AI Governance

A framework for building responsible, ethical, fair, and transparent AI

Are you ready for AI?

Business leaders must manage the associated risks as they scale their use of AI

In recent years, following technological breakthroughs and advances in development of machine learning (ML) models and management of large volumes of data, organizations are scaling their use of artificial intelligence (AI) technologies.

The use of AI and ML has gained momentum as organizations evaluate the potential applications of AI to enhance the customer experience, improve operational efficiencies, and automate business processes.

Growing applications of AI have reinforced concerns about ethical, fair, and responsible use of the technology that assists or replaces human decision-making.

Implementing AI systems requires careful management of the AI lifecycle, governing data, and machine learning model to prevent unintentional outcomes not only to an organization's brand reputation but, more importantly, to workers, individuals, and society. When adopting AI, it is important to have strong ethical and risk management frameworks surrounding its use.

"Responsible AI is the practice of designing, building and deploying AI in a manner that empowers people and businesses, and fairly impacts customers and society – allowing companies to engender trust and scale AI with confidence."

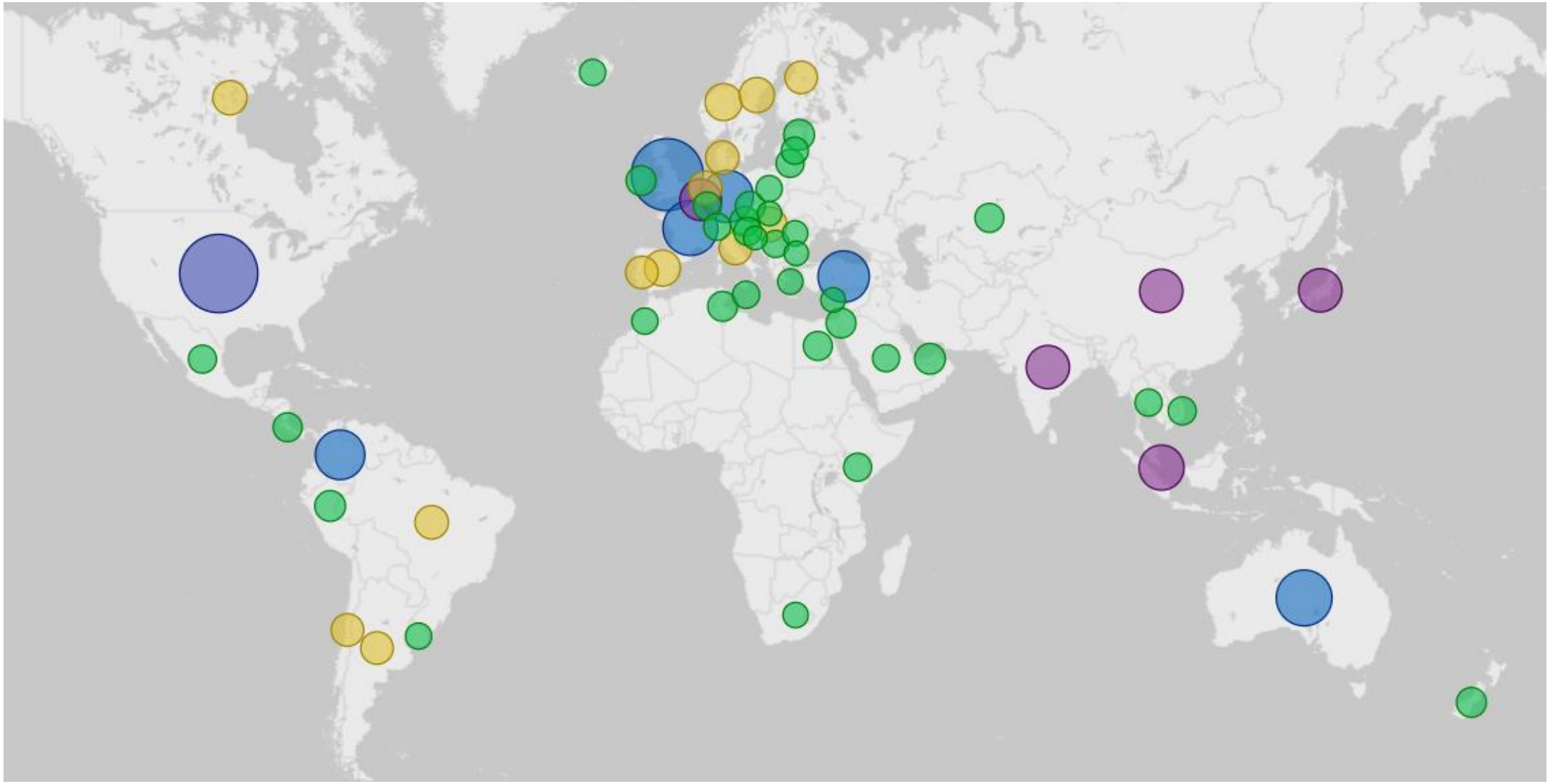
– World Economic Forum

Regulations and risk assessment tools

Governments around the world are developing AI assessment methodologies and legislation for AI. Here are a couple of examples:

