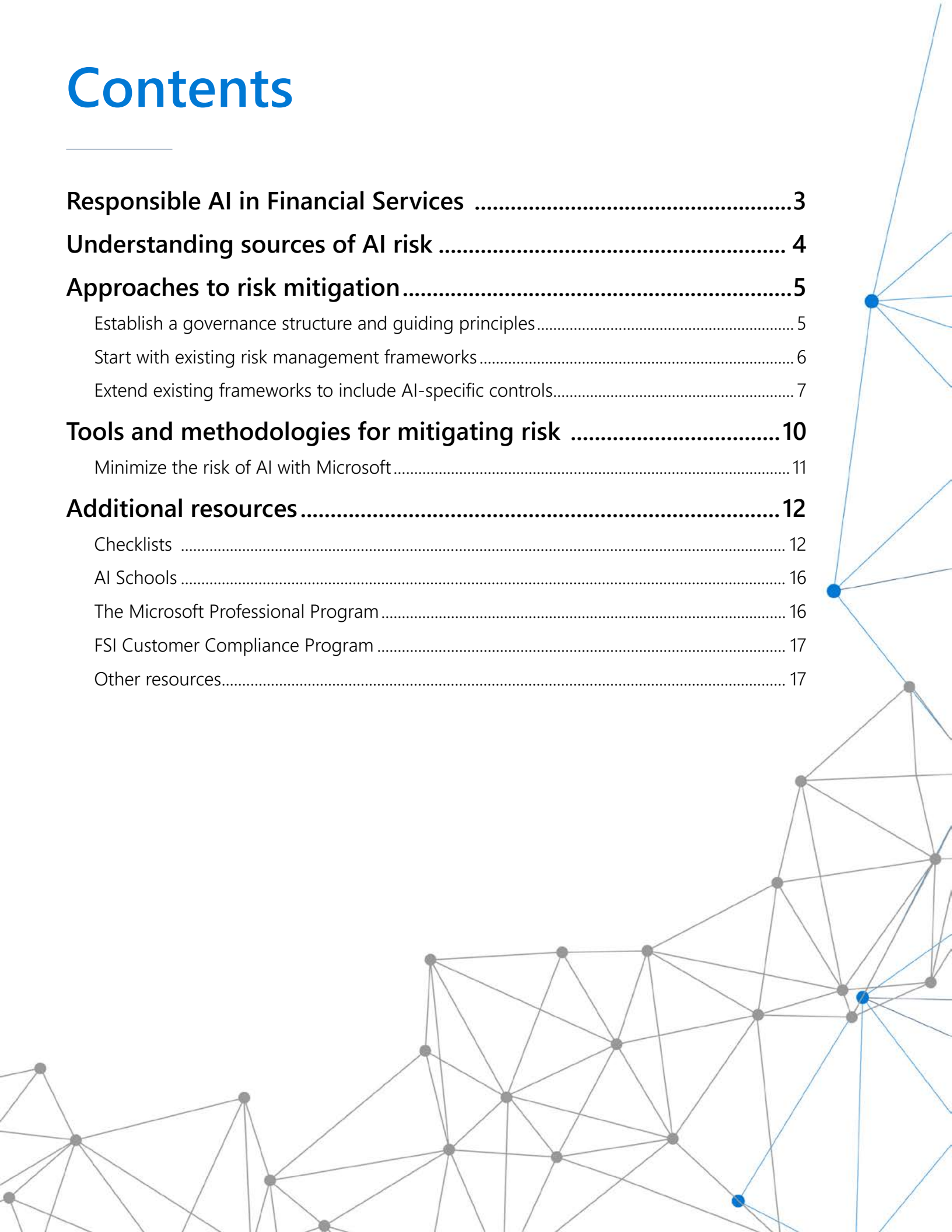
Responsible AI in financial services: governance and risk management

Driving transparency and accountability to mitigate AI risks

Daragh Morrissey, Director AI, Microsoft Worldwide Financial Services
Nick Lewins, Financial Services Lead, MIcrosoft Research

Microsoft

Contents



Responsible AI in Financial Services

Financial services organizations play a central role in the financial wellbeing of individuals, communities, and businesses. Every day, these companies make decisions that have a significant impact on people, such as approving or withholding credit, foreclosing on a mortgage, or deciding whether to pay out a life insurance claim. At the level of an individual company, these are serious responsibilities—and in aggregate can have significant impacts on economies and nations.

