Redbooks
ibm.com/redbooks

Simplify Your AI Journey: Ensuring Trustworthy AI with IBM watsonx.governance

Deepak Rangarao
Upasana Bhattacharya
Savitha Chinnappareddy PhD
Larry Coyne
David Cruz
Shuvanker Ghosh
Prem Piyush Goyal
Vasfi Gucer
Amna Jamal PhD
Warren Lucas
Karen Medhat
Bob Reno
Mohit Sharma
Mark Simmonds
Jasmeet Singh
Martijn Wiertz

Artificial Intelligence

Data and AI

IBM®
Redbooks

IBM Redbooks

Ensuring Trustworthy AI with IBM watsonx.governance

January 2025

SG24-8573-00

Contents

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at https://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

IBM®
IBM Cloud®
IBM Research®
IBM Watson®
OpenPages®
Redbooks®
Redbooks (logo) ®

The following terms are trademarks of other companies:

Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenShift, Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

RStudio, and the RStudio logo are registered trademarks of RStudio, Inc.

Other company, product, or service names may be trademarks or service marks of others.

Foreword

This trilogy of IBM® Redbooks® publications positions and explains IBM watsonx, the IBM strategic AI and Data platform. Each book focuses on one of the three main components of the watsonx platform:

- **IBM watsonx.ai:** A next-generation enterprise studio for AI developers to train, validate, tune, and deploy both traditional ML and new generative AI capabilities powered by foundation models.
- **IBM watsonx.data:** A fit-for-purpose data store built on an open-lakehouse architecture, optimized for different and governed data and AI workloads.
- **IBM watsonx.governance:** A set of AI governance capabilities enabling trusted AI workflows, helping organizations implement and comply with ever-changing industry and government regulations.

Organizations have long recognized the value that IBM Redbooks provide in guiding them with best practices, frameworks, clear explanations, and use cases as part of their solution evaluations and implementations.

This trilogy of books was only possible due to the close collaboration involving many skilled and talented authors that were selected from our IBM global technical sales, development, Expert Labs, Client Success Management, and consulting services organizations, using their diverse skills, experiences, and technical knowledge across the watsonx platform.

I would like to thank the authors, contributors, reviewers, and the IBM Redbooks team for their dedication, time, and effort in making these publications a valuable asset that organizations can use as part of their journey to AI.

I also want to thank Mark Simmonds and Deepak Rangarao for taking the lead in shaping this request into yet another successful IBM Redbooks project.

It is my sincere hope that you enjoy this watsonx trilogy as much as the team who wrote and contributed to them.

Steve Astorino, IBM General Manager - Development, Data, AI and Sustainability.

Preface

IBM® watsonx™ is the IBM strategic AI and Data platform. This book focuses on watsonx.governance, a key component of the platform.

IBM watsonx.governance offers a comprehensive solution for governing data and AI workloads within a secure and scalable environment. Built on an open architecture, it empowers organizations to manage data access, compliance, and security across hybrid multi-cloud deployments. IBM watsonx.governance simplifies data governance with built-in automation tools and integrates seamlessly with existing databases and tools, streamlining workflows and enhancing user experience

This IBM Redbooks publication provides a broad understanding of watsonx.governance concepts and architecture, and the services that are available in the product. In addition, several common use cases and scenarios are included that should help you better understand the capabilities of this product.

This publication is for watsonx customers who seek best practices and real-world examples of how to best implement their solutions while optimizing the value of their existing and future technology, AI, data, and skills investments.

> **Note:** Other books in this series are:
>
> - *Simplify Your AI Journey: Unleashing the Power of AI with IBM watsonx.ai*, SG24-8574
> - *Simplify Your AI Journey: Hybrid, Open Data Lakehouse with IBM watsonx.data,* SG24-8570

Authors

This book was produced by a team of specialists from around the world working with the IBM Redbooks, Tucson Center.

Deepak Rangarao is an IBM Distinguished Engineer and CTO responsible for Technical Sales-Cloud Paks. Currently, he leads the technical sales team to help organizations modernize their technology landscape with IBM Cloud® Paks. He has broad cross-industry experience in the data warehousing and analytics space, building analytic applications at large organizations and technical pre-sales with start-ups and large enterprise software vendors. Deepak has co-authored several books on topics, such as OLAP analytics, change data capture, data warehousing, and object storage and is a regular speaker at technical conferences. He is a certified technical specialist in Red Hat OpenShift, Apache Spark, Microsoft SQL Server, and web development technologies.

Upasana Bhattacharya is a Senior Product Manager for watsonx.governance, based in Markham, Canada. In this role she defines the product vision, guides its development, collaborating with cross-functional teams. In her previous role she was a Product Manager for Data and AI. Upasana holds a Bachelor of Arts in Economics and Foreign Affairs from the University of Virginia and an MBA from the McCombs School of Business at the University of Texas.

