develop commonly agreed professional standards for data science.⁷³ Concurrently, some professional bodies in the financial services space are turning their attention to codes of conduct for the use of data and emerging technologies.⁷⁴

Mechanisms for verifying process information: A third area of emerging work concerns the verifiability of process information for AI systems. This includes forms of independent certification for relevant norms and standards. Currently, declarations of adherence to norms and standards take the form of self-declared adherence ('self-certification'). However, in some contexts, stakeholders may place greater trust in such declarations if they are supported by independently administered certification or labelling schemes.

There is also an emerging literature on the role of auditors in examining system design, development, and deployment processes (including evaluation).⁷⁵ In contrast to certification, auditors may verify process information at a more detailed level. AI system auditors can be internal or external to the firm that is employing a given AI system.

Finally, growing research and development efforts are being dedicated to technical solutions that automate the generation and recording of process information. Software-generated 'audit trails' and related concepts can contribute to the reliability and verifiability of some types of process information, while at the same time reducing the cost of recording and making the information available to stakeholders.⁷⁶

5.4 Trade-offs

In concluding the discussion of transparency in this chapter, we note that there can be reasons for not making some types of information about AI systems accessible to certain stakeholders. Such reasons often play a prominent role in discussions about the disclosure of information to external stakeholders in particular. The applicability of such countervailing reasons is context dependent. In particular, these reasons, where relevant, do not speak against the disclosure of system logic and process information in a wholesale manner. Instead, they typically apply to the disclosure of specific types of information (eg specific aspects of system logic information rather than all types of system logic information) to specific types of stakeholders (eg customers rather than all external stakeholders), for specific types of use cases.

As highlighted in Section 5.2.3, disclosing information that is irrelevant or excessively detailed in response to stakeholders' questions may generate undue distrust. Avoiding 'information overload' is one possible reason against the disclosure of some types of information to certain stakeholders.

⁷² Such standards can take the form of process or product standards. For an overview of relevant initiatives, see OCEANIS 2021; Cihon 2019; Stiftung Neue Verantwortung 2020; National Institute of Standards and Technology 2019.

⁷³ Royal Statistical Society 2020; The Royal Society 2020.

⁷⁴ For example, see Chartered Insurance Institute 2019a; Chartered Insurance Institute 2019b.

⁷⁵ See, for example, Koshiyama et al. 2021.

⁷⁶ Brundage et al. 2020.

5 AI transparency

Three other potential reasons are worth noting:

- **Preventing system manipulation or ‘gaming’:** In some cases, firms employing AI systems may seek to protect certain aspects of information to prevent the subversion of these systems. In the case of fraud detection systems, for instance, preventing adversarial actors from finding ways to evade detection can speak against disclosing information about system logic or the data used to customers. Yet, this countervailing reason does not necessarily apply to the disclosure of the same information to regulators, or the disclosure of other types of information to customers.
- **Protecting commercially sensitive information:** Certain types of information may be considered commercially sensitive by the firm employing an AI system or by third-party providers involved in the system’s development. For example, an investment management firm that relies on proprietary AI systems to identify profitable investment opportunities has an interest to protect the competitive advantage enabled by these systems. Similarly, third-party providers may want to protect the IP contained in their products. As such, firms may be reluctant to disclose information that is central to their commercial success. Once again, however, this reason typically only applies to specific types of information (eg details of a system’s logic or proprietary source code) and their disclosure to certain stakeholders.
- **Protecting personal data:** Certain forms of information disclosure can conflict with firms’ obligation to protect personal data. This includes, most obviously, the direct sharing of personal data – be it data used in the development or the operation of AI systems – in ways that violate data protection legislation. In addition, where AI systems are trained with personal data, it may be possible to infer protected personal information through, for example, model inversion or membership inference attacks. While concerns about such attacks only apply in limited circumstances, they can speak against the disclosure of certain aspects of system logic information to stakeholders.