Financial services organizations are also increasingly using AI to optimize their operations. With the ability to ingest and analyze vast amounts of data, AI has the power to help financial organizations better understand their customers and identify fraud or security breaches more quickly and efficiently. Given the breadth of potential applications, it's unsurprising that the market for AI has exploded, with total spend forecasted to reach US$46 billion by 2020, with a quarter of this coming from the financial services industry.[1]

As AI takes on a bigger role in the financial services industry, it's essential that organizations use it responsibly and plan for unintended consequences. When an organization deploys AI, it may inadvertently deny people consequential services, amplify gender or racial biases present in training data, or violate data protection and privacy laws such as Europe's GDPR or the United States' ECOA*. In order to address these concerns, enterprises are finding the need to create internal internal controls to guide their AI efforts, whether they are deploying third-party AI solutions or developing their own.

* Equal Credit Opportunity Act

This paper is part one of a two-part series.

In part one, *Microsoft's perspective on responsible AI in financial services*, we explored Microsoft's ethical guiding principles for AI solutions and how they apply to financial services.

In this paper, we explore how to implement governance and risk management to foster the responsible use of AI.

The good news: financial services organizations can get a head start by leveraging existing risk management frameworks. This paper is intended to guide you on how to create an internal governance system for AI. While every company will have unique processes and controls, we'd like to share what we've learned to provide a starting point. It will take collaboration—between consumer advocates, financial institutions, technology companies, policymakers, and regulators—to maximize the benefits of AI while managing industry-wide risks.

Understanding sources of AI risk

As discussed in *Microsoft's Perspective on Responsible AI in Financial Services*, AI can create possible harms around fairness, reliability, safety, and privacy. These not only pose risk to the organization but to end users and society at large. These risks often result from issues with the data, the AI model itself, or the model's usage scenario.