- Responsible use of artificial intelligence (AI) guiding principles (Canada):
    1. **understand and measure the impact** of using AI by developing and sharing tools and approaches
    2. **be transparent** about how and when we are using AI, starting with a clear user need and public benefit
    3. **provide meaningful explanations** about AI decision-making, while also offering opportunities to review results and challenge these decisions
    4. **be as open as we can** by sharing source code, training data, and other relevant information, all while protecting personal information, system integration, and national security and defense
    5. **provide sufficient training** so that government employees developing and using AI solutions have the responsible design, function, and implementation skills needed to make AI-based public services better
- The Algorithmic Impact Assessment tool (Canada) is used to determine the impact level of an automated decision-system. It defines 48 risk and 33 mitigation questions. Assessment scores consider factors such as systems design, algorithm, decision type, impact, and data.
- The National AI Initiative Act of 2020 (DIVISION E, SEC. 5001) (US) became law on January 1, 2021. This is a program across the entire Federal government to accelerate AI research and application.
- Bill C-27, Artificial Intelligence and Data Act (AIDA) (Canada), when passed, would be the first law in Canada regulating the use of artificial intelligence systems.
- The EU Artificial Intelligence Act (EU) assigns applications of AI to three risk categories: applications and systems that create an **unacceptable risk**, such as government-run social scoring; **high-risk** applications, such as a CV-scanning tool that ranks job applicants; and lastly, applications **not** explicitly **listed as high-risk**.
- The FEAT Principles Assessment Methodology was created by the Monetary Authority of Singapore (MAS) in collaboration with other 27 industry partners for financial institutions to promote fairness, ethics, accountability, and transparency (FEAT) in the use of artificial intelligence and data analytics (AIDA).

AI policies around the world

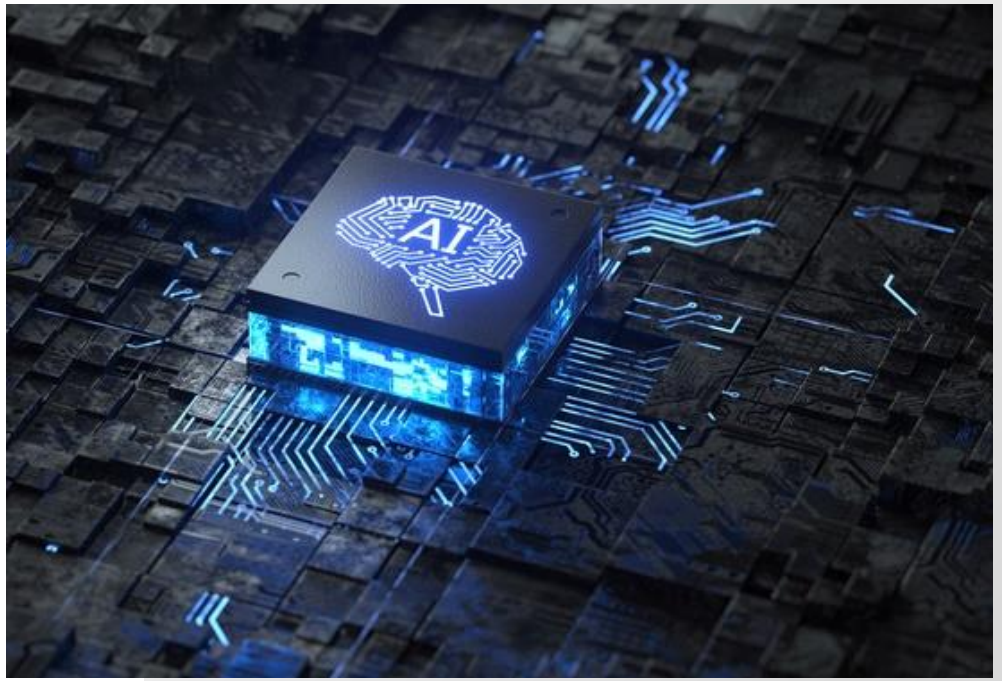


Source of data: OECD.AI (2021), powered by EC/OECD (2021), database of national AI policies, accessed on 7/09/2022, https://oecd.ai.

The need for AI governance

“To adopt AI, organizations will need to review and enhance their processes and governance frameworks to address new and evolving risks.”

– Canadian RegTech Association, Safeguarding AI Use Through Human-Centric Design, 2020



To ensure responsible, transparent, and ethical AI systems, organizations will need to review existing risk control frameworks and update them to include AI risk management and impact assessment frameworks and processes.

As ML and AI technologies are constantly evolving, the AI governance and AI risk management frameworks will need to evolve to ensure the appropriate safeguards and controls are in place.

This applies not only to the machine learning models and AI system custom built by the organization’s data science and AI team, but it also includes AI-powered vendor tools and technologies. The vendors should be able to explain how AI is used in their products, how the model was trained, and what data was used to train the model.

AI governance enables management, monitoring, and control of all AI activities within an organization.

Key concepts

Info-Tech Research Group defines the key terms used in this document as follows:

Machine learning systems learn from experience and without explicit instructions. They learn patterns from data, then analyze and make predictions based on past behavior and the patterns learned.

Artificial intelligence is a combination of technologies and can include machine learning. AI systems perform tasks that mimic human intelligence, such as learning from experience and problem solving. Most importantly, AI makes its own decisions without human intervention.

We use the definition of **data ethics** by the Open Data Institute: "Data ethics is a branch of ethics that considers the impact of data practices on people, society and the environment. The purpose of data ethics is to guide the values and conduct of data practitioners in data collection, sharing and use."

Algorithmic or machine bias is systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others. Algorithmic bias is not a technical problem. It's a social and political problem, and in the context of implementing AI for business benefits, it's a business problem.

↓ Download the blueprint *Mitigate Machine Bias* for detailed discussion on bias, fairness, and transparency in AI systems

Key concepts – explainable, transparent and trustworthy

"Responsible AI is the practice of designing, building and deploying AI in a manner that empowers people and businesses and fairly impacts customers and society – allowing companies to engender trust and scale AI with confidence" (CIFAR).

