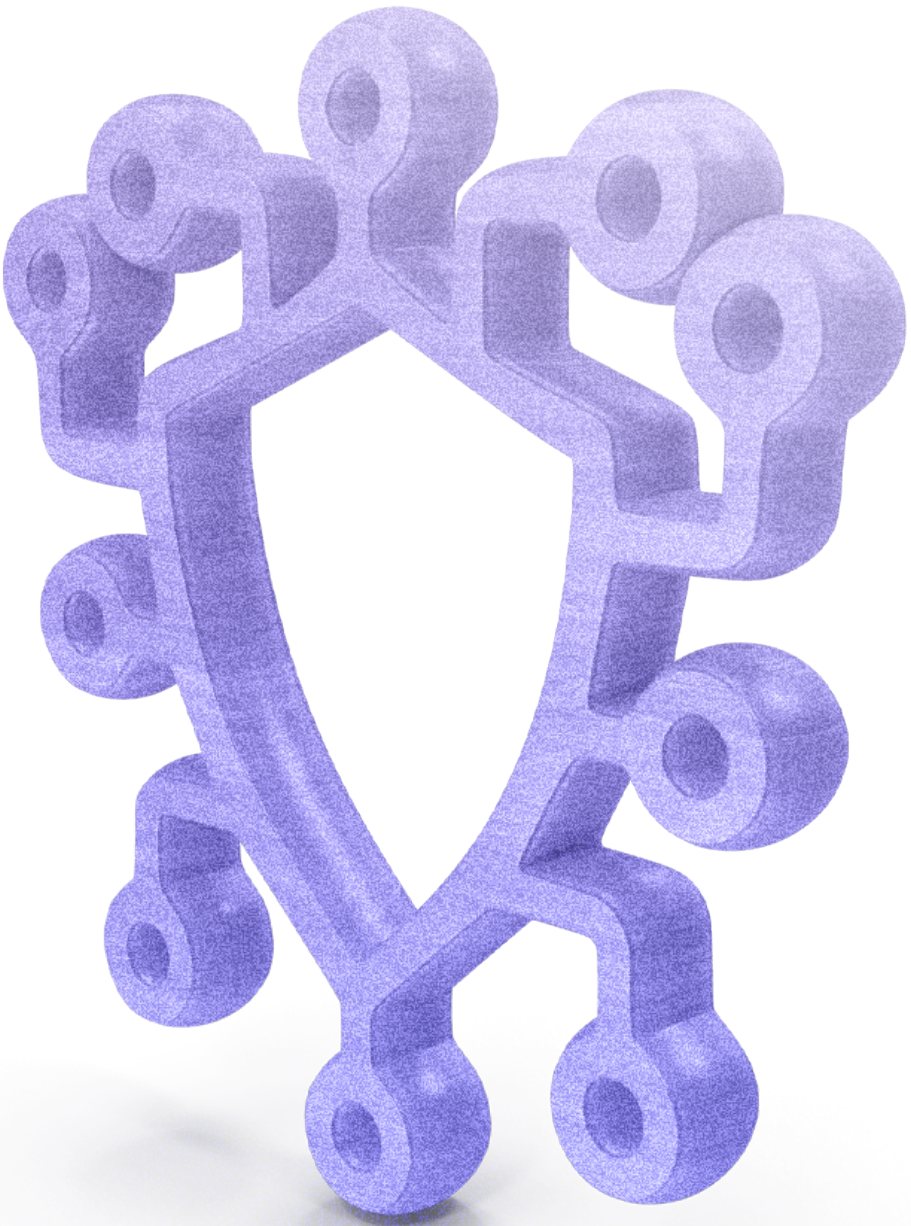
FS-ISAC

Responsible AI Principles



February 2024

Contents

EXECUTIVE SUMMARY

Artificial intelligence offers financial services organizations extraordinary opportunities.

It also precipitates decisions with profound impact on legal compliance, business success, and, most importantly, stakeholder trust. Many of those decisions apply to cybersecurity. Among the most nuanced – and consequential – involve the ethical usage of artificial intelligence (AI).

Such ethical considerations exist in a dynamic cyber ecosystem in which use cases are novel and regulations vary. FS-ISAC's AI Risk Working Group analyzed this complex landscape and developed a framework incorporating five principles – security and resiliency, explainability, privacy, fairness, reliability, and accountability – core to the responsible use and management of AI.

Designed to be actionable and practical, the framework's purpose is to help the financial services cyber community make sound decisions aligned with their institutions' values, capitalize on the opportunities AI affords – and sustain the trust the sector relies on.