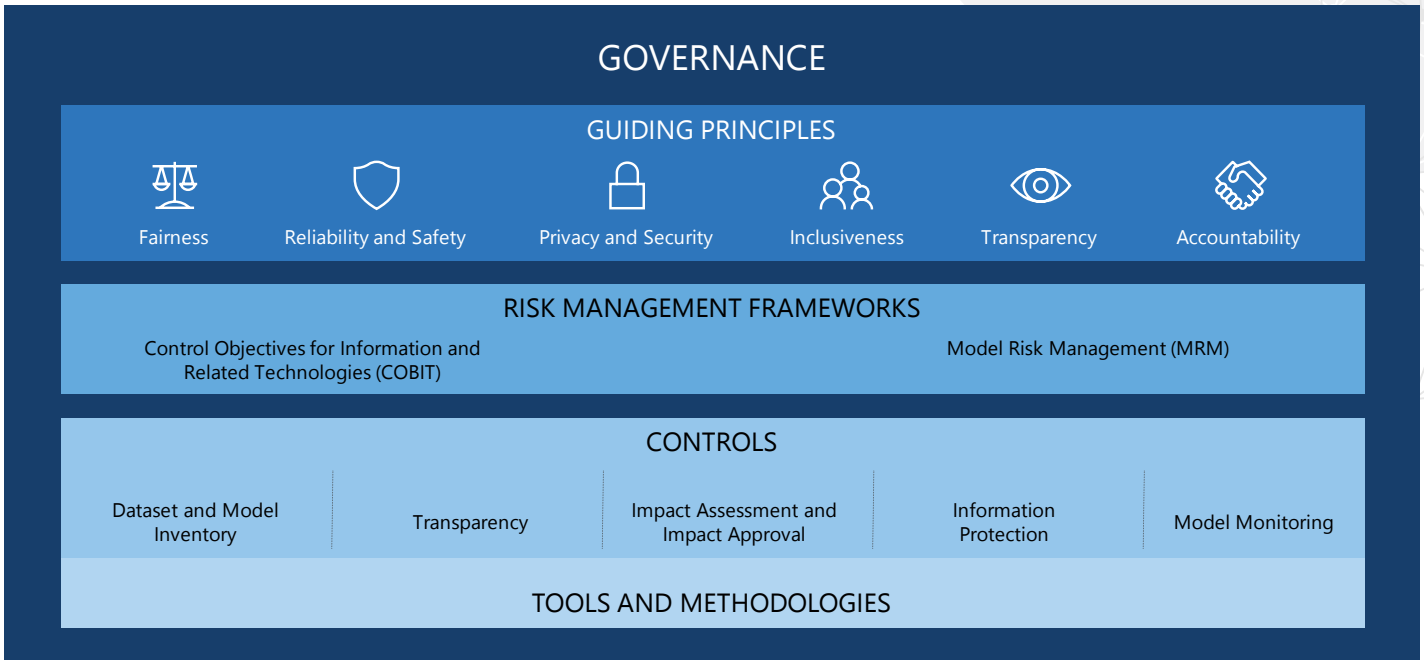
- **Data:** Before machine learning models are used, they are trained to recognize patterns using "training data." It's important for organizations to be careful and considerate about what training data is used and how it is structured. Flawed training data will create flawed AI models. If there's only data from one age group or one time of year, for example, the AI model will end up skewed or biased. Risks can arise if the data has errors, lacks critical variables or historical depth, is an insufficient sample size, or doesn't match the deployment context of the model.
- **Model:** AI models themselves can also create risks if they are not well-designed, which could cause them to make incorrect approximations, choose an inappropriate objective to optimize, or using a variable as a proxy for another in a way that unintentionally introduces bias. It's important for organizations to consider potential issues before deploying a model. After it's deployed, they can monitor model performance and accuracy during the training process and on an ongoing basis, and re-train the model periodically. AI models can "drift" or decline in performance over time, and even if a model is performing well it should be updated periodically to take advantage of more recent training data.
- **Usage scenario:** Each usage scenario may contain potential harms and should be subjected to risk assessment and governance approval. It's important to ensure that each AI model is only used for the purpose for which it was designed and approved.

With such inherent risks, it's imperative that organizations maintain oversight and control of AI solutions to ensure compliance with regulations and ethical principles

Approaches to risk mitigation

At Microsoft, we see several key components for migitating AI risks. The first is to establish a governance structure and guiding principles. From there, companies can leverage risk management frameworks, controls, and tools and methodologies to manage AI systems. In this section, we'll highlight our perspective on each of these components as they relate to AI, with the hopes of providing concrete examples you can use in your organization.

Whereas the logic inside most IT systems is programmed explicitly, an AI system learns its logic by extrapolating from examples. This extrapolated logic is called a "model."