The AI system is considered **trustworthy** when people understand how the technology works and when we can assess that it's safe and reliable. We must be able to trust the output of the system and understand how the system was designed, what data was used to train it, and how it was implemented.

Explainable AI, sometimes abbreviated as XAI, refers to the ability to explain how an AI model makes predictions, its anticipated impact, and its potential biases.

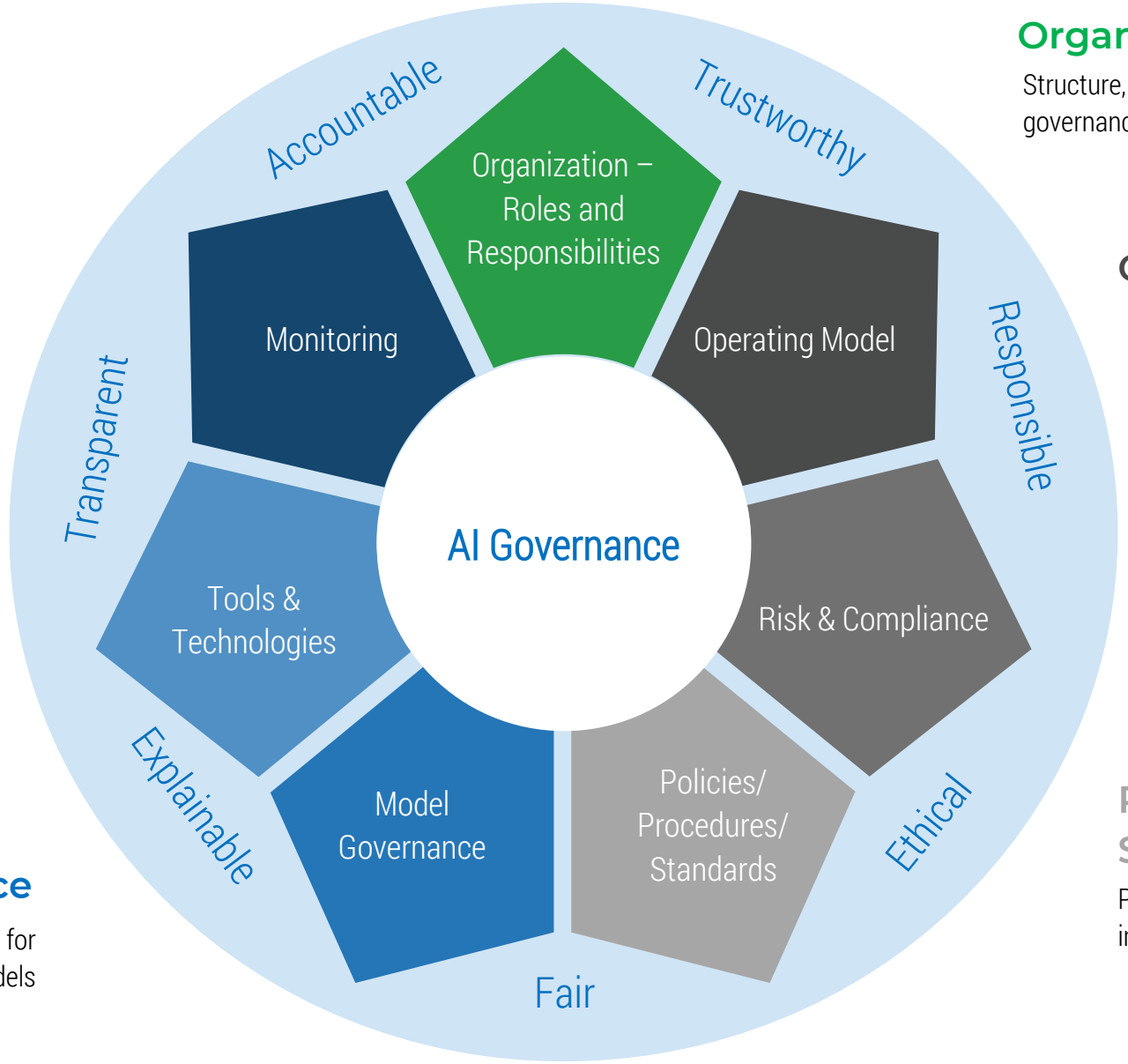
Transparency means communicating with and empowering users by sharing information internally and with external stakeholders, including beneficiaries and people impacted by the AI-powered product or service.

68% [of Canadians] are concerned they don't understand the technology well enough to know the risks.

77% say they are concerned about the risks AI poses to society

– TD, 2019

AI Governance Framework



Organization

Structure, roles, and responsibilities of the AI governance organization

Operating Model

How AI governance operates and works with other organizational structures to deliver value

Risk & Compliance

Alignment with corporate risk management and ensuring compliance with regulations and assessment frameworks

Policies/Procedures/ Standards

Policies and procedures to support implementation of AI governance

Model Governance

Ensures accountability and traceability for AI/ML models

Tools & Technologies

Tools and technologies to support AI governance framework implementation

Monitoring

Monitoring compliance and risk of AI/ML systems/models in production

Key components of AI governance

The AI Governance Framework and program will:

- Define accountability and responsibility for AI.
- Help to define the AI risk management framework.
- Support the ethical, transparent, and fair use of AI.
- Define a framework to support ML/AI model governance.

The AI Governance Framework helps organizations to govern, monitor, and adopt AI practices and systems.

The **key components** of the AI Governance Framework include:

- AI organization, AI ownership
- Operating model
- Policies/procedures/standards to support the AI governance program
- Risk and compliance
- Model governance
- Tools and technologies to support AI
- Monitoring of AI system in production

The AI Governance Framework should also define a set of metrics and key performance indicators (KPIs) that can be used to measure the success of the framework implementation.