A more detailed discussion of these trade-offs is beyond the scope of this report.⁷⁷ The applicability and implications of transparency trade-offs depend on context and vary between AI use cases. It is worth noting that, regardless of the applicability of different countervailing reasons, large segments of the information that is of interest to stakeholders will remain unaffected.

⁷⁷ For further introductory discussions, see ICO 2020; ICO and The Alan Turing Institute 2020.

6 Conclusion

This report provided an introduction to the use of AI in the context of financial services, examined its potential ethical and regulatory implications, and set out the role of AI transparency in addressing these implications.

We conclude by briefly summarising the report's key takeaways and highlighting areas in need of further work.

6 Conclusion

As Chapter 2 underlined, the field of AI has a decades-long history and is characterised by methodological connections to statistics and other disciplines with long-standing applications in financial services. Recent technological advancements discussed under the heading of AI build on these connections and often take incremental forms rather than representing a seismic shift in paradigms. Underpinning these advancements are the analytically distinct innovation elements of ML, non-traditional data, and automation.

These elements – used on their own or in combination with each other – can pose challenges from the perspective of responsible innovation. In particular, data quality issues, novel characteristics of models, changes in the structure of technology supply chains, and increases in the scale of impacts associated with AI systems give rise to concerns in the six areas identified in Chapter 3. AI ethics principles recognise these challenges and seek to steer the responsible design, development, and deployment of AI systems.

Across the diverse landscape of use cases in financial services, AI has the potential to enable significant benefits as well as to lead to serious harms. Chapter 4 highlighted some of these benefits and harms in five distinct areas: consumer protection, financial crime, competition, the stability of firms and markets, and cybersecurity. Realising benefits and preventing harms depends on ensuring and demonstrating that AI systems are trustworthy and used responsibly.

Transparency plays a foundational role for the responsible adoption of AI. This includes the availability of information about AI systems to internal stakeholders involved in decisions about their design, development, and deployment as well as to external stakeholders such as customers and regulators. Across different stakeholders, transparency needs can comprise information about AI systems' logic and information about the processes surrounding the systems' design, development, and deployment. In its different forms, transparency is crucial to ensuring and demonstrating AI systems' trustworthiness and responsible use.

Against this background, there are important ensuing questions where future research could make significant contributions. Two areas, in particular, are worth highlighting.

6 Conclusion

First, Chapter 5 provides a solid conceptual foundation for defining expectations and making decision about AI transparency. It does not, however, act as a guide for how to implement AI transparency in practice. An important area for further work consists of applying the conceptual framework presented here to specific use cases of AI in financial services and determining the concrete forms that AI transparency should take.

Second, this report provides a comprehensive mapping of the potential challenges and concerns related to the use of AI in financial services. It does not, however, assess the extent to which existing regulatory arrangements or industry practices are adequate for addressing these challenges and concerns. Further work is needed to answer questions about the possible need for changes to regulatory requirements or modifications to the risk and control frameworks used by firms. This important work will also need to consider the question of whether adjustments that may be needed should take the form of AI-specific rules and frameworks or the form of changes to provisions and arrangements that are broader in scope, such as general model risk management practices.

AI is already having transformative impacts on the delivery of financial services. Its role is set to increase further in the years to come. Like in other sectors, firms in financial services, regulators, consumers, and society at large are confronted with an evolving landscape of promising technological innovations and newly emerging challenges and risks. This report's contribution is to equip stakeholders with the understanding needed to navigate this landscape in pursuit of responsible and socially beneficial innovation.

Appendix: ML approaches

This appendix describes the primary ML approaches mentioned in Chapter 2: supervised learning, unsupervised learning, and reinforcement learning.

Appendix: ML approaches

A.1 Supervised learning

In the case of supervised learning, the task of the ML algorithm is to infer the value of a predefined target (or output) variable based on known values of feature (or input) variables.⁷⁸ The existence of labelled data (ie data with known values for the target in question)⁷⁹ is a prerequisite for supervised learning. The learning process consists of developing a model of the relationship between feature variables and target variables based on labelled training data. This process is also known as 'model training'. Following a successful training phase (confirmed through a testing phase that also relies on labelled data), the resulting model can be applied to unlabelled data to infer the most likely value of the target variable. This is known as the inference phase.