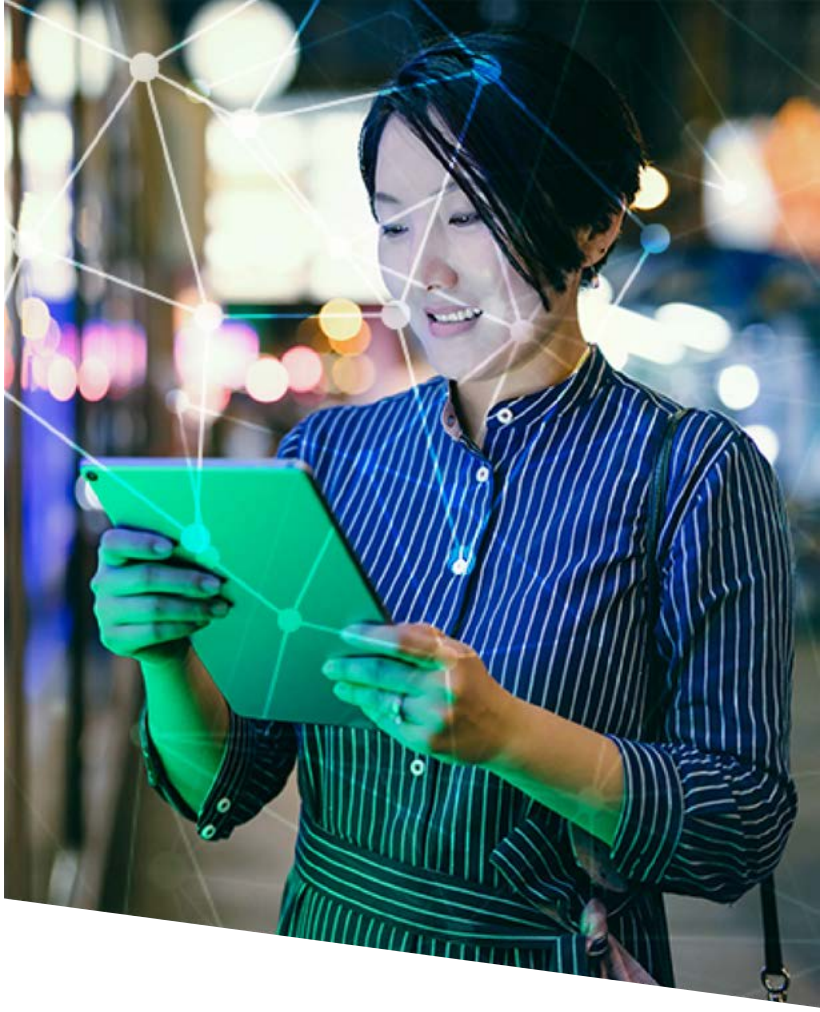
Establish a governance structure and guiding principles

Many financial institutions are building out governance functions for AI, often called AI Centers of Excellence (COEs). To be successful, COEs need solid executive sponsorship enable a coalition of the willing across the organization to collaborate effectively. There are several common structures for these teams, which can either be centralized, decentralized, or a hybrid.

COEs perform many functions, including identifying opportunities to enhance the company's operations with AI, creating and landing the organization's AI vision, and measuring the business impact of each AI solution. However, the first step that we recommend for an AI governance system is to establish guiding ethical principles that will serve as a foundation for all of the company's activities related to AI. After they are defined, COEs can then develop and enforce policies, standards, and enterprise risk management frameworks that align to the organization's guiding principles surrounding AI.

With guiding principles and standards in place, a governance system for AI could perform functions such as:

- Recruit internal champions who can promote and evangelize responsible AI practices across the organization
- Lead change management processes to ease AI adoption, including training and readiness programs for employees
- Provide a mechanism for employees to suggest AI use cases, and submit projects for ethical review
- Work with technical teams to create a programmatic, repeatable, and responsible approach to developing AI solutions to streamline the solution development lifecycle
- Determine the nature and extent of risks that the organization is willing to take on in the pursuit of its goals



However, before developing new risk management processes, organizations can benefit from evaluating how AI fits into their existing frameworks.

Start with existing risk management frameworks

Financial services organizations don't have to start from scratch to create new risk management frameworks for AI.

As a baseline for AI risk management, organizations can apply general IT controls including Control Objectives for Information and Related Technologies (COBIT). Since AI systems use machine learning models, generalized Model Risk Management (MRM) is a good starting point as well. Banks operating under Basel II IRB Advanced Status for Credit Risk will be familiar with Model Risk Management (MRM), as will insurers using actuarial models for insurance risk.[2] MRM is already widely used to govern statistical financial models, but it applies to AI models as well. For example, organizations can leverage existing regulatory guidance on managing model risk from the OCC's Supervisory Note 11-7. Less material models may only require a subset of these controls.

In many cases, existing frameworks are capable of addressing issues raised by the use of AI. However, when the use of AI raises concerns not addressed by an existing framework—like intelligibility or bias—then additional controls and processes need to be implemented.