AI ownership

Who owns an AI system or AI/ML model within an organization?

To answer this question, we need to understand:

- What business problem or business opportunity we are addressing with AI?
- Who within the organization will define requirements for an AI system?
- Who will define how AI will create business value? How will this value be measured?
- How will the output of the system or model be used by the organization? What business unit or a department will use the prediction generated by the model? Who will take ownership of the AI system and its capabilities when it is in production?
- What business processes will be impacted by AI?
- Who will be responsible for identifying what data can and cannot be used by AI?

The AI owner will be ultimately accountable for ensuring the AI is responsible, ethical, transparent, fair, and trustworthy.

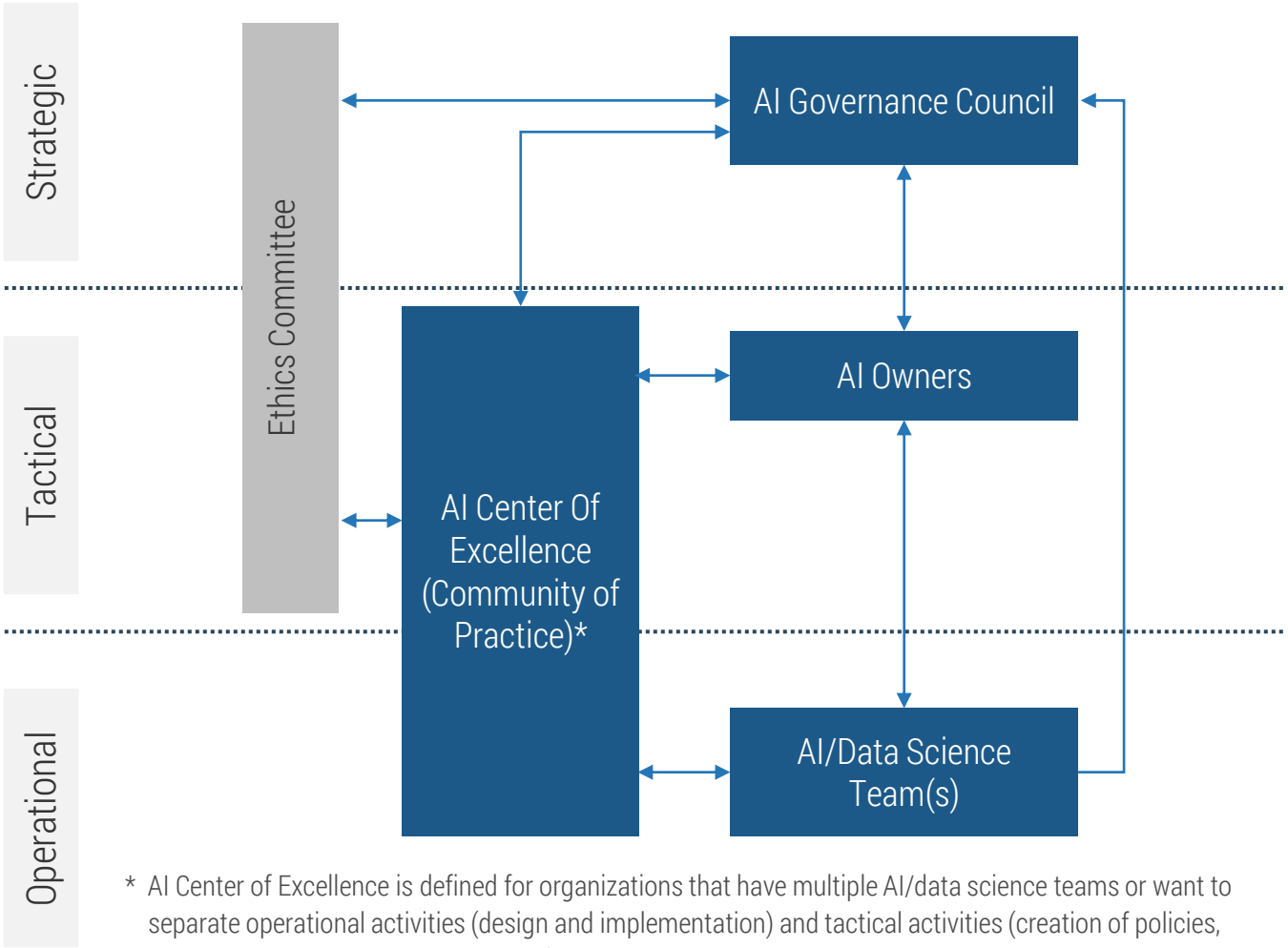
The AI owner has the highest interest not only in the outcome and the use of the AI but also in ensuring the potential risks are addressed, communicated, and monitored through AI lifecycle.

The AI owner is typically a senior business leader who:

- Is the head of a business unit or the head of a line of business that uses AI.
- Is positioned to accept accountability for the AI system or ML model.
- Understands the risks associated with the use of AI and works with the team to address the risks.
- Holds authority and influence to effect change, including across business processes and systems, needed to ensure ethical and responsible use of AI.

AI governance organization

There is no one-size-fits-all AI governance structure. Organizations need to identify roles and responsibilities at strategic, tactical, and operational levels; establish an AI governance council; and identify all groups that support AI initiatives. Maturity, the size of organization, and enterprise governance structure will influence AI governance structure.



- Alignment with organizational goals
- Enterprise risk assessment
- Alignment with ethical standards.
- Final decision authority for resolving any AI-related issues.