Figure 14 summarises these processes.

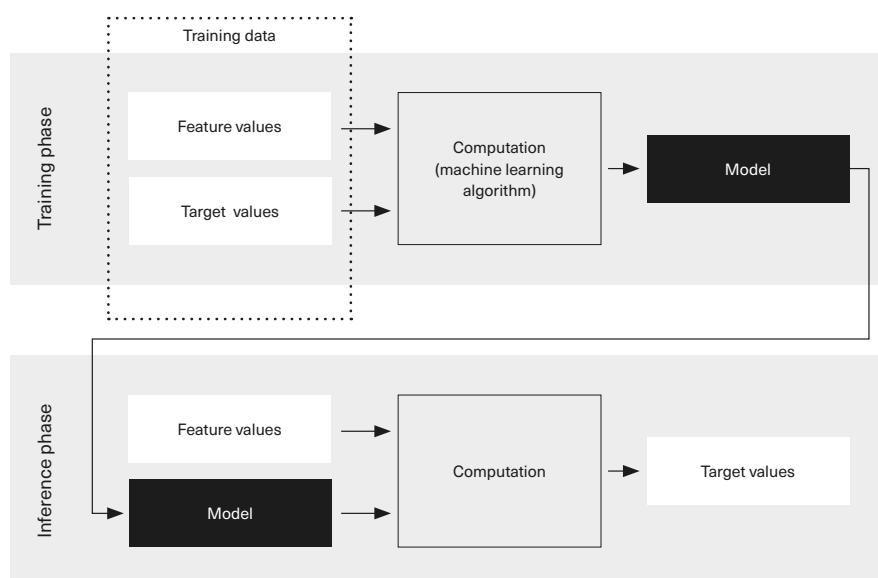


Figure 14: Simplified supervised learning process

Supervised learning can solve two main types of analytical problems:

- **Regression problems**, where the target of interest is a continuous variable. Examples include predicting future stock prices or insurance costs.
- **Classification problems**, where the target of interest is a categorical variable. This includes problems in which the target value is binary (eg a financial transaction being fraudulent or non-fraudulent) as well as multi-class problems that involve more than two categories. For example, classification can serve to assess the probability of customers defaulting on loan repayments.

⁷⁸ The distinction between target and features corresponds to the distinction between an independent variable and dependent variables in traditional statistics terminology.

⁷⁹ In other words, labelled training data consists of a set of known input-output pairs.

Appendix: ML approaches

A.2 Unsupervised learning

Unsupervised learning involves the identification of patterns and relationships in data without there being a pre-defined relationship of interest. In contrast to supervised learning, this approach does not rely on labelled training data. As a result, unsupervised learning can be more exploratory, although the results are not necessarily any less actionable.

Unsupervised learning is particularly useful in contexts where labelled data does not exist or is expensive to produce. This approach can solve problems such as:

- **Cluster analysis**, where the aim is to group units of observation based on similarities and dissimilarities between them. Examples of tasks where cluster analysis can help include customer segmentation exercises.
- **Association analysis**, where the goal is to identify salient relationships between variables within a dataset. Association rules (ie formal if-then statements) typically describe such relationships. These rules can translate into insights such as 'customers that are interested in X also tend to be interested in Y and Z'. Association analysis is used for tasks such as product recommendations and customer service management.

A.3 Reinforcement learning

Reinforcement learning relies on the concept of an 'agent' exploring an environment. The task of the agent is to identify an optimal action or sequence of actions (the target of interest) in response to its environment. The learning process does not rely on examples of 'correct answers'. Instead, it relies on a reward function that provides feedback on the actions taken. The agent aims to maximise its reward and thus improve its performance through an iterative process of trial and error.