Extend existing frameworks to include AI-specific controls

While existing risk management frameworks are a good starting point, AI models present specific challenges because they rely on extrapolated logic instead of hard-coded rules, and they have the capacity to continuously learn. This can result in unintended consequences around fairness or changes in reliability over time. AI also relies on data for training and predictions, which can create data privacy and security risks. Compounding these issues is the fact that AI solutions sometimes function as a black box, where users don't understand how the system's outputs were derived from its inputs, making it harder to identify and resolve these issues.

AI governance systems will need to thoughtfully address concerns raised by AI in line with their organization's guiding principles, COE, and use cases. While every organization will have its own unique risk management framework, we recommend that governance systems consider the following key controls to mitigate AI risks:

Dataset and model inventory

- Scrutinize data sets used for AI applications to ensure they are representative of the wider population relevant for the use case.
- Train design teams to be aware of, and mitigate, biases within datasets, models, and applications.
- Ensure everyone working on the model understands where the datasets came from and what they contain and update the team on any dataset changes that take place.
- Implement a database of all AI models, versions, datasets, documentation, use cases, and their level of materiality.
- Create a workflow for tracking the approval status of each AI model.

Transparency

- Assess what level of transparency is needed for each AI application (based on the materiality of its potential impact on customers, interdependency with other applications, and regulatory obligations) and take this into account during design.
- Based on this assessment, build the following components of transparency into your AI application:
  - **Traceability:** Clearly document goals, definitions, design choices, and any assumptions made during the development process. Also document the provenance, source, and quality of initial training datasets and additional datasets used for re-training.
  - **Communication:** Be forthcoming about when, why, and how you choose to build and deploy AI, as well as the system's limitations.
  - **Intelligibility:** Intelligibility refers to the ability of people to understand and monitor the technical behavior of an AI system. But simply publishing the underlying algorithms and datasets rarely provides meaningful transparency, as these can be largely incomprehensible to most people, particularly with more complex systems like deep neural networks. Luckily, a number of promising approaches to achieving intelligibility are emerging. Some facilitate understanding of key characteristics of the datasets used to train and test models. Others focus on explaining why individual outputs were produced or predictions were made. Even more offer simplified but human-understandable explanations for the overall behavior of a trained model or entire AI system. Explore the range of available intelligibility approaches and select those that most effectively provide the information about the system or its components that each stakeholder needs to understand in order to meet their goals.

Impact assessment and impact approval

- Test and document the impact and quality of data, AI models, and use cases.
- Generally, this would be done in an Impact Assessment document that usually includes assessment of materiality, inherent risk, controls and residual risk, and impact on fairness, security, and privacy.
- Where possible, this would leverage existing organizational policy frameworks such as privacy policy/DPIA (data protection impact assessment), data quality policy, risk management framework, or information classification policy.
- Ensure that auditors can reproduce and verify the analysis performed for the impact assessment.
- Have the impact and residual risk approved by a governing body with the appropriate organizational delegation. This could be an AI ethics committee, an AI model risk committee, and/or the accountable executive for the line of business.

Information protection

- Ensure sensitive information can only be used by authorized staff for authorized purposes.
- Establish preventative and detective controls against insider conduct risk. For example, the risk of data scientists misusing the training data by retrieving the sensitive data on famous people or people they know.

Model monitoring

- Ensure that all AI applications are subject to a suitable control framework and audit process throughout their lifecycle.
- Monitor AI model performance in production against baseline standards (this should be done more frequently for highly material models).
- Establish trigger thresholds to prompt additional impact assessments when original data relied on for the assessment changes or model performance degrades.