- Oversight of AI initiatives
- Development of AI/ML standards, policies, and procedures
- Development of AI architecture
- Alignment on tools and technologies used by AI/data science teams
- Alignment between AI/ML/data science teams across the organization

- Review data sets for compliance with risk, privacy, ethics, and legal requirements
- Design AI/ML solutions
- Monitor AI/ML solution performance and compliance in production

* AI Center of Excellence is defined for organizations that have multiple AI/data science teams or want to separate operational activities (design and implementation) and tactical activities (creation of policies, standards, and community of practice).

AI Governance Council – roles and responsibilities

The AI Governance Council acts as an ultimate approval and decision-making body.

- Align AI/ML goals and the scope of AI systems implementation with strategic business objectives/ initiatives.
- Ensure AI value and value drivers are understood.
- Define prioritization criteria for AI projects.
- Ensure the risk management framework is in place and the risk assessment is performed for AI implementations.
- Review data sets and approve the use of data and/or data attributes for AI/ML modeling.
- Review and approve (if required) the use of AI/ML models and solutions.
- Review and assess compliance of the proposed AI system or model with privacy and ethics regulations, external and internal guidelines, and policies.

AI Governance Council:

- AI Owners (Business Stakeholders)
- AI/Data Science Lead
- Data Governance Lead
- Privacy Lead
- Security Lead
- AI Center of Excellence Lead
- CDO and/or CIO

AI Center of Excellence – roles and responsibilities

The AI Center of Excellence acts as a central body for promoting AI/ML best practices, providing AI/ML/data science teams across the organization with support, helping with knowledge sharing, and driving standardization across all teams working on AI initiatives.

- Define and develop standards, policies, and procedures to promote guidelines and best practices in implementing responsible AI that complies with privacy and ethics regulations, external and internal guidelines, and policies.
- Build guardrails for the use of AI systems.
- Assist development teams in conducting initial risk and compliance assessment of datasets and models.
- Standardize AI architecture, tools, and technologies, including ML operations (MLOps) tools, across the organization.
- Define training opportunities, develop training plan, and conduct training.
- Collaborate with the AI/data science teams and AI governance council to resolve any issues or concerns related to either data or model use.
- Collaborate with Enterprise Architecture (EA) to ensure alignment with EA practices.
- Develop prototypes or proofs of concept and share the results with the community.
- Align with IT on the priorities and tool adoptions.

Note: AI Center of Excellence is defined for organizations that have multiple AI/data science teams or want to separate operational activities (design and implementation) and tactical activities (creation of policies, standards, and community of practice).

AI Center of Excellence:

- AI/Data Science Lead
- Data Scientists
- Data Engineers
- MLOps Specialists
- AI/ML Developers
- IT Stakeholders
- Data Governance Stakeholders (e.g. Data Stewards and Data Custodians)
- Product Managers

AI/data science team – roles and responsibilities

The organization might have multiple AI/data science teams embedded within business units, working on business domain–specific solutions

Members of AI/data science teams:

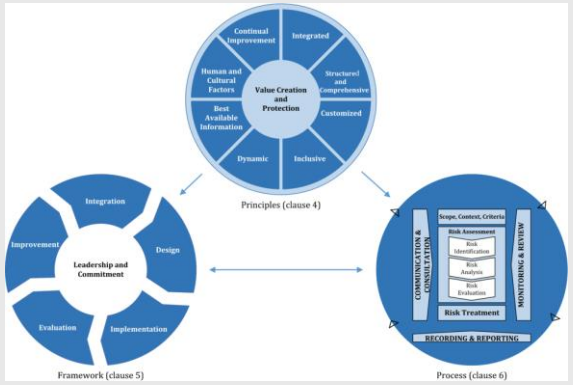
- Design, implement, and deploy the solution.
- Define and implement model governance.
- Align the use and management of data sets with data governance.
- Document model cards and data sheets for data sets.
- Conduct initial compliance assessment of the proposed AI system or model with privacy and ethics regulations, external and internal guidelines, and policies.
- Collaborate with the AI Center of Excellence and AI Governance Council to resolve any issues or concerns related to either data or model use.
- Implement access control for the AI/ML system/model.

AI/Data Science Team:

- AI/Data Science Lead
- Data Scientists
- Data Engineers
- Data Analysts
- MLOps Specialists
- AI/ML Developers
- Technical Leads
- Quality Leads
- Operations Team

Operating model – alignment with other governance structures

Corporate Governance
- Enterprise Risk Management Framework
- Ethics Board/Committee



ISO 31000:2018(en) Risk management

Align

AI Governance

Data Governance

Align

IT Governance
- IT Risk Council
- Change Advisory Board
- Tools and Technologies
- DevOps
- Software Development Lifecycle

Info-Tech Insight

Take advantage of other governance structures within the organization. Make sure there is a clear separation of roles and responsibilities.

AI governance – operating model

The AI governance operating model defines the design and operational implementation of a governance structure that delivers measurable business value to the organization while managing model lifecycle and risks and ensuring trustworthy and ethical AI.

- Design and implementation of the AI governance operating model is driven by:
  - The **scope of AI** initiatives.
  - The maturity and **structure of AI/data science practice.** Consider the number of your teams working on AI and data science initiatives, their geographical distribution, and the presence and level of involvement of citizen data scientists. All these factors will play an important role in how you structure your AI governance operating model.
  - Your organization's **enterprise risk management** framework. Some organizations might find it more effective to adapt parts of the existing risk management structures to adopt AI-related risk controls or integrate the AI risk management framework with the existing enterprise framework.
  - The **enterprise governance** structure and its components. Some organizations must align the AI Governance Council with the Enterprise Risk Committee or coordinate with the IT Risk Council.
- Alignment with the **IT Steering Committee** on processes and practices around tool implementation, including MLOps tools and platforms.
- AI Governance will have to coordinate agenda and processes with **Enterprise Ethics Committee/Board**.
- Regardless of the specifics of implementing the AI governance organization, it requires a home and an operating model to ensure its sustainability and evolution. The key is to determine what structure will work best in your organization, considering the maturity of the AI practice and other organizational groups.