INTRODUCTION

Artificial intelligence ushered in a new era of innovation and transformative potential within the financial services sector. As institutions harness the power of AI to enhance customer experiences, streamline operations, and drive financial insights, the institutions and their leaders must navigate the complex landscape of ethical considerations.

These considerations have significant business implications. The ethical development, implementation, use, and governance of AI systems instills confidence in AI solutions, prevents regulatory scrutiny and penalties, and can avert financial losses due to misunderstood or misused AI outputs. Most importantly, these principles foster the trust of customers, investors, and regulators, which is paramount in the financial services industry.

Senior leadership is crucial to these efforts. They set the tone for the ethical use of AI through their support and decisions. Their awareness of AI's use and downstream impact ensures alignment with their decisions. Their accountability and involvement are necessary for the successful implementation, use, and governance of AI. And by keeping the staff members who use AI in a feedback loop, they align AI usage with their expectations.

In this paper, we explore the foundational areas of responsible AI deployment within financial institutions. Our aim is to furnish a holistic framework that empowers financial institutions to align their AI practices with the highest level of ethics and trustworthiness. The framework requires a thoughtful, adaptive approach rooted in an understanding of the potential consequences of AI usage. In this evolving ecosystem, simple directives are unlikely to serve AI practitioners and stakeholders as effectively as well-established and commonly agreed upon principles.

With the dynamic nature of this ecosystem in mind, the principles in this document were designed to provide a basis for decision-making even as technology changes over time.

PRINCIPLES

I. Safe, Secure, and Resilient AI Systems

AI systems must be trustworthy and demonstrably safe, secure, and resilient to enhance stakeholder trust, mitigate risks, and ensure that AI systems contribute positively to organizational objectives. Manual and automated policies and procedures are equally essential to the security and resiliency of AI systems.

Proposed Approach

> **Risk assessment and mitigation:** Conduct comprehensive risk assessments for AI systems, identifying potential risks and vulnerabilities. Develop mitigation strategies and integrate them into the AI development lifecycle to minimize the likelihood of unintended consequences.

> **Robust testing and validation:** Test and validate AI systems to ensure they are reliable and support only intended operations. This includes scenario-based testing and validation against various use cases and potential disruptions.

> **Algorithmic transparency:** Prioritize the transparency of AI algorithms to ascertain that decision-making processes are understandable and interpretable. This transparency fosters trust among stakeholders and enables effective accountability.

> **Data quality and bias mitigation:** Conduct a meticulous data quality assessment and implement measures to mitigate biases in AI systems. Use high-quality, diverse datasets to enhance fairness and accuracy.

- Discuss boundaries that define appropriate system behavior.
- Define acceptable levels of bias and data quality as basic guidelines for the organization.

> **Security by design:** Implement protection mechanisms that prevent unauthorized access and the inappropriate use of data and models. Such measures enable AI systems to maintain confidentiality, integrity, and availability. Controls must be in place to protect against data poisoning, prompt injection, the exfiltration of models, training data, or other intellectual property through AI system endpoints. Embed and integrate secure development practices and controls into the AI development lifecycle to ensure system security. These practices include stringent security reviews, static and dynamic testing, code audits, and compliance with industry standards.

> **Continuous monitoring and maintenance:** Establish ongoing monitoring and maintenance protocols to ensure the continued health and performance of AI systems. Install rapid response mechanisms to address emerging risks or issues.

- Monitor data quality and bias.
- Monitor model degradation.
- Meet internal thresholds and explainability requirements related to external regulators, customers, etc.
- Align model testing and production environments and enforce separation of duties.

> **Incident response and recovery plans:** Develop comprehensive incident response and recovery plans to effectively manage and recover from unexpected events and disruptions. These plans help ensure minimal downtime, preserve system integrity, and protect institutions from unexpected or adversarial use of the model or data.

> **Regulatory compliance:** Vigilantly adhere to relevant laws, regulations, and industry standards governing AI deployment. Compliance with data privacy regulations and ethical guidelines must be a priority.

> **Continuous learning and improvement:** Foster a culture of continuous learning and improvement in AI practices. Regular training and skill development empowers teams to adapt to evolving AI challenges.

> **Transparency and documentation:** Incorporate transparency and accountability into AI practices. Thoroughly document AI systems, decision-making processes, and risk management strategies.

> **Data governance:** Holistically manage the lifecycle of data used in the system for training and input/output. Consider issues involving, but not limited to, high-level design of data flows, mitigating privacy and compliance concerns, confidentiality of queries, safeguarding sensitive data, preserving the provenance and lineage of training data, and monitoring the quality of the data used and generated.