Reinforcement learning is useful in settings where optimal actions (ie correct answers) are unknown. In these settings, labelled training data is unobtainable or risks leading to suboptimal results if analysts use supervised learning. The conceptual structure of the approach also makes it relevant to problem types that have a sequential or dynamic nature. Examples include problems in robotics or game playing.

A lot of work on reinforcement learning occurs in the context of fundamental research. This includes research dedicated to general AI. Reinforcement learning is less prevalent in business contexts compared to other ML approaches. Business applications that attract the most attention lie outside financial services and include autonomous vehicle and other forms of robotic engineering. Within financial services, possible applications include trading or trade execution, and dynamic pricing.

Glossary

Adaptivity. A property of dynamic systems which relates to their ability to update themselves based on new data.

AI ethics principles. A set of principles that seek to steer the responsible design, development, and deployment of AI systems.

AI lifecycle phases. The different stages in an AI system's design, development, and deployment that can be thought of – collectively – as the phases of an AI system's lifecycle.

Algorithm. A pre-defined series of steps that need to be followed in order to solve a computational problem.

Artificial intelligence (AI). The 'science of making computers do things that require intelligence when done by humans.' This is a commonly used definition of AI proposed by Marvin Minsky.

Automated machine learning (AutoML). The automation of processes related to developing or maintaining ML models.

Automation. A situation where the use of technology reduces or removes the role of humans in performing tasks and processes.

Computer vision. A subfield of AI research focused on the processing of visual data by machines.

Counterfactual explanations. Information about the conditions under which an AI system would yield a certain output.

Data poisoning attacks. A type of attack whereby dynamic models are exposed to data that is intended to 'retrain' them with the aim of reducing their effectiveness.

Dimensionality. The number of input variables that a model relies on.

Dynamic system. An AI system that relies on an ML model that, once deployed, continues to adapt in response to new data that becomes available during operation. Dynamic system contrasts with **static system**.

Explainability methods. A collection of strategies and tools that can shed light on a system's logic indirectly rather than obtaining such information through an analysis of a formal representation of the system's logic.

External transparency. A type of transparency where information about an AI system is accessible to external stakeholders (ie actors external to the firm employing the AI system that have a significant relationship with the firm deploying the system or may be affected by the AI system's use).

Feature variable. A variable upon which a model relies to generate outputs, also known as an input variable. Feature variable contrasts with **target variable**.

General AI. The idea of AI systems that have universal abilities on par with those of the human mind. These abilities include the versatility to learn and perform any intellectual task that humans are capable of. General AI remains an ambition rather than a reality. General AI contrasts with **narrow AI**.

Glossary

Human-in-the-loop. A type of arrangement in which humans actively confirm the execution of actions or decisions recommended by an AI system. See also **human-on-the-loop**.

Human-on-the-loop. A type of arrangement in which humans play a supervisory role and can override the execution of actions or decisions recommended by an AI system. See also **human-in-the-loop**.

Input variable. A variable upon which a model relies to generate outputs, also known as a feature variable. Input variable contrasts with **output variable**.

Inscrutability. A property of some highly complex models which makes it impossible even for experts with high levels of specialised technical knowledge to obtain a complete understanding of the relationships between model inputs and outputs.

Internal transparency. A type of transparency where information about an AI system is accessible to internal stakeholders (ie individuals within the firm that is employing the AI system).

Labelled data. In the context of ML problems that involve a specific target (or output) variable of interest, this term refers to datasets that contain known values for the target (or output) variable in question. Labelled data contrasts with **unlabelled data**.

Machine learning (ML). The development of AI systems that are able to perform tasks as a result of a 'learning' process that relies on data. ML is at the core of recent advances in the field of statistical AI and contrasts with approaches and methods that rely on embedding explicit rules and logical statements into code. Prominent ML approaches include **supervised learning**, **unsupervised learning**, and **reinforcement learning**.

Model opacity. A model's property of making it difficult to understand 'how it works' and, in particular, the relationship between model inputs and model outputs.