Info-Tech Insight

The design of your AI governance operating model will depend on the organization's risk management framework, the structure of the AI/data science practice, and the scope of AI implementations.

AI governance – processes

Some of the necessary processes that will enable AI governance operations might include:

- Formal **data use approval process** that will define how the data sets and/or specific data attributes will be used in AI systems. This process should define what AI governance components and teams are involved in the process, steps of the review process, and what information should be gathered and reviewed to provide decisions. It also defines who is responsible for providing an approval and when the approval is required.
- Formal **model review and approval process** that will define the model lifecycle and how and when the model needs to be reviewed and approved by the AI Center of Excellence community of practice and/or AI Governance Council. Not all models will have the same review and approval process; it will depend on the model use, model complexity, and preliminary risk assessment.
- **Monitoring and oversight** process(es) that will define how the organization will make sure that the policies and standards are followed to address potential AI risks.
- Formal **peer review** process for scientific papers published by the AI/data science team.

Potential AI risks

Key potential risks of AI can come from data used by AI models (training data and data in production), model design, application of AI system, and lack of AI governance.

- **Poor data quality** and **inadequate data** used for model training:
  - Inadequate data can create biases in the AI system. Training data that lacks context or has inherent bias can lead to a biased model.
  - Data quality can impact model capabilities and performance.
- Model response to **changes to the input data in production**. The model learning in production is exposed to unexpected new data.
- Risks associated with **model design and the data used:**
  - Lack of **transparency** of the model, which leads to lack of trust in AI and affects AI adoption.
  - A **biased** model that does not comply with **ethics** policies and regulations.
- Risks of running the model in production:
  - Lack of **explainability of the model**, which is the ability to explain how the model makes predictions, what attributes were used by the model for prediction, and how these attributes contribute to the prediction. It is important to monitor model explainability in production for critical applications.
  - **Adversarial attacks***, including but not limited to adversarial input and model extraction:
    - **Adversarial input** is when the input data from external systems is manipulated so that a model makes a false prediction. Examples of adversarial input include a sticker on a stop sign that prevents a self-driving car from detecting the sign and stopping, or a spam email designed to resemble a normal email so the spam filter does not classify the email as spam.
    - **Model extraction** attacks allow an adversary to efficiently steal the model through prediction APIs by collecting data and training a substitute model using this data, hence replicating the functionality of a target model.

* Refer to MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems) for a landscape of threats to machine learning systems.

AI risk management framework

Organizations need to review existing risk control frameworks and update them to include AI risk management frameworks and processes.

- Most financial institutions follow the Three Lines of Defense Model, developed by the Institute of Internal Auditors.
- Some organizations implement the framework, principles, and processes of the ISO 31000 Risk Management Standard, developed by the International Organization for Standardization, to integrate risk management into the organization's activities and functions.
- Regardless of what enterprise risk management framework is implemented by the organization, AI governance needs to be aligned with the enterprise risk management framework and take advantage of its established processes and structure.
- AI governance needs to ensure that AI risk management implements sufficient oversight and effectively challenges the proposed use of the AI system, evaluates risks through the project lifecycle, and monitors production use of the AI system.
- An organization can start with creating an AI impact assessment methodology.
- The AI risk management framework should include an auditing process for third-party products. These can include any off-the-shelf AI-powered vendor products and prebuilt models.
- The risk management framework should include the definition of the risk categories to differentiate high-impact and high-risk AI models and systems from low-risk ones.

Define policies and procedures to support implementation of AI governance

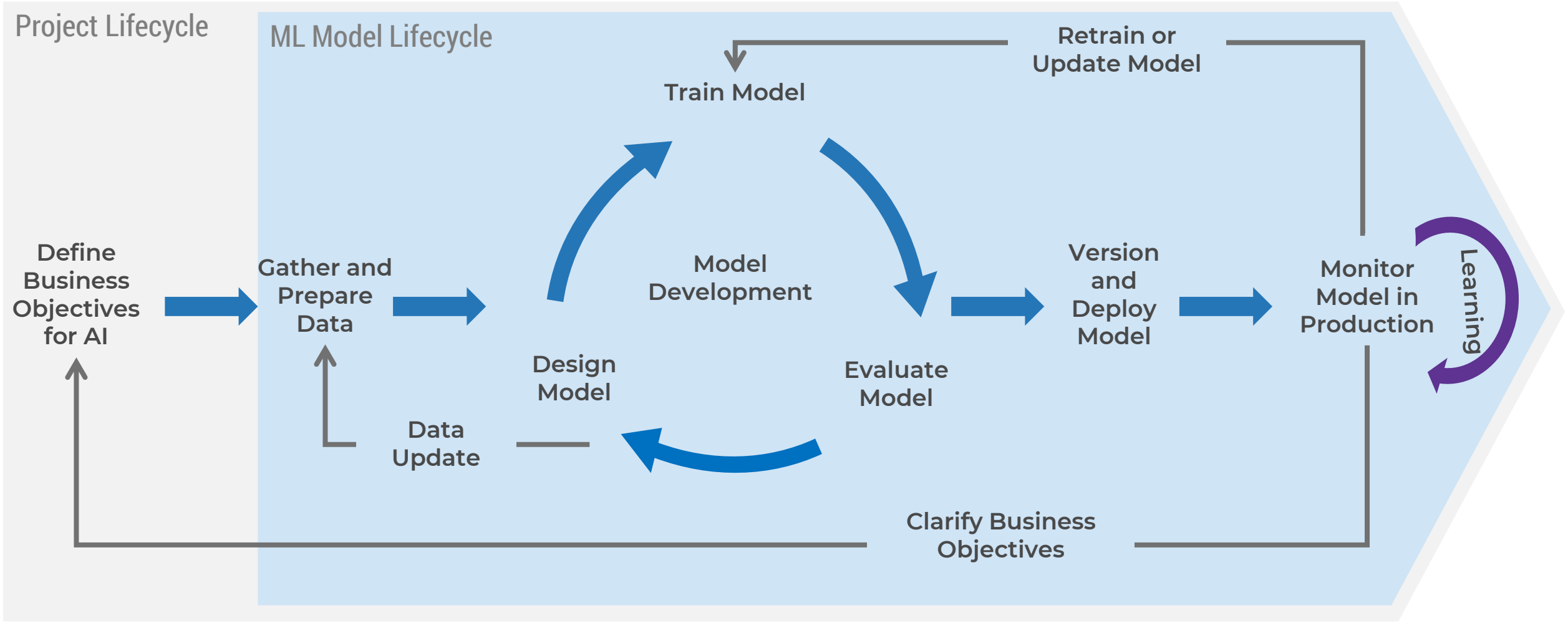
Create an umbrella AI Governance Policy first to serve as a general use policy that all other AI policies will agree with.