▶ Ethical AI Principles in Practice — Problems and Solutions

Chatbot data leakage threatens privacy

GenAI clients, such as chatbots, are susceptible to data leakage and injection attacks when prompts are not validated prior to submitting to the model. Adding a data firewall between the GenAI clients and the language model anonymizes the personally identifiable information and sanitizes scripts to protect from injection attacks.

Training data corruption impacts AI output

Training data can be manipulated to corrupt the model state and output. To restore the systems after an incident, maintain copies of validated training data, models, and configurations and disaster recovery procedures.

Users diverge from ethical principles

Developers and users of AI could fail to follow the ethical principles detailed in this document or in the organization's policies. Draft an ongoing awareness program to provide regular training on responsible use and development of AI to both users and developers of AI-based systems.

II. Explainable and Interpretable AI Systems

Stakeholders may be dubious about non-transparent systems, and skepticism can undermine trust in AI outputs. Trust is paramount in financial services. Therefore, this approach advocates for a commitment to transparency with an emphasis on clarity in AI operations to sustain user and stakeholder confidence. The approach is informed by industry forerunners and governmental standards and aims to create AI systems that are technically sound, understandable, and accountable to a diverse cohort of stakeholders. The proposal for implementing transparent and interpretable AI systems underscores the necessity for stakeholder confidence, facilitates collaboration, and empowers users to make informed decisions based on AI-generated insights.

Proposed Approach

> **Transparency and interpretability:** Prioritize transparency and interpretability in AI systems with decision-making processes that are understandable to users, consumers, developers, watch groups, and other external parties.

> **Explanations for decisions:** Focus on developing AI systems that not only provide transparent decisions but also offer comprehensible explanations. This approach helps users understand the rationale behind AI-based decisions. Frequently review those explanations to ensure they remain accurate.

> **Process documention:** Establish mechanisms for transparency and documentation. Enhance interpretability by documenting AI system processes, algorithms, and decision logic.

> **Interdisciplinary collaboration:** Collaborate with experts from various domains, such as AI, ethics, and design. Incorporating diverse perspectives facilitates the development of more interpretable AI systems.

> **Human-centered design:** Incorporate human-centered design practices. Designing AI systems with the user in mind encourages AI model outputs and decisions that are comprehensible and can be effectively communicated.

> **Explainability techniques:** When applicable, leverage explainability techniques such as feature importance analysis, model visualization, and rule-based models. These techniques contribute to explainability and interpretability.

▶ Ethical AI Principles in Practice — Problems and Solutions

Traders can't justify AI-based decisions

In algorithmic trading, traders using AI systems that don't employ explainability techniques may not be able to justify the trades to clients or regulators. Explainable AI systems can communicate the key variables influencing trading decisions, such as market volatility or trade volume, making the AI system's operations more transparent to the end user.

Reliability of brokerage's market forecasting algorithm can't be validated

If an AI algorithm used to forecast market movements is not transparent, the basis for its predictions becomes unclear, and stakeholders cannot validate its reliability. For example, if the model uses high-dimensional data, stakeholders would require a simplified representation of its decision process, possibly through dimensionality reduction techniques, to understand influential market factors.

Bank can't explain its process for AI-directed fraudulent transaction blocks

If an AI system designed to detect fraudulent transactions doesn't document its process, banks may not be able to provide required information during audits or to customers disputing a transaction block. An explainable and reliable AI system would have the capability equivalent of an audit trail that logs the decision process, including the data points considered and the fraud detection rules applied.

III. Privacy-Enhanced AI Systems

Balancing transparency with "need to know" access nurtures stakeholder trust, and ensures compliance with evolving privacy regulations. Just as importantly, judiciously allocating access regarding the inputs and outputs of AI systems, standards, and controls can also ensure or improve the quality of the data in use, maximizing the quality of results. By reviewing existing privacy controls prior to incorporation into new AI solutions, organizations may reduce new risks at no additional cost, or gain greater confidence in decisions regarding investment in new controls. Moreover, implementing privacy-enhanced AI systems elevates them to or above information security standards for the sector, where data privacy concepts are already a pillar. When implementing AI we encourage applying our currently robust standards and controls and expanding them as needed.