Narrow AI. AI systems whose abilities are limited to certain pre-defined tasks for which they were developed. Narrow AI contrasts with **general AI**.

Natural language generation. An area of NLP focused on producing written or spoken human language.

Natural language processing (NLP). A subfield of AI research focused on the processing of written or spoken human language by machines. It includes the areas of **speech recognition**, **natural language understanding**, and **natural language generation**.

Natural language understanding. An area of NLP focused on recognising meaning in human language.

Non-traditional data. Data that firms have not used historically to perform a given task. In some instances, the data in question did not exist or was not previously accessible. In other cases, it was available but went unused due to a lack of technical capabilities. Non-traditional data contrasts with **traditional data**.

Optical character recognition. The recognition of characters from images of handwritten or printed text.

Glossary

Outlier. A data point which differs significantly from the majority of observations within a given dataset.

Output variable. The variable of interest in an ML context, also known as a model's target variable. Output variable contrasts with **input variable**.

Parameter. A coefficient in the model reflecting the weight assigned to a given feature variable.

Process transparency. A type of transparency where stakeholders have access to information that relates to the processes surrounding an AI system's design, development, and deployment.

Reinforcement learning. An ML approach which relies on the concept of an 'agent' exploring an environment. The task of the agent is to identify an optimal action or sequence of actions (the target of interest) in response to its environment.

Representativeness. An aspect of data quality which means that the composition of a dataset used in a modelling task provides an adequate representation of the real world for the intended purpose.

Speech recognition. An area of NLP focused on the processing of spoken human language by machines.

Static system. An AI system that, once deployed, does not evolve further unless it is deliberately updated, for example by replacing the model it relies on. Static system contrasts with **dynamic system**.

Statistical AI. A subfield of AI that relies on bottom-up, data-driven systems. The capabilities of such systems are not the result of the rule-based application of encoded human knowledge but instead arise from the analysis of data. Statistical AI contrasts with **symbolic AI**.

Structured data. Data that is organised, formatted, and stored in machine-readable formats. Structured data contrasts with **unstructured data**.

SUM values. A set of ethical values that 'support, underwrite, and motivate' responsible and reflective AI innovation practices.

Supervised learning. The use of ML methods to develop models that serve to infer the value of a predefined target (or output) variable based on known values of feature (or input) variables. During the development process, labelled data is used to 'train' the model by analysing the relationships between feature and target values in this data.

Symbolic AI. A subfield of AI that relies on translating human knowledge and logical statements into explicitly programmed rules. Symbolic AI contrasts with **statistical AI**.

System logic. The operational logic of an AI system or, in colloquial terms, the system's 'inner workings.'

System transparency. A type of transparency where stakeholders have access to information that relates to the operational logic of a given AI system or, in colloquial terms, information about the system's 'inner workings.'

Glossary

Target variable. The variable of interest in an ML context, also known as a model's output variable. Target variable contrasts with **feature variable**.

Traditional data. Data that firms have been using historically to perform a given task. Traditional data contrasts with **non-traditional data**.

Transparency. An AI ethics principle which relates to stakeholders having access to relevant information about a given AI system. See also **process transparency**, **system transparency**, **internal transparency**, and **external transparency**.

Unlabelled data. In the context of ML problems that involve a specific target (or output) variable of interest, this term refers to datasets that include values for the feature (or input) variables, but that do not include values for the target (or output) variable. Unlabelled data contrasts with **labelled data**.

Unstructured data: Data that is not organised and stored in a structured format. Unstructured data contrasts with **structured data**.

Unsupervised learning. The use of ML methods to identify patterns and relationships in data without there being a pre-defined relationship of interest and without relying on labelled training data.

Voice sentiment analysis. The analysis of human voice data with the aim of recognising sentiments and emotions.