- The AI Governance Policy is created and approved by the AI Governance Council and establishes the AI governance organization, provides directions for AI risk management, and sets the foundation for all other AI policies.
- The AI Governance Policy should contain explicit definitions of AI terminology as well as AI frameworks and components of the AI governance organization.
- Review existing policies and standards, as they might cover many AI risk and compliance use cases. Update any existing policies to define AI-specific scenarios.
- Examples of the policies might include:
  - Use of Data for AI/ML and Data Science Projects
  - AI Impact Assessment Policy

Info-Tech Insight

When defining your AI policies, processes, and standards, ensure they are relevant, serve a purpose, and/or support the risk management framework and the use of AI in the organization.

AI/Machine Learning Model Lifecycle



Info-Tech Insight

Model governance should cover the entire machine learning model lifecycle.

Model governance

Model governance helps bring greater accountability and traceability for AI/ML models through the model lifecycle.

Data

- What are the data sets (assets) that will be used for AI/ML? What business problem will the planned AI system solve and how will this data be used to solve the problem and achieve business goals?
- Are there any limitations on the use of data sets or some of the attributes?
- Who are the related data owners? Who will the AI team have to work with to answer any questions they have about data?
- How and where was the data obtained from? Does the team have the rights to use that data? How can data access and use problems be resolved?
- Are there legislations (e.g. GDPR), policies, or regulations that guide or dictate how that data can be used?
- What are the risks associated with the use of the data?

Model

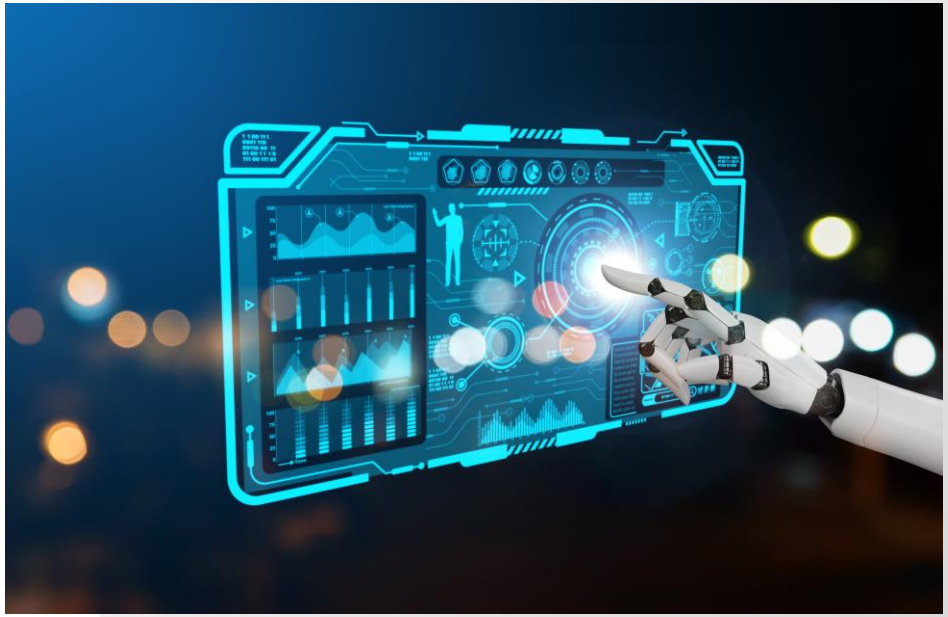
- How was the model created, tested, trained, deployed, and evaluated?
- Define how the model performance, accuracy, and explainability are assessed and monitored in production.
- How is the model metadata captured through the model lifecycle?
- Define the reporting on the model metrics.
- What are the process and the tools used to monitor model risk and compliance in production?
- What is the process of updating and/or retraining the model that is currently in production?
- How is model versioning implemented?

Tools and technologies to support AI governance

There are a variety of different tools, technologies, and platforms that can be used to support AI governance. The key is to use the right tools and platforms and avoid tool proliferation.