Proposed Approach

> **Implement/ensure controls before the AI system is built:** In accordance with NIST 800-30 and related standards, consider AI implementation in the context of a System Development Life Cycle (SDLC). As such, address foreseeable privacy concerns within the earliest, pre-build phases of the SDLC to minimize the risks associated with data privacy lapses. In the case of NIST, the phases and related risk management activities would be:

SDLC Phases	Phase Characteristics	Support from Risk Management Activities
Phase 1: Initiation	The need for an IT system is expressed and the purpose and scope of the IT system is documented	Identified risks are used to support the development of the system requirements, including security requirements and a security concept of operations (strategy)
Phase 2: Development or acquisition	The IT system is designed, purchased, programmed, developed, or otherwise constructed	The risks identified during this phase can be used to support the security analyses of the IT system that may lead to architecture and design tradeoffs during system development

> **Treat privacy as a necessity constructed into systems before they are deployed:** AI systems should be meticulously designed to prioritize the protection and preservation of individual privacy during the collection and utilization of data. At a minimum, this means:

- Obtain user consent when including the information of specific individuals in the customer-facing AI systems.
- Prioritize user awareness and control, ensuring that data collection and usage are transparent and well-regulated.
- Implement measures that shield sensitive user information from unauthorized access and misuse.
- Adopt techniques such as data anonymization and data minimization. The privacy resilience of AI systems is further fortified by reducing the amount of personal information collected, used, and stored.
- Integrate privacy risk management into every stage of AI development and deployment. Conduct privacy impact assessments to identify potential privacy risks, enabling the implementation of better targeted mitigations.

> **Consider regulatory requirements (GLBA, CCPA, etc.):** Perform model evaluations for open source models that are included in the solution being developed or purchased. Keep in mind that some solutions leverage existing models that were developed separately.

- Drive a resilient security framework that prioritizes privacy defense. This framework will be instrumental in preventing unauthorized access to data, and protects the privacy foundations of AI systems.
- Consider whether the process or user of the AI system really needs the personally identifiable information (PII) in scope. Block PII by default with exceptions for adding it.

> **Adhere to industry standards for privacy:** While the current state of AI and its substantial industry impacts are new, the concern for data privacy is not. Embrace the principles of existing industry standards while scoping and designing/purchasing AI systems. Industry standards include (but are not limited to) privacy best practices established by ISC2.

ISC2 privacy best practices assure that the data used by/results from AI is:

- > Obtained fairly and lawfully
- > Used only for the original specified purpose
- > Adequate, relevant, and not excessive to purpose
- > Accurate and up to date
- > Accessible to the subject
- > Kept secure
- > Destroyed after its purpose is served

> **Establish ethical guidelines and policies for data collection, usage, and sharing:** This approach ensures that user privacy rights are protected while harnessing the potential of AI-powered insights. AI should never violate the privacy of others, unintentionally or otherwise.

> **Enforce guidelines and policies:** Deploy controls like Data Leakage Prevention (DLP) and continuous monitoring controls that already exist; these can be reconfigured to reduce privacy-related risks surrounding AI platforms as well.

IV. Fairness with Harmful Bias Managed in AI Systems

AI systems are often trained on historical data, which can include bias. An early chatbot, for example, returned misogynistic and racist replies because it was trained with offensive language[1,2]*. Such outputs can cause reputational damage, undermine trust among stakeholders, drive unfair business decisions, and subvert the values and public commitments made by financial services firms. The implementation of this proposal addresses and manages biases and directs AI systems to deliver unbiased, equitable, and just outcomes in alignment with organizational values. It helps meet regulatory requirements regarding fair and unbiased outcomes in decision-making and helps develop trust between the organization, its stakeholders, and government institutions. Not adhering to this approach can result in regulatory scrutiny and fines, reputational risk, and loss of consumer trust.*

Proposed Approach

> **Bias detection and mitigation:** Integrate bias detection and mitigation techniques into the AI development lifecycle. These measures systematically identify and rectify biases in data and algorithms so that AI systems do not perpetuate discriminatory outcomes.

> **Ethical AI development:** Root development approaches in ethical AI development practices. Scrutinize training data and models to detect and address biases, ensuring that AI systems are devoid of unfair biases.

> **Diverse and representative data:** Source a comprehensive range of demographic and cultural perspectives in training data to mitigate bias and promote fairness.

> **Regular audits and reviews:** Institute a regimen of regular audits and reviews of AI systems.These periodic assessments will identify any existing or emergent disparities, enabling corrective actions to maintain fairness.

> **Fairness metrics:** Implement fairness metrics during model training and evaluation.These metrics should provide quantifiable insights into bias levels, enabling the institution to assess and address concerns proactively.

> **User feedback:** Actively seek user and stakeholder feedback. This input is a valuable source of insights into potential biases or unfair outcomes, and can guide continuous improvement efforts.

> **Fairness impact assessments:** Incorporate fairness impact assessments in the same vein as privacy impact assessments. Fairness assessments evaluate the potential impact of AI systems on fairness and equity, driving informed decisions and risk mitigations.

> **Transparency and accountability:** Transparently communicate the institution's approach to the use, risk management, governance, and ethics for AI, such as bias detection and mitigation. Formally document these approaches in an acceptable use policy, establishing clear accountability of AI users within development and deployment processes.