Bibliography

- Arnold, Matthew, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, et al. 2019. [“FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity.”](#) IBM Journal of Research and Development 63, no. 4/5.
- Asher, Nicholas, Soumya Paul, and Chris Russell. 2020. [“Adequate and Fair Explanations.”](#)
- Ashmore, Rob, Radu Calinescu, and Colin Paterson. 2019. [“Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges.”](#)
- Association for Financial Markets in Europe (AFME). 2018. [“Considerations on the Ethical Use of Artificial Intelligence in Capital Markets.”](#)
- Autorité de Contrôle Prudentiel et de Résolution. 2020. [“Governance of Artificial Intelligence in Finance.”](#)
- Autorité de la Concurrence and Bundeskartellamt. 2019. [“Algorithms and Competition.”](#)
- Bender, Emily M., and Batya Friedman. 2018. [“Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.”](#) Transactions of the Association for Computational Linguistics 6: 587–604.
- Bhatt, Umang, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. [“Explainable Machine Learning in Deployment.”](#) Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.
- Bracke, Philippe, Anupam Datta, Carsten Jung, and Shayak Sen. 2019. [“Machine Learning Explainability in Finance: An Application to Default Risk Analysis.”](#) Staff Working Paper No. 816 Bank of England.
- Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, and Peter Eckersley. 2018. [“The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.”](#)
- Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 2020. [“Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.”](#)
- Buchanan, Bonnie G. 2019. [“Artificial Intelligence in Finance.”](#) The Alan Turing Institute.
- Burrell, Jenna. 2016. [“How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.”](#) Big Data & Society 3, no. 1: 1–12.
- Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. [“Explainable AI in Credit Risk Management.”](#) Computational Economics 57: 203–16.
- Cambridge Centre for Alternative Finance and World Economic Forum. 2020. [“Transforming Paradigms: A Global AI in Financial Services Survey.”](#)
- Centre for Data Ethics and Innovation. 2020. [“Bias Mitigation.”](#)
- Chartered Insurance Institute. 2019a. [“Digital Ethics Companion: A Practical Guide.”](#)
- . 2019b. [“Digital Ethics: A Companion to the Code of Ethics.”](#)
- Chen, Chaofan, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. [“This Looks Like That: Deep Learning for Interpretable Image Recognition.”](#) Advances in Neural Information Processing Systems. Vol. 32.

Bibliography

- Chen, Chaofan, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. [“An Interpretable Model with Globally Consistent Explanations for Credit Risk.”](#)
- Chen, Chaofan, and Cynthia Rudin. 2018. [“An Optimization Approach to Learning Falling Rule Lists.”](#) Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR.
- Cihon, Peter. 2019. [“Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development.”](#) Center for the Governance of AI.
- CMA. 2018. [“Pricing Algorithms.”](#)
- . 2021. [“Algorithms: How They Can Reduce Competition and Harm Consumers.”](#)
- Council of Europe. 2019. [“CAHAI - Ad Hoc Committee on Artificial Intelligence.”](#)
- De Nederlandsche Bank. 2019. [“General Principles for the Use of Artificial Intelligence in the Financial Sector.”](#)
- Doshi-Velez, Finale, and Been Kim. 2017. [“Towards a Rigorous Science of Interpretable Machine Learning.”](#)
- Du, Mengnan, Ninghao Liu, and Xia Hu. 2019. [“Techniques for Interpretable Machine Learning.”](#) Communications of the ACM 63, no. 1: 68–77.
- European Banking Authority. 2020. [“EBA Report on Big Data and Advanced Analytics.”](#) EBA/REP/2020/01.
- Ezrahi, Ariel, and Maurice E. Stucke. 2016. Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy. Cambridge, MA: Harvard University Press.
- FCA. 2015. [“Consumer Vulnerability.”](#) Occasional Paper No.8.
- . 2018a. [“Price Discrimination in the Cash Savings Market.”](#) Discussion Paper DP18/6.
- . 2018b. [“Price Discrimination in Financial Services: How Should We Deal with Questions of Fairness?”](#) Research Note.
- . 2018c. [“Fair Pricing in Financial Services.”](#) Discussion Paper DP18/9.
- . 2018d. [“6 Evidential Questions to Help Assess Concerns about Fairness in Price Discrimination.”](#)
- . 2019a. [“Mortgages Market Study: Final Report.”](#) Market Study MS16/2.3.
- . 2019b. [“General Insurance Pricing Practices: Interim Report.”](#) Market Study MS18/1.2.
- . 2019c. [“Robo Advice – Will Consumers Get with the Programme?”](#) Insight Blog.
- . 2020. [“General Insurance Pricing Practices: Final Report.”](#) Market Study MS18/1.3.
- FCA and Bank of England. 2019. [“Machine Learning in UK Financial Services.”](#) Research Note.
- FinRegLab. 2019. [“The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings.”](#)
- . 2020a. [“The Use of Cash-Flow Data in Underwriting Credit: Market Context & Policy Analysis.”](#)
- . 2020b. [“The Use of Cash-Flow Data in Underwriting Credit: Policy Overview.”](#)
- Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. [“Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.”](#) Berkman Klein Center for Internet & Society.