Tools and methodologies for mitigating risk

Microsoft is investing in a wide range of tools and methodologies to help you mitigate AI risks. Below is a partial list of Microsoft's publicly available tools and publications:

Key control for material AI deployments	Tools and Methodologies
Dataset and model inventory	Keeping a database of AI models allows you to reproduce them and document their versions, uses, and governance approvals over time. • DevOps in Azure Machine Learning (MLOps) makes it easier to track and reproduce models and their version histories. MLOps offers centralized management throughout the entire model development process (data preparation, experimentation, model training, model management, deployment, and monitoring) while providing mechanisms for automating, sharing, and reproducing models.
Transparency	Transparency is a key part of the impact assessment. The tools below can be used for interpreting AI models: • InterpretML is an open-source package for training interpretable models and explaining black box systems. It includes several methods for generating explanations of the behavior of models or their individual predictions (including Explainable Boosting Machine (EBM)), enabling developers to compare and contrast explanations and select methods best suited to their needs. • Model Interpretability is a feature in Azure Machine Learning that enables model designers and evaluators to explain why a model makes the predictions it does. These insights can be used to debug the model, validate that its behavior matches objectives, check for bias, and build trust. • Datasheets for datasets is a paper proposing that dataset creators should include in a datasheet for their dataset, such as training datasets, model inputs and outputs, and model features. Like a datasheet for electronic components, a datasheet for datasets would help developers understand if a specific dataset is appropriate for their use case. • Local Interpretable Model-agnostics Explanations (LIME) provides an easily understood description of a machine learning classifier by perturbing the input and seeing how the predictions change.
Impact assessment and impact approval	Performing an impact assessment and getting it approved by the accountable executive is a key control. The tools below can be used for interpreting AI models for the impact assessment:

	• Methodology for reducing bias in word embedding helps reduce gender biases by modifying embeddings to reduce gender stereotypes, such as the association between receptionist and female, while maintaining potentially useful associations, such as the association between the words queen and female. • A reductions approach to fair classification provides a method for turning any common classifier AI model into a "fair" classifier model according to any of a wide range of fairness definitions. For example, consider a machine learning system tasked with choosing applicants to interview for a job. This method can turn an AI model that predicts who should be interviewed based on previous hiring decisions into a model that predicts who should be interviewed while also respecting demographic parity (or another fairness definition).
Information protection	Protecting sensitive data elements is a key control and we have a range of tools to help with this: • Securing the Future of Artificial Intelligence and Machine Learning at Microsoft provides guidance on how to protect algorithms, data, and services from new AI-specific security threats. While security is a constantly changing field, this paper outlines emerging engineering challenges and shares initial thoughts on potential remediation. • Homomorphic encryption is a special type of encryption technique that allows users to compute on encrypted data without decrypting it. The results of the computations are encrypted and can be revealed only by the owner of the decryption key. To further the use of this important encryption technique, we developed the Simple Encrypted Arithmetic Library (SEAL) and made it open source. • Multi-party computation (MPC) allows a set of parties to share encrypted data and algorithms with each other while preserving input privacy and ensuring that no party sees information about other members. For example, with MPC we can build a system that analyzes data from all three hospitals without any of them gaining access to each other's health data. • Differential Privacy, a mathematical definition of privacy invented by Cynthia Dwork in 2006 at Microsoft Research Labs, enables data scientists to learn about general characteristics of populations while guaranteeing the privacy of any individual's records. • Artificial Intelligence and the GDPR Challenge, a whitepaper authored by repretantatives from Microsoft's Corporate, External, & Legal Affairs (CELA), addresses issues of AI explaniability and provides considerations surrounding GDPR requirements for AI fairness in credit scoring and insurance underwriting.
Model monitoring	AI model monitoring is an ongoing control to check for model performance degradation. We have capabilities for this in Azure Machine Learning: • DevOps in Azure Machine Learning (MLOps) helps teams monitor model performance by collecting application and model telemetry. These features can help banks to audit changes to their AI models, automate testing, and reproduce model outputs.

Minimize the risk of AI with Microsoft

The benefits of AI in the financial services industry are too great to ignore. With the technology, many organizations can improve their customer experience, empower their employees to do better work, and optimize internal operations to increase efficiency and lower costs.

However, along with its differentiated value, AI brings many unique risks and challenges to financial services organizations. To avoid unintended consequences, mitigate risk, and minimize bias, organizations must leverage data and AI responsibly.