▸ Ethical AI Principles in Practice — Problems and Solutions

Bank uses biased AI outputs in mortgage loan decision

AI datasets that incorporate decades of mortgage approval and rate decisions may include data reflecting unfair outcomes. That data trains the AI, so its outputs may direct mortgage loan officers toward unfair decisions. Use bias detection and mitigation techniques, test data with fairness metrics, and review the system regularly.

Investor chatbot makes inappropriate comments

Chatbots that advise on investment decisions use data drawn from large language models. That data may contain inappropriate terminology and concepts. Thoroughly test data and design guardrails for chatbots and other customer interfaces.

Insurance firm's AI produces racially homogenous advertising images

Biased training data in generative AI models design images that lack diversity or create offensive imagery. One generative AI model [3] was found to consistently produce images of White doctors for prompts requesting African doctors because its training data included only White doctors. Incorporate a range of perspectives in training data, and solicit user and stakeholder feedback to guide continuous improvement efforts.

V. Ensuring Valid and Reliable AI Systems

Consistently reliable and valid AI outputs direct effective decisions and reinforce stakeholders' trust in the organization. AI systems – especially those with no or limited human input – are better equipped to minimize risks, deliver consistent outcomes, and maintain performance excellence when they are trained, validated, monitored, transparent, and undergo continuous improvement. This approach supports stakeholder confidence and the financial services' responsibility for considered AI deployment. This section covers robust training, validation, monitoring, transparency, and continuous improvement in AI systems.

Proposed Approach

> **Robust model training:** Commit to comprehensive model training that emphasizes reliability and safety. Conduct rigorous training and meticulous testing to ensure AI systems' behavior remains consistent and accurate across diverse scenarios. A well-defined method for training models should be clear and demonstrable. A governance process or forum that supports the review of the scope of the model and its training is recommended.

> **Data quality assurance:** Validity and reliability hinge on the quality and accuracy of training data. Use stringent data quality assurance practices so that the data utilized is representative, relevant, and free from errors that could undermine system performance.

Ensuring data quality includes:

> Defining use cases for accuracy
> Defining acceptable levels of success from model training and datasets
> Ensuring data diversity and validity used for training

> **Quality assurance standards:** Align practices with quality assurance standards specific to AI development. These standards direct AI system usage toward best practices, industry standards, and established guidelines.

> **Risk assessment:** Assess potential risks associ-

ated with AI deployment based on the Responsible AI Principles. The assessment should include security, privacy, regulatory compliance, etc. incorporated in the design of the AI system. To the extent possible, the assessment should consider the potential unintended outputs the AI system may generate. These assessments empower the development of proactive measures to mitigate risks and safeguard the validity and reliability of AI systems. Share the results of the risk assessment with the relevant organizational governing bodies.

> **Testing and validation:** Conduct comprehensive functional and performance testing to ensure the AI system consistently meets its intended objectives. Well-constructed simulations may be required to fully test and validate the results. Thresholds for success for all testing should be defined proactively.

> **Model monitoring and maintenance:** To sustain the validity and reliability of AI systems, establish comprehensive model monitoring and maintenance protocols. Address evolving challenges and maintain consistent performance with regular updates and proactive maintenance.

> **Continuous monitoring:** Vigilant monitoring in real-world conditions facilitates early detection of deviations from expected behavior (e.g., model drift), enabling swift corrective actions that support validity and reliability. Consider developing monitoring metrics and thresholds for AI accuracy, drift, and other such concerns. These metrics help the organization's AI governing body ensure AI system operation within acceptable thresholds.

> **Feedback and improvement:** Gather and document feedback from users and stakeholders. Their insights identify areas for improvement so that AI systems can be fine-tuned to enhance validity and reliability.

▶ Ethical AI Principles in Practice — Problems and Solutions

CEO and Board discussions don't cover AI ethics

A meticulous, regular review of the findings and recommendations presented by the AI Ethics Committee ensures that ethical considerations are seamlessly integrated into critical decisions, such as those related to loan approval. The active participation of the institution's CEO and Board of Directors in AI-related discussions will also set an example others follow.

The AI system's training dataset isn't effective for its use case

A manual task or automated process that evaluates training data before it is ingested into a model aligns it with the type of data that will meet the use cases for the AI system, with documentation regarding the method and steps taken to perform this validation.

The AI system's users have a poor customer experience

A feedback webform regarding the AI system will give end users a platform to provide regular, useful information about their customer experience. By reviewing the feedback on a regular basis, the governing body at the organization can implement changes to enhance the system.