Savitha Chinnappreddy, PhD is a Senior AI Engineering Manager at IBM with over 17 years of experience in AI and Data Analytics. She holds a PhD in AI and Data Analytics and is currently pursuing a post-doctorate focused on Human & AI Collaboration: Governance strategies for trustworthy AI & Safe AI systems. She has extensive Experience in managing and scaling large AI and Data Science teams, she has worked closely with architecture and infrastructure teams to establish compliant pipelines for AI and analytics, delivering impactful solutions to global customers. With 11 publications in esteemed journals and conferences, as well as holding a patent, she is also an active guest speaker and participant in faculty development programs, committed to sharing her knowledge and inspiring the next generation of AI professionals.

Larry Coyne is a Project Leader at the IBM International Technical Support Organization, Tucson, Arizona, center. He has over 35 years of IBM experience, with 23 years in IBM storage software management. He holds degrees in Software Engineering from the University of Texas at El Paso and Project Management from George Washington University. His areas of expertise include client relationship management, quality assurance, development management, and support management for IBM storage management software.

David Cruz is a Data Scientist and AI Engineer working under IBM's Client Engineering team. In this role, David has been dedicated to the Federal Market where he works to implement a wide range of AI solutions for federal clients. In his prior role, he worked under the Data Science Elite team where he gained skills with IBM platforms for Governance, namely IBM OpenScale, and this has translated into a growing skill set with watsonx governance. He is constantly working to implement the cutting edge of AI and AI Governance technology, and has written various blog posts on topics ranging from Unsupervised Learning techniques, to RAG how-to guides for beginners.

Shuvanker Ghosh is a certified Executive Architect and Worldwide Platform Leader for Data and AI in Worldwide Solution Architecture in IBM Technology Expert Labs. With 18 years of experience at IBM, he serves as a trusted advisor to clients, offering thought leadership on IBM's Data and AI portfolio. He guides organizations in their responsible AI journey, helping them adopt best practices. His current focus is on defining solution blueprints and architectural patterns that assist clients in addressing their business challenges through responsible and trustworthy AI solutions. He possesses extensive expertise in the IBM Data and AI portfolio, including the watsonx platform and Cloud Pak for Data. Shuvanker has successfully led and delivered complex programs that involve multiple teams, providing technical management, architecture, technology thought leadership, and software development methodologies and processes. His experience spans various industries, including retail, finance, insurance, healthcare, telecommunications, and government

Prem Piyush Goyal is a problem solver with extensive experience in developing cutting-edge technologies at IBM. Specializing in full-stack development, cloud-based microservices, and AI solutions, he has worked on high-impact projects like IBM Watson® Data Platform and IBM Watson OpenScale. His expertise spans Python, JavaScript, React, Kubernetes, and AI-driven solutions like Explainable AI and Concept Drift Detection. Passionate about building transparent and scalable AI, he continually enhances user experience and optimizes performance for enterprise applications. His innovative mindset and problem-solving abilities help drive trust and transparency in AI systems.

Vasfi Gucer leads projects for the IBM Redbooks team, leveraging his 20+ years of experience in systems management, networking, and software. A prolific writer and global IBM instructor, his focus has shifted to storage and cloud computing in the past eight years. Vasfi holds multiple certifications, including IBM Certified Senior IT Specialist, PMP, ITIL V2 Manager, and ITIL V3 Expert.

Amna Jamal PhD is a seasoned Data and AI Subject Matter Expert (SME) at IBM, boasting over 8 years of expertise in data management and data science. With a Ph.D. in Engineering from the National University of Singapore, she brings a wealth of knowledge and experience to the field, driving innovation and excellence in the intersection of data and artificial intelligence.

Warren Lucas is a member of IBM Expert Labs. Prior to his time at IBM, Warren has spent nearly a decade working in Regulatory Compliance, Operational Risk, and Model Risk Governance supporting a number of Fortune 50 companies in their efforts to redesign and implement internal governance processes. As a Solution Architect, Warren has specialized in Governance Console (IBM OpenPages®) for over seven years, where he has personally performed development, design, advisory, and configuration within the platform. Warren has a current patent submission for a novel approach in governance and confidence assessments in large language models (LLMs); he holds a degree in Quantitative Economics.