We encourage all financial services organizations to create their own guiding principles and structure for continual oversight. AI is constantly evolving, and we're continually learning how to implement responsible AI practices alongside industry leaders. We are eager to collaborate with you to help the financial services industry harness the vital power of AI while mitigating the risks and uncertainties it brings.

Additional resources

Checklists

To help you consider how to implement ethical principles in your own organization, we developed the following recommendations:

Fairness

- ☐ **Understand the scope, spirit, and potential uses of the AI system** by asking questions such as, how is the system intended to work? Who is the system designed to work for? Will it work for everyone equally? How can it harm others?
- ☐ **Attract a diverse pool of talent.** To the best of your ability, ensure the design team reflects the world in which we live by including team members that have different backgrounds, experiences, education and perspectives. It is also crucial to recruit diverse perspectives from outside the team, as groupthink and internal politics may lead to unintended bias or risk.

- ☐ **Identify potentially harmful sources of bias in datasets** by evaluating where the data came from, understanding how it was organized, and testing to ensure the model produces fair outcomes. Bias can be introduced at every stage in creation, from collection to modeling to operation.
- ☐ **Identify bias in machine learning algorithms** by leveraging tools and techniques that improve the transparency and intelligibility of models.
- ☐ **Leverage human review and domain expertise.** Train employees to understand the meaning and implications of AI results to ensure that they are ultimately accountable for decisions that leverage AI, especially when AI is used to inform consequential decisions about people. Finally, include relevant subject matter experts (such as those with consumer credit expertise for a credit scoring AI system) in the design process and in deployment decisions.
- ☐ **Research and employ best practices, analytical techniques, and tools** from other institutions and enterprises to help detect, prevent, and address bias in AI systems.

Reliability and Safety

- ☐ **Understand your organization's AI Maturity** by taking Microsoft's AI Ready Assessment. Use the results to determine which AI technologies will fit your organization's current maturity level and how your organization can best take advantage of AI.
- ☐ **Develop processes for auditing AI systems** in order to evaluate the quality and suitability of data and models, monitor ongoing performance, and verify that systems are behaving as intended based on established performance measures.
- ☐ **Provide detailed explanation of system operation** including design specifications, information about training data, training failures that occurred, potential inadequacies with training data, and the inferences and significant predictions generated.
- ☐ **Design for unintended circumstances** such as accidental system interactions, the introduction of malicious data, or cyberattacks.
- ☐ **Involve domain experts in the design and implementation processes**, especially when AI is being used to help make consequential decisions about people.
- ☐ **Conduct rigorous testing during AI system development and deployment** to ensure that systems can respond safely to unanticipated circumstances, don't have unexpected performance failures, and don't evolve in unexpected ways. AI systems involved in high-stakes scenarios that affect human safety or large populations should be tested both in lab and real-world scenarios.

- ☐ **Evaluate when and how an AI system should seek human input for impactful decisions or during critical situations.** Consider how an AI system should transfer control to a human in a manner that is meaningful and intelligible. Design AI systems to ensure humans have the necessary level of input on highly impactful decisions.
- ☐ **Develop a robust feedback mechanism for users to report performance issues** so that they can be resolved quickly.

Privacy and Security

- ☐ **Comply with relevant data protection, privacy, and transparency laws** like GDPR or the California Privacy Act by investing resources in developing compliance technologies and processes or working with a technology leader during the development of AI systems. Develop processes to continually check that the AI systems are satisfying all aspects of these laws.
- ☐ **Design AI systems to maintain the integrity of personal data** so that they can only use personal data during the time it's required and for the defined purposes that have been shared with customers. Delete inadvertently collected personal data or data that is no longer relevant to the defined purpose.
- ☐ **Protect AI systems from bad actors** by designing AI systems in accordance with secure development and operations foundations, using role-based access, and protecting personal and confidential data that is transferred to third parties. Design AI systems to identify abnormal behaviors and to prevent manipulation and malicious attacks. Learn more about how to protect against new AI-specific security threats by reading our paper, Securing the Future of Artificial Intelligence and Machine Learning at Microsoft.
- ☐ **Design AI systems with appropriate controls** for customers to make choices about how and why their data is collected and used.