The following capabilities must exist in tools and technologies that support AI governance:

- Data analysis and data visualization, including tools for creating reports and dashboards
- Data cataloging
- Model management, including capturing model metadata and model versioning
- MLOps technologies, tools, and platforms that allow you to monitor model performance in production, visualize metrics, and capture model performance data for further (or near real-time) analysis
- Role-based access to the models and datasets



Monitoring AI systems in production

Monitoring AI systems in production for explainability is critical

- In production, the model can be monitored to detect **data drift** but also to maintain **data compliance** to ensure the data and model predictions are still fair, unbiased, and explainable.
- AI systems learn based on the new data that is received while the model is in production, and each new data iteration dynamically optimizes the AI system. For example, a chatbot learns and updates its response with each user interaction. AI governance needs to ensure that the critical models that learn in production should have additional safety protocols and continuous monitoring with thresholds and notifications clearly defined and implemented.
- The production environment should be monitored for adversarial robustness to ensure that the model has not been compromised by adversarial attacks.
- Model monitoring usually would include:
    - Controlling model performances over time.
    - Detecting data drift.
    - Automatically retraining models in case of performance drift.
    - Logging and auditing queries sent to the models.

"For machine learning models, ***data drift*** is the change in model input data that leads to model performance degradation. Monitoring data drift helps detect these model performance issues. Data drift is one of the top reasons model accuracy degrades over time."

– Azure Machine Learning Documentation

Research Contributors and Experts

Irina Sedenko

Research Director
Info-Tech

Ellie D. Norris

Associate Director
Merck

Bibliography

"A Global AI in Financial Services Survey." *Cambridge Centre for Alternative Finance (CCAF) at the University of Cambridge Judge Business School and the World Economic Forum*, Jan. 2020. Web.

"Algorithmic Impact Assessment tool." Government of Canada, 1 Nov. 2022. Web.

Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS). "Artificial Intelligence Risk & Governance." *The Wharton School, The University of Pennsylvania,* n.d. Accessed August 2022.

"Attacking Machine Learning with Adversarial Examples." *OpenAI*, 24 Feb. 2017. Accessed August 2022.

Bill C-27: Artificial Intelligence and Data Act (AIDA). 1st Reading, 44th Parl., 1st sess. Library of Parliament, 16 June, 2022. *Parliament of Canada.* Accessed August 2022.

Buchanan, Will, et al. "Detect data drift (preview) on datasets." *Azure Machine Learning Documentation, Microsoft*, 2022. Web.

"Directive on Automated Decision-Making." *Government of Canada*, 1 May 2021. Web.

"Ethics, Transparency and Accountability Framework for Automated Decision-Making. Guidance." *Office for Artificial Intelligence, UK Government*, 13 May 2021. Web.

European Commission. "Artificial Intelligence Act." *The AI Act,* 2021. Web.

Floridi, Luciano, and Mariarosaria Taddeo. "What is data ethics?" The Royal Society Publishing, vol. 374, no. 2083, 28 Dec. 2016. Accessed July 2022.

G20 Trade Ministers and Digital Economy Ministers. "G20 Ministerial Statement on Trade and Digital Economy." *Ministry of Foreign Affairs of Japan,* 2019. Accessed August 2022.

Giardino, Elisa. "The mirage of a global framework for AI governance." *Medium,* 7 Nov. 2020 Accessed July 2022.

"Glossary of Terms." *Open Data Institute*, n.d. Accessed July 2022.

"Invitation to Pilot AI Verify: AI Governance Testing Framework & Toolkit." Infocomm Media Development Authority (IMDA), 25 May 2022. Accessed August 2022. "Three Common Problems With The Three Lines Of Defense Framework." Forbes, 6 July 2020. Accessed August 2022.

"ISO 31000 Risk Management." ISO, n.d. Web.

"ISO 31000:2018(en) Risk management — Guideline." ISO, 2018. Web.

Jagielski, Matthew, and Nicolas Papernot. "In Model Extraction, Don't Just Ask 'How?': Ask 'Why?'." *cleverhans-blog,* 21 May 2020. Web.

Liu, Shengyi. "Model Extraction Attack and Defense on Deep Generative Models." *Journal of Physics: Conference Series*, 2189, 012024, 5 Dec. 2021. Web.

MITRE ATLAS™, The MITRE Corporation. https://atlas.mitre.org/

Molnar, Christoph. *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*. 12 July 2022. Web.

Bibliography

"Moving Beyond Principles. Addressing AI Operational Challenges." *Canadian Regulatory Technology Association (CRTA)*, Feb. 2022.

National Artificial Intelligence Initiative, *US Government*. https://www.ai.gov/. Accessed August 2022.

OECD.AI , *OECD*. https://oecd.ai. Accessed 9 July 2022.

"Pan-Canadian AI Strategy." *CIFAR*, n.d. Web.

"Recommendation of the Council on Artificial Intelligence." *OECD Legal Instruments,* 2019. Accessed August 2022.

"Responsible AI in Financial Services – Enhancing Trust While Adopting New Capabilities." *TD*, 2019. Web.

"Safeguarding AI Use Through Human-Centric Design." Cana*dian Regulatory Technology Association (CRTA)*, May 2020.

"The IIA's Three Lines Model. An update of the Three Lines of Defense." *The Institute of Internal Auditors (IIA),* July 2020. Web.

"William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Division E—National Artificial Intelligence Initiative Act of 2020." *Congress.gov,* 3 Dec. 2020. Accessed August 2022.

Zhu, Wei. "4 steps to developing responsible AI." *World Economic Forum*, 20 June 2019. Web.


INFO~TECH
RESEARCH GROUP