Karen Medhat is a Customer Success Manager Architect in the UK and the youngest IBM Certified Thought Leader Level 3 Technical Specialist. She is the Chair of the IBM Technical Consultancy Group and an IBM Academy of technology member. She holds an MSc degree with honors in Engineering in AI and Wireless Sensor Networks from the Faculty of Engineering, Cairo University, and a BSc degree with honors in Engineering from the same faculty. She co-creates curriculum and exams for different IBM professional certificates. She also created and co-created courses for IBM Skills Academy in various areas of IBM technologies. She serves on the review board of international conferences and journals in AI and wireless communication. She also is an IBM Inventor and experienced in creating applications architecture and leading teams of different scales to deliver customers' projects successfully. She frequently mentors IT professionals to help them define their career goals, learn new technical skills, or acquire professional certifications. She has authored publications on Cloud, IoT, AI, wireless networks, microservices architecture, and Blockchain.

Bob Reno is a Principal Technical Sales Specialist with over 30 years of experience in Data Warehousing, Analytics, and AI. As a member of the IBM World Wide Data and AI Technical Sales team, Bob is a watsonx.governance leader working with customers to enable their organizations to embrace responsible AI. Bob has contributed to the creation of several IBM Certification Tests and written several workshops in the watsonx, Cloud Pak for Data and Data Warehousing space to enable customers and the IBM Technical Community. Prior to joining IBM, Bob has held roles as a Developer, Technical Architect, and Director of Data Warehousing and Analytics.

Mohit Sharma is an AI engineering lead on the Client Engineering watsonx team in Bangalore, India. Prior to this, Mohit was associated with IBM consulting, and worked on client production projects involving classical ML and deep learning. Mohit has around 14 years of experience in AI, and worked at Hewlett Packard, Wipro (where he conceptualized the Holmes AI platform) and Accenture before joining IBM in 2018. An AI practitioner having experience in design and development of AI-based solutions using both open-source and commercial technologies, Mohit is interested in both data and the science behind it. He has 4 published patents to his credit, and has filed his first patent at IBM.

Mark Simmonds is a Program Director in IBM Data and AI. He writes extensively on AI, data science, and data fabric, and holds multiple author recognition awards. He previously worked as an IT architect leading complex infrastructure design and corporate technical architecture projects. He is a member of the British Computer Society, holds a Bachelor’s Degree in Computer Science, is a published author, and a prolific public speaker.

Jasmeet Singh is a watsonx Client Success Manager with 17 years of experience in IT and 8 years of experience in watsonx Technologies with IBM Technology Expert Labs team with 4 years focused as watsonx.governance and AI Governance SME. He has delivered high-quality implementations of AI Governance with big named IBM clients. Jasmeet holds 5 patents in AI field and holds an MS degree in Cybersecurity from NC A&T State University.

Martijn Wiertz is the Technical Sales Leader for IBM watsonx.governance in the EMEA region. In this role, he combines his technical, analytical and industry knowledge to help clients understand and validate the unique value that the solution can bring to help them enact responsible AI. He has more than 25 years of experience in the field of advanced analytics, experiencing all major developments in the evolution of the industry first hand. Prior to his current role, Martijn was the global technical sales lead for IBM's solution to combat financial crimes and he was Director, Enterprise Solutions at SPSS Inc. when that company was acquired by IBM.

Thanks to the following people for their contributions to this project:

- **Steve Astorino**, IBM General Manager - Development, Data, AI and Sustainability

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  https://www.linkedin.com/groups/2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/subscribe

- Stay current on recent Redbooks publications with RSS Feeds:

  https://www.redbooks.ibm.com/rss.html

1

Challenges and opportunities in AI governance

In 2024, the topic of governance of Artificial Intelligence (AI) has grown enormously in both attention and in adoption. This chapter provides a definition of AI governance and describes the key challenges and opportunities it presents.

This chapter has the following sections:

1.1 What is AI governance?

There is not one formal definition of AI governance that is generally accepted. Many organizations have created a definition that focuses on specific elements of the overall AI governance picture. The one that, in the opinion of the author, captures a broad spectrum succinctly best is this one from the Data & AI Alliance:

> *"A system of rules, practices, processes and tools that help an organization use AI in alignment with its values and strategies, address compliance requirements and drive trustworthy performance.[1]"*

In other words, it is about two things. First, it is about the rules that an organization sets for themselves to make sure their use of AI is profitable, compliant, secure and fair. And second, it is about the methods that an organization applies to ensure and document that those rules are followed using tools as the enabler.

Note: Some clients also refer to AI governance as responsible governance. While AI governance focuses on the broader framework, responsible governance specifically highlights the ethical and societal implications.

AI governance comprises a set of activities that run in parallel to the operational process of creating, deploying and maintaining AI assets. Table 1-1 contrasts the typical activities related to different aspects of an AI solution. Many of the activities in the *AI governance* column will be discussed in more detail in the following chapters of this Redbooks publication.