Bibliography

- G20. 2019. [“G20 Ministerial Statement on Trade and Digital Economy.”](#)
- Geburu, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. [“Datasheets for Datasets.”](#) Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning.
- Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2019. [“Explaining Explanations: An Overview of Interpretability of Machine Learning.”](#) The 5th IEEE International Conference on Data Science and Advanced Analytics.
- Government Office for Science. 2011. [“Crashes and High Frequency Trading.”](#)
- . 2012. [“The Future of Computer Trading in Financial Markets: An International Perspective.”](#) Final project report.
- Henderson, Peter, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. [“Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning.”](#)
- High-Level Expert Group on Artificial Intelligence. 2019. [“Ethics Guidelines for Trustworthy AI.”](#) European Commission.
- Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. [“The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards.”](#)
- Hong Kong Monetary Authority. 2019a. [“High-Level Principles on Artificial Intelligence.”](#)
- . 2019b. [“Consumer Protection in Respect of Use of Big Data Analytics and Artificial Intelligence by Authorized Institutions.”](#)
- Hu, Xiyang, Cynthia Rudin, and Margo Seltzer. 2020. [“Optimal Sparse Decision Trees.”](#) Advances in Neural Information Processing Systems. Vol. 32.
- Huynh, Dong, Sophie Stalla-Bourdillon, and Luc Moreau. 2019. [“Provenance-Based Explanations for Automated Decisions.”](#)
- ICO. 2016. [“Data Protection Impact Assessments.”](#)
- . 2019. [“Guide to the General Data Protection Regulation \(GDPR\).”](#)
- . 2020. [“Guidance on AI and Data Protection.”](#)
- ICO and The Alan Turing Institute. 2020. [“Explaining Decisions Made with AI.”](#)
- IEEE. 2020. “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems.” <https://ethicsinaction.ieee.org/>.
- International Competition Network. 2020. [“Scoping Paper: The Impact of Digitalization in Cartel Enforcement.”](#)
- Kelley, Patrick Gage, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. [“A ‘Nutrition Label’ for Privacy.”](#) Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09.
- Kirilenko, Andrei A, and Andrew W Lo. 2013. [“Moore’s Law versus Murphy’s Law: Algorithmic Trading and Its Discontents.”](#) Journal of Economic Perspectives 27, no. 2: 51–72.
- Koshiyama, Adriano, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, et al. 2021. [“Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms.”](#)