VI. Enhancing Accountability and Transparency in AI Systems

Prioritizing ethical AI practices assures financial services customers, regulators, and investors that decisions incorporating AI input are data-driven as well as ethically sound. This section outlines a comprehensive approach to promote accountability and transparency throughout the AI lifecycle. The proposed guidance includes clear responsibility, governance, documentation, response plans, stakeholder engagement, transparency in design, ethical considerations, communication, impact assessments, and external audits. Implementing these approaches will enhance stakeholder trust and emphasize risk management. Business continuity and harm minimization in case of AI-related incidents is more probable. The approach adds diverse perspectives to leaders' decision-making while impact assessments give it insights, optimizing AI systems' benefits while mitigating associated risks. Lastly, external validation through audits or assessments adds credibility.

Proposed Approach

> **Governance and Oversight**

- Senior leadership ownership: To foster a culture of ethical AI within financial institutions, senior leadership and management must take ownership of AI-related decisions. This entails active engagement in understanding AI projects, acknowledging their ethical implications, and providing essential guidance and oversight throughout the AI lifecycle.
- Establish governance structures: Develop and maintain well-defined governance structures tailored specifically for AI systems to provide a comprehensive framework that outlines the roles, responsibilities, and reporting lines concerning AI development, deployment, and ethical considerations. It is essential to identify the key stakeholders who will play active roles in AI decision-making.
- Define roles and responsibilities: Make a concerted effort to specify the duties of data scientists, AI developers, project managers, compliance officers, and any other relevant personnel. Effective governance hinges on the definition of roles and responsibilities for individuals engaged in AI projects. Financial institutions should clearly define each role's tasks related to ethical oversight.

> **Clear Responsibility**

- Accountability for development and deployment: Prioritize clear accountability for people engaged in the development, maintenance, security, and deployment of AI systems in the financial institution. This includes roles such as data scientists, developers, and project managers. These individuals must take ownership of the outcomes and ethical considerations related to their AI systems. By holding AI developers and maintainers accountable, financial services institutions ensure that technology is used responsibly and that any unintended consequences or ethical issues are addressed promptly.
- Third-party responsibility: When collaborating with third-party vendors or developers for AI solutions, financial institutions must ensure that these external parties share responsibility for the AI systems they provide. Contracts and agreements should be structured to clearly outline the roles, responsibilities, and accountability of third parties concerning the maintenance of ethical AI practices. This practice aligns external AI solutions with the institution's ethical standards and regulatory requirements, effectively reducing risks associated with AI outsourcing.
- User responsibility: Actively foster a culture of ethical responsibility among all users of AI solutions within the financial services organization. This is achieved through comprehensive training and guidelines provided to employees and stakeholders on the responsible use of AI tools. Users should be well-informed about the ethical considerations related to AI and should understand their role in upholding ethical practices. This proactive approach not only minimizes the risk of unintended consequences but also nurtures a culture of ethical AI within the institution.
- Security, responsibility, and oversight: In addition to ethical responsibility, financial services institutions must place paramount importance on security, responsibility, and oversight. Security teams should ensure that AI systems meet stringent security standards, protect sensitive data, and guard against potential vulnerabilities. Regular security audits, threat assessments, and risk management practices should be integral components of the security framework. By adhering to internally defined security protocols, financial institutions proactively address security concerns and defend AI systems and sensitive information.

- Multiple levels of reviews: Make decisions in multiple tiers: at the local level or the implementation team, the regional level or the business unit, and the global or corporate level. Ethical decisions should not be made in isolation. Ensure that all within the firm use a consistent approach and know what others are doing. This way, decisions are made once transparently, employees learn from one another, and tough choices do not have to be re-litigated.

> **Documentation**

- Implement comprehensive documentation: Establish a robust documentation process for all AI-related endeavors, encompassing development, deployment, and decision-making phases. This involves meticulously recording pertinent details such as shared definitions, model training specifics, data sources, algorithm choices, and ethical considerations. Financial institutions should provide a rationale behind significant decisions related to AI systems that is thoroughly documented and transparent.
- Maintain an AI usage inventory: Establish and maintain a comprehensive inventory of AI applications in use within the institution. This inventory should encompass the purpose of each AI system, the sources of data it relies upon, significant ethical considerations, and the owners responsible for its operation.

> **Transparency in Design**

- Designing for transparency: Prioritize transparency in the design of AI systems when they are being developed. Ensure that decision-making processes and algorithms are clear and understandable, even to non-technical stakeholders such as customers and regulators. Incorporate features that explain how the AI system reaches its conclusions, emphasizing fairness, transparency, and accountability.