Table 1-1 Contrast between typical Governance and Operations activities

Aspect of an AI solution	Typical AI governance activities	Typical AI operations activities
Foundation models	Review and approve new foundation models before they're applied in your use cases.	Acquire and host foundation models so they can be used when developing AI solutions.
Use cases	Assess and review new use cases before starting development. List the expected risk mitigation and compliance measures that the technical and other stakeholders need to apply. Determine the level and extent of governance based on the risk profile of the use case.	Create an initial solution design and architecture to support assessment and review.
AI asset development	Document the technical characteristics and development process of the AI assets. Review the developed AI assets for adequate risk mitigation and regulatory compliance. Approve for deployment.	Create and evaluate the AI assets that are needed to deliver a use case.

[1] Data & Trust Alliance, IBM - The urgency of AI governance, 2023

Aspect of an AI solution	Typical AI governance activities	Typical AI operations activities
AI asset deployment	Document the technical characteristics and development process of the AI deployments. Review the deployed assets for adequate risk mitigation and compliance.	Deploy the AI assets to enable their day-to-day use.
AI asset post-deployment	Manage issues and incidents according to compliance and risk mitigation plans. Review and approve requests for changes to AI assets, including any changes to the required risk mitigation and compliance measures. Review and approve the decommissioning of AI assets. Manage periodic attestations about use cases and AI assets.	Setup automated monitoring of deployed AI assets. Manage the AI assets from a technical perspective. Create and evaluate new versions of AI assets as needed.
AI embedded in enterprise applications	Review and approve new embedded AI capabilities before they're applied in your use cases.	<not typically involved since these capabilities come to the organization prebuilt>
Compliance with legal obligations across use cases	Gather evidence to demonstrate compliance with obligations that go across individual use cases (for example: “AI literacy” in the EU AI Act)	<not typically involved>
Regulator interactions	Manage inbound (for example: a request for information) interactions with regulators. Manage outbound (for example: registering high-risk AI use cases) interactions with regulators.	<not typically involved unless additional technical information is required for a specific AI solution.>
Regulatory change	Proactively assess how regulatory proposals related to AI might impact the organization. Review regulatory changes and determine with use cases are affected. Manage the activities to bring use cases and AI assets into compliance with the changed regulation.	<not typically involved unless the regulatory change requires additional technical work.>
Risk control evaluations	Assess effectiveness of AI risk controls.	<not typically involved>
Implement AI governance policies	Create and maintain compliance libraries, plans and assessment templates. Create and maintain risk/control libraries, plans and assessment templates.	<not typically involved>

Like any business practice, AI governance can benefit from applying the people, process and technology (PPT) framework. When applied to AI governance, this framework helps organizations ensure that their AI systems are aligned with their strategic goals, are developed and deployed responsibly, and deliver value to stakeholders. Here is a breakdown of how the PPT framework might apply to AI governance[2]:

- People: The first element of the PPT framework focuses on the people involved in the AI governance lifecycle, this includes technical experts, legal advisors, compliance officers,

[2] Based on IBM watsonx.ai chat interface, using the IBM Granite 13B Chat v2 model

business leaders, ethicists, and other stakeholders. In the context of AI governance, people play a crucial role in ensuring that AI systems are designed and used ethically, transparently, and responsibly. This involves:

  - Building diverse and inclusive teams
  - Providing adequate training and support
  - Fostering a culture of ethical AI practices

- Process: The second element of the PPT framework emphasizes the importance of well-defined and repeatable processes for the governance of AI systems. In the context of AI governance, processes help organizations manage risks associated with AI, ensure compliance with regulations, and maintain the quality and reliability of AI solutions. This includes:
  - Defining AI principles and policies
  - Establishing clear governance structures
  - Defining roles and responsibilities
  - Defining decision-making frameworks
  - Implementing robust change management processes
- Technology: The third element of the PPT framework focuses on the technology infrastructure and tools used to enact the principles, structures and processes in day-to-day work. In the context of AI governance, AI governance platforms provide a centralized and integrated view of AI systems, enabling organizations to govern the lifecycle of AI solutions, manage AI risks and ensure compliance. These platforms typically include features such as:
  - An inventory of use cases and AI assets
  - Automated workflows
  - Risk and legal assessments
  - Compliance plans
  - Issue management
  - Reporting and dashboarding

1.2 Governance as a key enabler for realizing AI value

AI governance serves as a critical enabler for organizations to unlock the full value of AI in a responsible and trustworthy manner. As organizations adopt AI to achieve ambitious goals, governance provides the structure needed to scale these efforts responsibly and effectively. Executives are pushing for the adoption of AI in enterprise contexts, which differs significantly from consumer-focused applications.

Scaling AI to meet these objectives is only possible when governance frameworks are firmly in place. Without consistent governance, organizations risk operational inefficiencies, compliance failures, brand reputation, and ethical concerns that can hinder AI adoption.