- ☐ **Ensure your AI system maintains anonymity** by de-identifying personal data.
- ☐ **Conduct privacy and security reviews** for all AI systems.
- ☐ **Research and implement industry best practices** for tracking relevant information about customer data, accessing and using that data, and auditing access and use.

Inclusiveness

- ☐ **Comply with laws regarding accessibility and inclusiveness** such as the Americans with Disabilities Act, the Communications and Video Accessibility Act, and the European Union laws and U.S. regulations that mandate the procurement of accessible technology.
- ☐ **Use the** Inclusive Design toolkit to help system developers understand and address potential barriers in a product environment that could unintentionally exclude people.
- ☐ **Have a diverse group of people test your systems** to help you determine whether the system can be used as intended by the broadest possible audience. The testing group should include people with differing perspectives, skill levels, and disabilities to ensure your model and solution are usable by as many people as possible.
- ☐ **Consider commonly used accessibility standards** to help ensure your system is accessible for people of all abilities.

Transparency

- ☐ **Document goals, definitions, design choices, assumptions, system behavior, and model limitations** throughout all stages of the AI lifecycle.
- ☐ **Share model and dataset documentation** with stakeholders so they gain a full understanding of both and can decide if they are appropriate for their use case.
- ☐ **Improve model intelligibility** by leveraging simpler models and generating intelligible explanations of the model's behavior that address the particular goals and needs of key stakeholders.
- ☐ **Train employees on how to interpret AI outputs** and ensure that they remain accountable for making consequential decisions based on the results.

Accountability

- ☐ **Set up internal review boards** to provide oversight and guidance on the responsible development and deployment of AI systems.
- ☐ **Ensure your employees are trained** to use and maintain the solution in a responsible and ethical manner and understand when the solution may require additional technical support.
- ☐ **Keep humans with requisite expertise in the loop** by reporting to them and involving them in decisions about model execution. When automation of decisions is required, ensure they are able to inspect, identify, and resolve challenges with model output and execution.
- ☐ **Put in place a clear system of accountability and governance** to conduct remediation or correction activities if models are seen as behaving in an unfair or potentially harmful manner.

AI Schools

To help organizations get started with AI, Microsoft provides the following training resources.

The AI Business School provides practical advice to help business decision-makers formulate and execute an AI strategy.

The AI School is targeted at developers, technical teams, and non-technical employees looking to learn more. It has the following tracks:

- Conversational AI
- AI Services
- Machine Learning
- Autonomous Systems
- Responsible AI

The Machine Learning and Fairness webinar, part of a larger Microsoft Research series, provides guidance on how to detect and mitigate biases while developing and deploying machine learning systems.

The Microsoft Professional Program

The Microsoft Professional Program was created to help you gain technical job-ready skills and get real-world experience through online courses, hands-on labs, and expert instruction. It includes introductory courses to AI and Data Science. With each training course, the student gets access to an Azure Sandbox to build AI models and complete AI labs.

FSI Customer Compliance Program

Through our FSI Customer Compliance Program, we engage regularly with global regulators and leading FSI institutions on a range of matters including AI ethics, risk management and community expectations on adopting AI.

Other resources

Other noteworthy, globally-leading efforts include the guidelines created by the Monetary Authority of Singapore, the articles of GDPR concerning fully automated algorithmic decision-making, and Leading your organization to responsible AI, a whitepaper published by McKinsey & Company.

www.weforum.org/reports/the-new-physics-of-financial-services-how-artificial-intelligence-is-transforming-the-financial-ecosystem

www.ukfinance.org.uk/system/files/AI-2019_FINAL_ONLINE.pdf

www.iif.com/Publications/ID/3525/Machine-Learning-in-Credit-Risk-2nd-Edition-Summary-Report

Microsoft