> **Ethical Considerations**

- Financial services institutions' ethical considerations: Ensure that AI systems align with societal values, respect human rights, adhere to trust-based best practices in the financial services sector, and reflect the organization's core values and cultural norms. This commitment is fundamental to promoting ethical and responsible AI deployment within an institution and is foundational to an accountability strategy in AI implementation.
- Ethics committee formation for advancing responsible AI: Institute an ethics committee accountable for guiding the organization through the complex landscape of AI development and ensuring that ethical principles and industry best practices are rigorously upheld. In the pursuit of ethical and responsible AI deployment within financial institutions, the establishment of an ethics committee is foundational.
- Negative public perception and legal consequences: Avoid negative public perception and legal repercussions resulting from the failure to incorporate ethical considerations into AI deployment. Such incidents can damage the institution's reputation and lead to regulatory investigations and fines.

> **Stakeholder Engagement**

- Building trust through collaboration: Actively engage with stakeholders to ensure that AI deployment aligns with the expectations and values of users, customers, and the wider community affected by AI systems. Collaboration establishes channels for continuous communication, serving as a bridge between the institution and its stakeholders.
- AI governance team: Create a team to serve as the hub for stakeholder engagement. Their function is to seek perspective from a diverse group of stakeholders, including users, customers, regulators, and the public. Through surveys, forums, and direct communication, they can gather valuable input, feedback, and concerns related to AI systems.
- Customer support team: Facilitate the customer support team's ability to channel customer experiences and concerns with AI applications to the appropriate role. Because end users often interact directly with AI applications and with customer support teams, this function is an essential link in stakeholder engagement.
- Public relations team: Focus the public relations team on communicating the institution's AI initiatives and gathering feedback from the wider public. This function plays a crucial role in managing external perceptions and ensuring transparency in AI deployments.

> **Impact Assessments**

- Anticipating consequences for informed decision-making: Use impact assessments to understand and evaluate the potential consequences, both positive and negative, arising

from AI systems. Such assessments are a cornerstone of responsible AI deployment in financial services institutions, encompassing a wide array of impacts, from fraud potential to socio-demographic biases, and can be conducted internally or externally. Impact assessments provide the critical insights needed for informed decision-making and risk management.

- External impact assessments: Consider obtaining the evaluation of a neutral third party. While impact assessments are often conducted internally, external assessments performed by third parties bring an additional layer of objectivity and expertise to the financial service company's evaluation process. External auditors or specialized firms can conduct comprehensive assessments, providing an unbiased perspective on potential impacts and offering recommendations for risk mitigation.

> **Response Plans**

- Comprehensive response plans: Implement a comprehensive response plan to safeguard the institution and stakeholders. It can be instrumental in effectively managing AI-related risks and incidents. The plan should identify potential risks, foster cross-functional collaboration, create scenario-based strategies, prioritize transparency, and ensure continuous improvement. With this response, the institution can navigate the complexities of AI deployment while maintaining ethical standards, regulatory compliance, and stakeholder trust.
- Model bias oversight guidance: Develop a response plan that outlines the steps to be taken in case of identified model bias.This plan should include cross-functional collaboration between data scientists, compliance experts, and business executives. In the event of significant bias, consider corrective actions such as retraining the model with a diverse dataset and external audits to validate fairness.
- Security incident guidance: Create a response plan for security incidents involving cybersecurity experts, legal representatives, and communication specialists. Ensure that the plan addresses incident assessment, containment, and transparent communication with affected parties, including customers, regulators, and the public. Conduct a post-incident review for continuous improvement.
- Ethical concerns raised by stakeholders guidance: Establish an ethics response team to investigate and address ethical concerns raised by stakeholders. This team should engage with stakeholders, assess the impact of their concerns, and prioritize transparent communication. Collaboration with the ethics committee can provide expert insights into ethical dilemmas.

> **Communication**

- Fostering accountability and transparency: Support channels of effective, transparent communication. Communication ensures that stakeholders, both internal and external, are well in-formed about AI capabilities, limitations, and any deviations from the originally intended use.

> **External Audits**

- Demonstrating accountability and transparency to external auditors: Consider third-party external audits or independent assessments to promote accountability, transparency, and the financial institution's commitment to ethical principles. External audits can offer an outside perspective of an organization's AI practices, providing a clear picture of adherence to accountability and transparency standards

▶ Specific scenarios in which transparent actions are essential

Customer expresses concerns about the ethical implications of an AI-driven credit scoring model

The response plan should guide the institution in addressing these concerns transparently and taking appropriate actions to rectify any ethical issues.

An AI-driven chatbot provides inaccurate responses due to system drift

The response plan should outline steps to analyze and rectify the drift while transparently informing users about the issue.

An institution fails to meet regulatory requirements in its AI-based anti-money laundering system

The response plan should involve transparent communication with regulators, corrective actions, and collaboration with external auditors to validate compliance.