This section addresses three common concerns that people have about AI governance as an enabler of AI value.

1.2.1 Concern 1: Governance is a brake on AI

In some ways this is a true statement: governance does set boundaries in the application of AI and then holds people to that.

But compare it to a car: what would happen if your car doesn't have brakes? How would you even make it out of your street without the precise control to stop at the first intersection? How would you drive on a highway if you could not make an emergency stop or reliably take an exit? How would you park your car? Without brakes you would actually never be able to drive fast.

We have developed a governance system to make sure millions of people can get in cars, trucks and buses each day and get to their destination safely. This includes traffic laws, driver's licenses, speed limits, road signs, technical safety devices and more.

Extending that traffic metaphor to AI, AI governance helps you:

- Educate your employees about the rules of the road so they don't need a huge level of oversight every time they get in their car.
- Train your employees how to drive their AI safely (for example through an Ethics by Design method)
- Determine when it is safe to go fast (typically large numbers of AI use cases that an organization agrees are low- to no-risk).
- Identify those use cases where extra care is required, and clearly lay out what the obligations are.
- Implement guardrails in the extra dangerous hairpin turns to catch AI mishaps even if the driver is momentarily distracted.
- Maintain your brakes by adjusting your policies and procedures as AI continues to evolve.

Like brakes, AI governance, when applied properly, allows an organization to go fast most of the time.

1.2.2 Concern 2: Governance does not scale

That is absolutely true in the sense that *manual and ad-hoc* governance doesn't scale.

Through the people, process, and technology approach referred to earlier, organizations can make their AI governance more systematic and use technology to automate and scale.

IBM takes this approach in our own AI governance, and we now create 100s of new and governed AI use cases per quarter. Governance can absolutely scale to the level of a complex global company with large AI ambitions.

Software can help scale AI governance through:

- A *single enterprise inventory* of AI use cases and assets that all parties can use for their specific purposes. Note that this does not mean that all AI development needs to be done on a single platform.
- *Discovery* of any AI that is already in use but not yet registered (shadow AI).
- *Integration with systems of record* - depending on the type of business you are in, you will already have an inventory of your IT systems, a product catalog and other systems of record. Use technology to leverage the information you already capture there so you do not have to redo that for governance purposes.

- Automated *reports and dashboards* for specific roles and users.
- *Self-service methods* to business owners to register the relevant information about their use cases. Such a data gathering can be maintained by and combine the needs of specialists (legal, ethics, security, etc.).
- Automated *assessment* of the risk mitigation and compliance requirements for a use case.
- Automated *workflows* to involve specialists, facilitate reviews, resolve issues and manage escalation points.
- A standardized method for technical teams to *capture metadata* about the AI that they are developing.
- Standardized *monitoring* capabilities that are mapped to the legal obligations and corporate risk.
- Integration with *AI platforms and processes* to avoid duplicate work for the technical teams.
- Integration with any *adjacent tooling* where governance requirements might be fulfilled (for example: security monitoring of AI systems).

1.2.3 Concern 3: Governance does not contribute to value generation

Governance is often approached, and justified, as a way to avoid losses such as fines, damage to brand reputation or negative operational impact.

While this *loss avoidance* can be of substantial value to an organization, authors from the Notre Dame - IBM Tech Ethics Lab make the case for a holistic ROI framework that also incorporates a *value generation* perspective[3] that includes amongst others:

- Building unique organizational capabilities to take advantage of market opportunities more quickly.
- Ability to attract and retain talent through the organization's reputation.
- A sincere commitment to values-based leadership.

1.3 Challenges with governance of enterprise AI

With the rise in popularity of ChatGPT[4] and other consumer AI tools, organizations have been looking for ways to apply these capabilities also to their business environment. Like other technologies, transitioning from customer to business usage comes with specific challenges. This chapter describes the main challenges as they relate to the governance of an organization's AI initiatives.

1.3.1 Generative AI has changed the governance game

While business teams are smitten by the potential for business improvement, and the technical specialists are enamored by the latest technological breakthroughs, generative AI is different (enough) from earlier forms of AI to approach them thoughtfully.