Bibliography

- Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. [“Quantifying the Carbon Emissions of Machine Learning.”](#)
- Lee, Michelle Seng Ah, and Jatinder Singh. 2020. [“Spelling Errors and Non-Standard Language in Peer-to-Peer Loan Applications and the Borrower’s Probability of Default.”](#)
- . 2021. [“The Landscape and Gaps in Open Source Fairness Toolkits.”](#) Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.
- Leslie, David. 2019. [“Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector.”](#) The Alan Turing Institute.
- Li, Oscar, Hao Liu, Chaofan Chen, and Cynthia Rudin. 2018. [“Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network That Explains Its Predictions.”](#) Thirty-Second AAAI Conference on Artificial Intelligence.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [“Model Cards for Model Reporting.”](#) Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19.
- Mittelstadt, Brent, Chris Russell, and Sandra Wachter. 2019. [“Explaining Explanations in AI.”](#) Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19.
- Monetary Authority of Singapore. 2019. [“Principles to Promote Fairness, Ethics, Accountability and Transparency \(FEAT\) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector.”](#)
- Money and Mental Health Policy Institute. 2019. [“Data Protecting: Using Financial Data to Support Customers.”](#)
- National Institute of Standards and Technology. 2019. [“U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools.”](#)
- OCEANIS. 2021. [“Global AI Standards Repository.”](#)
- OECD. 2017. [“Algorithms and Collusion: Competition Policy in the Digital Age.”](#)
- . 2018. [“Personalised Pricing in the Digital Era.”](#)
- . 2019. [“Recommendation of the Council on Artificial Intelligence.”](#)
- Office of Fair Trading. 2013a. [“Personalised Pricing: Increasing Transparency to Improve Trust.”](#)
- . 2013b. [“The Economics of Online Personalised Pricing.”](#)
- Parliamentary Office of Science and Technology. 2020. [“Interpretable Machine Learning.”](#) 633 POSTNOTE.
- Raji, Inioluwa Deborah, Andrew Smart, Rebecca White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. [“Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing.”](#) Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.
- Raji, Inioluwa Deborah, and Jingying Yang. 2020. [“ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles.”](#)

Bibliography

- Royal Statistical Society. 2020. [“Professional Standards to Be Set for Data Science.”](#)
- Rudin, Cynthia. 2019. [“Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.”](#) *Nature Machine Intelligence* 1, no. 5: 206–15.
- Selbst, Andrew D., and Solon Barocas. 2018. [“The Intuitive Appeal of Explainable Machines.”](#) *Fordham Law Review* 87, no. 3: 1085–1139.
- Select Committee on Artificial Intelligence. 2019. [“AI in the UK: Ready, Willing and Able?”](#) House of Lords.
- Sokol, Kacper, and Peter Flach. 2020a. [“One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency.”](#) *KI - Künstliche Intelligenz* 34, no. 2: 235–50.
- . 2020b. [“Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches.”](#) Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.
- Sokolovska, Nataliya, Yann Chevaileyre, and Jean-Daniel Zucker. 2018. [“A Provable Algorithm for Learning Interpretable Scoring Systems.”](#) Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR.
- Stiftung Neue Verantwortung. 2020. [“AI Governance through Political Fora and Standards. Developing Organisations: Mapping the Actors Relevant to AI Governance.”](#)
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. [“Energy and Policy Considerations for Deep Learning in NLP.”](#) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
- The Royal Society. 2019. [“Explainable AI: The Basics.”](#)
- . 2020. [“What Makes a Data Scientist? Why Professional Skills and Behaviours Are More Important than Ever.”](#)
- UK Finance and KPMG. 2020. [“Ethical Principles for Advanced Analytics and Artificial Intelligence in Financial Services.”](#)
- Université de Montréal. 2018. [“Montréal Declaration for a Responsible Development of Artificial Intelligence.”](#)
- VDE and Bertelsmann Stiftung. 2020. [“From Principles to Practice: An Interdisciplinary Framework to Operationalise AI Ethics.”](#)
- Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. [“Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.”](#) *Harvard Journal of Law & Technology* 31, no. 2: 841–87.
- World Economic Forum. 2018. [“The New Physics of Financial Services: Understanding How Artificial Intelligence Is Transforming the Financial Ecosystem.”](#)
- . 2019. [“Navigating Uncharted Waters: A Roadmap to Responsible Innovation with AI in Financial Services.”](#)



turing.ac.uk
@turinginst