> The AI Governance Team should conduct surveys among bank customers to discover their preferences regarding AI-driven chatbots and engage with regulatory bodies to ensure AI applications align with industry standards. Customer support agents should actively seek feedback from customers using AI-powered financial advisory tools and relay valuable insights to the AI development team for system improvements. The public relations team should conduct public forums and webinars to educate the community about the institution's AI-driven fraud detection systems and gather feedback on the system's effectiveness, addressing concerns related to privacy and data security.

▶ Ethical AI Principles in Practice — Problems and Solutions

No documentation exists for a bank's credit scoring dataset

In the development of an AI-driven credit scoring model, meticulous documentation should encompass datasets, machine learning algorithms, data preprocessing, and strategies addressing bias. Morever, all financial services organizations should have an AI usage inventory covering diverse AI applications, including chatbots, fraud detection models, and credit risk assessment systems. The inventory should document data sources, adherence to ethical guidelines during development, and designate personnel responsible for ongoing operational oversight.

The exchange employees who use the AI system don't understand how it makes decisions

The staff responsible for AI system utilization should have access to comprehensive documentation and explanations of AI algorithms and decision-making processes. This transparency ensures that employees can effectively work with AI systems, and can comprehend how the system influences daily tasks.

A credit union customer is angry that an AI model rejected their loan application

Consider informing customers that an AI model was used in formulating responses. For instance, when employing AI for loan approval, the institution can provide customers with explanations regarding why their loan application was approved or denied by the AI model and allow them to request reconsideration by a human as appropriate. This helps foster trust and informed decision-making.

A fintech can't validate its AI system's privacy safeguards meet regulatory requirements

Meeting regulatory requirements may necessitate a higher degree of transparency. Institutions should be prepared to provide detailed documentation and transparent insights into AI systems to demonstrate compliance with regulations, encompassing information on data sources, model training, and fairness considerations.

CONCLUSION

As thorough as the authors have been in this work, we know technology and controls will evolve, regulations will advance, and overall sentiment on AI will change over time. However, an approach using time-honored principles and existing industry standards are foundational to safe, effective, ethical AI development and usage in the financial services industry.

We know that our industry will grow our morisprudence as we implement these principles, and we hope that growth only enhances the trust and confidence that are the key to our sector.

The views and opinions of the contributors are not necessarily those of their employers.

Group Chair

Benjamin Dynkin, Chair

Contributors

Lincoln Guy, *Bank of Hope*

Benjamin Dynkin, *Wells Fargo*

Lisa Matthews, *Ally Financial Inc.*

Daniel Paula, *MUFG Bank*

Daniel Sorek, *Goldman Sachs*

Kapil Pruthi, *TIAA-CREF*

Mike Silverman, *FS-ISAC*

Brad Allison, *Aflac Inc.*

REFERENCES AND RESOURCES

1 https://www.bbc.com/news/technology-35902104

2 https://en.wikipedia.org/wiki/Tay_(chatbot)

3 https://www.npr.org/sections/goatsandsoda/2023/10/06/1201840678/ai-was-asked-to-create-images-of-black-african-docs-treating-white-kids-howd-it-

National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework. (U.S. Department of Commerce, Washington, D.C.), NIST Special Publication (NIST SP) 1270, Draft, January 2023. https://doi.org/10.6028/NIST.SP.1270-draft

The White House. (2023, October 30). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

UNESCO. (2022). Recommendation on the ethics of artificial intelligence (Document code SHS/BIO/PI/2021/1). https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Board of Governors of the Federal Reserve System. (2011, April 4). SR 11-7: Guidance on Model Risk Management. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Stahl, B. C., Schroeder, D., & Rodrigues, R. (2023). Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges. Springer. https://link.springer.com/book/10.1007/978-3-031-17040-9

Hsu, M. J. (2023, June 16). Tokenization and AI in Banking: How Risk and Compliance Can Facilitate Responsible Innovation. Remarks to the American Bankers Association (ABA) Risk and Compliance Conference. https://www.occ.gov/news-issuances/speeches/2023/pub-speech-2023-64.pdf

Microsoft Corporation (2023) Embrace Responsible AI Principles and Practices. (Microsoft, Redmond, WA), Microsoft Learn Training Module, [2023]. Available at: https://learn.microsoft.com/en-us/training/modules/embrace-responsible-ai-principles-practices/

Google LLC (2023) Responsible AI: Applying AI Principles with Google Cloud. (Google, Mountain View, CA), Google Cloud Skills Boost Course, [2023]. Available at: https://www.cloudskillsboost.google/course_templates/388

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.48550/arXiv.1811.10154

Contact

fsisac.com
media@fsisac.com