These differences impact the governance of the AI in the following ways:

[3] On the ROI of AI Ethics and Governance Investments: From Loss Aversion to Value Generation

[4] OpenAIs consumer chatbot

- Generative AI enables *use cases in new areas* of an organization increasing the complexity of governance - more use cases to review, more (types of) users to educate, more room for accidental errors.
- Generative AI carries *amplified and new risks* such as bias in generated content, intellectual property concerns, hallucination, energy consumption and misinformation.
- Generative AI has a *different supply chain* - most organizations do not train their own large language models, but leverage pre-trained models from commercial vendors and the open-source community. These external models need to be vetted before being applied in use cases. See Chapter 4, “Onboarding a new foundation model” on page 41 for more on this topic.
- Generative AI introduces *different assets to govern* - organizations now routinely work with prompt templates, foundation models (FMs), and other AI assets. These assets come with different metadata and different ways to measure their performance and robustness. Chapter 6, “Governing the end-to-end lifecycle of an AI asset” on page 61 for more on this topic.
- Generative AI is *evolving rapidly*. ChatGPT broke through to the mainstream in 2023, in 2024 many organizations turned to retrieval-augmented generation (RAG) solutions and AI assistants in various forms, and 2025 looks to be the year of agentic AI. Each of these developments also impacts how AI is governed. To help keep up, look for AI governance partners with first-hand experience in delivering these types of cutting-edge AI solutions.

Having said all that, generative AI is not replacing machine learning (ML) - traditional machine learning remains a core technology in businesses and continues to deliver value. Consistent governance is essential for both ML and generative AI to address overlapping and unique challenges effectively.

1.3.2 Bring together diverse stakeholder perspectives

AI governance is not just a matter for the technical teams. It is an organizational team sport that must be informed by a wide range of perspectives to address the diverse challenges associated with AI systems.

Figure 1-1 shows different perspectives typically involved in the AI-related “rules” referred to in 1.1, “What is AI governance?” on page 2.

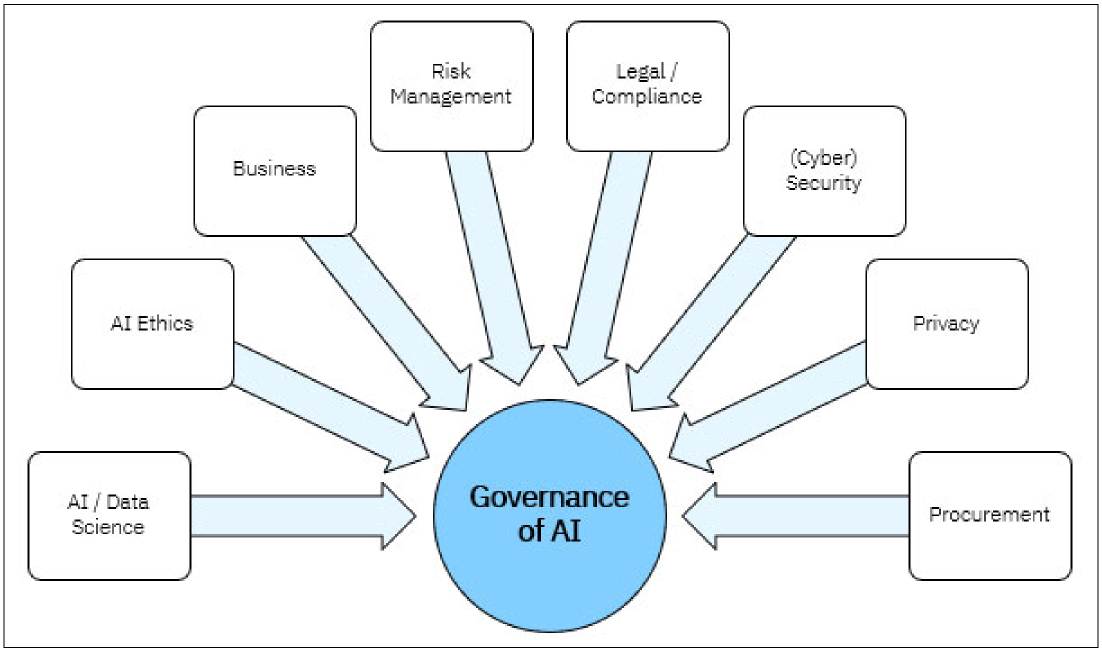


Figure 1-1 Comprehensive AI governance involves different stakeholder perspectives

Even if an organization does not have a dedicated department for each of these, the more these perspectives are represented in an AI governance framework, the more effective it will be. Table 1-2 gives examples of which of the AI governance activities from in paragraph 1.1 are commonly informed by which perspectives. Organizations will need to create their own version of such a mapping and define the roles and responsibilities of all involved parties.

Table 1-2 Mapping stakeholder perspectives to AI governance activities

Governance activity	**Data Science / AI**	**AI Ethics**	**Business unit**	**Legal / compliance**	**Risk management**	**(Cyber) Security**	**Privacy**	**Procurement**	**Other**
Assess and approve foundation models	X	X		X	X	X	X	X	
Assess and approve use cases	X	X	X	X	X	X	X		
Govern the development of AI assets	X	X	X	X	X	X	X		
Govern the deployment of AI assets	X			X	X	X	X		
Govern the post-deployment of AI assets	X		X	X	X	X			
Compliance with legal obligations across use cases (such as AI literacy			X	X					X
Assess and approve AI embedded in enterprise applications		X	X	X	X	X	X	X	
Regulator interactions				X					
Regulatory change				X					
Risk control assessment		X			X	X			

Given this multi-dimensional nature of AI governance, organizations should be aware of common gaps:

- Communication gaps: AI engineers are not legal experts and vice versa, and the same goes for the other stakeholders. To overcome this gap, consider:
  - Education programs to create a common base of understanding, without everyone having to know the full extent of the other's roles.
  - Automated workflows to integrate everyone's contributions into a unified framework.
- Technology gaps: Individual teams will be tempted to create or deploy tooling for their specific piece of the AI governance puzzle. While their efforts are of course with the best of intentions, it does create an increasing integration problem as the company matures to a more interconnected approach.

To overcome this gap, look for tooling that provides integrated support all of the activities and roles listed in the table above. Limit the use of bespoke integration to only those situations where generally available AI governance tooling does not have the integrated capabilities.

In summary, the challenge lies not in providing isolated capabilities to each group but in ensuring that all these tools and workflows are interconnected effectively. For example:

- Do the technical teams have a clear view of the compliance requirements they will specifically need to fulfill as they develop an AI solution?
- Do risk teams have a clear view of the exposure created by using the same AI models across AI solutions that are created in-house and those that are purchased as part of an application?
- Does the governance solution actively support technical teams in gathering technical documentation details in an easy way?
- When the technical teams update an AI solution, will the legal teams see those changes and be able to re-assess them for any new compliance obligations?

A holistic approach enables seamless governance, fostering trust, accountability, and alignment across the organization.

1.3.3 Technical complexity is increasing

As AI evolved from machine/deep learning to generative AI, the technical complexity has increased in the following ways:

More complex relationship between use cases and AI assets

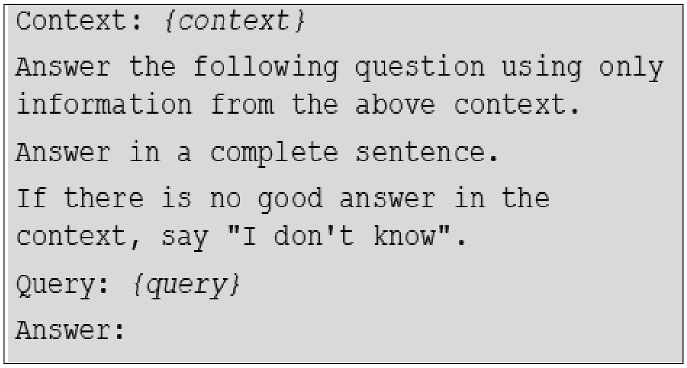
In machine learning and deep learning projects, typically there was a very tight one-to-one relationship between the use case and the AI assets (models). Models were trained to perform a specific task, and if you had a different task, you would train another model. The use case comes first, and then the model(s) follow.

With generative AI and foundation models, this is typically no longer the case as these models are now pre-trained to handle a large variety of use cases. Not only does this mean a one-to-many relationship, but the relationship is also reversed: the model is already there before any specific use cases are considered. One result of this is that the use case has become more important as a governed item in its own right, separate from the (foundation) models applied to deliver a use case. Chapter 5 describes the governance considerations around use cases

More AI assets

In machine learning and deep learning projects, typically the asset created was a trained model. Depending on the use case and the data, these can take many forms but they're generally all referred to as *models*.

With generative AI and foundation models, there are other types of assets used to make a specific use case come to life. The primary way to interact with a foundation models is through a prompt, which is typically text-based instructions on what you want the model to do. For many enterprise use cases, these are not one-off but repeated interactions that take the form of a parametrized prompt template. For example, for a retrieval-augmented generation use case, an LLM is prompted each time a user query comes in. The prompt always has the same structure and core instructions, with parameters for the user's query and the context data that should be used to answer the query. See Figure 1-2 on page 10.



```
Context: {context}
Answer the following question using only
information from the above context.
Answer in a complete sentence.
If there is no good answer in the
context, say "I don't know".
Query: {query}
Answer:
```

Figure 1-2 Example of a parametrized prompt template for a RAG use case

So instead of *training models*, a lot of projects now involve *creating prompts*.

Additionally, organizations might decide to fine-tune a pre-trained model to make it fit their specific needs better. There are different methods to do this that result in what is effectively a new model. It is derived from the base model but is a new object in its own right that needs the appropriate amount of oversight.

Lastly, other AI techniques such as prompt chaining (where different prompts are executed in a defined sequence) might result in other assets to be governed.

Chapter 6, "Governing the end-to-end lifecycle of an AI asset" on page 61 describes the governance considerations for different AI assets

More sources of AI

In machine learning and deep learning projects, organizations usually train their own models, potentially with assistance from external service providers, using either their data or trusted third-party data.

With the advent of Generative AI and foundation models, these models are typically sourced from third party as pre-trained models, such as IBM's Granite family of models. However, organizations often lack adequate insight into the data used to train these external foundation models. Therefore, it is crucial for organizations to conduct a thorough assessment and review of the foundation model before adopting it for their own use. Chapter 4, "Onboarding a new foundation model" on page 41 describes the governance considerations when bringing in these externally developed models.

These multi-purpose foundation models also spur on your technology providers to embed more and more AI in any enterprise applications that an organization might use. Chapter 4, “Onboarding a new foundation model” on page 41 describes the governance considerations for AI embedded in business applications. Another trend to watch is the prebuilt AI capabilities in devices such as laptops and smartphones.

1.3.4 Regulatory and risk complexity is increasing

In this section we review regulatory compliance management and operational risk management.

Regulatory compliance management

The AI regulatory space is very dynamic: There are many AI-specific legislations in force or in progress across the world, there are general regulations that also apply to AI and there are several options for voluntary commitment schemes.

Especially for organizations that operate across multiple regions, the regulatory challenge is becoming more complex quickly and one cannot expect everybody in an organization to be a regulatory expert for AI. As a result, the role of compliance will become more prominent in the “team sport” described in 1.3.2, “Bring together diverse stakeholder perspectives” on page 7.

Luckily, regulatory compliance management is a well-established discipline, which can now also be applied to AI. This includes actions such as:

- Defining compliance requirements and obligations
- Implementing legal assessment tools
- Connecting AI use cases with mandates
- Defining compliance plans
- Managing legal obligations across use cases (for example, AI literacy)
- Managing regulator interactions
- Managing regulatory change
- Regulatory reporting
- Documenting and reporting on compliance status

As AI is set to impact more and more business processes, AI governance becomes more than just a legal check box. Organizations will need to define/update their risk frameworks and enterprise policies; especially as generative AI and agentic AI bring amplified and new risks to an organization.

Operational risk management

Operational risk management is also a well-established discipline, which can now also be applied to AI. This includes actions such as:

- Defining AI risks and mitigation strategies
- Implementing libraries of risks and controls
- Implementing AI risk identification tools
- Connecting AI use cases with business processes
- Defining test plans
- Managing loss events, loss impacts and loss recoveries
- Documenting and reporting on risk assessments and mitigation strategies

1.4 An example of legislation and standards related to AI

Like AI in general, legislation and standards around AI are evolving quickly. In this chapter we cover some headlines and refer the reader to the respective sources of these legislations and standards for their current state and planned further enhancements.

1.4.1 AI-specific legislation

Across the world, AI-specific legislation is being developed or has been enacted. This paragraph describes the key characteristics of one example.

Regulation (EU) 2024/1689 - The EU AI Act

The European AI Act went into force on August 1st, 2024, and is applicable to any organization placing AI on the market in the European Union (EU), regardless of their home base. That means that organizations headquartered elsewhere in the world do have to comply with the Act for AI systems that impact EU citizens, as consumers, employees or other roles.

The AI Act regulates two things:

1. “General purpose AI models” (the legal terminology for what we call foundation models in this publication).
   - A special category is models with “*systemic risk*” - the really large models that can be used so widely that their potential impact is very extensive.
2. “AI systems”
   - Some uses of AI are classified as *prohibited*, they are not allowed to be placed on the market in the EU. Examples are AI systems that use subliminal techniques, enact forms of social scoring or employ untargeted scraping of facial images from the internet or CCTV footage.
   - Some uses of AI are classified as *high risk*, they are allowed but come with a set of requirements to ensure they don't violate the rights of EU citizens. Some of these defined uses are industry-specific (for example in utilities, financial services and public sector) and some are horizontal (for example certain use cases in HR or education).
   - Some uses of AI are classified as having *transparency* risk, they are also allowed and come with obligations to disclose that a person is interacting with an AI system, or that content is created by an AI system.

An organization’s legal obligations depend on the classification of the system or the model, and the role that an organization plays with respect to that system or model (such as provider, deployer, others).

The Act defines penalties for:

- Bringing prohibited AI systems onto the market.
- Not meeting the obligations for high-risk use cases and/or general purpose AI models.
- Providing incorrect, incomplete or misleading information to notified bodies or national competent authorities in reply to a request.

For large enterprises, the fines can be as high as 35 million euro or 7% of its total worldwide annual turnover for the preceding financial year, whichever is higher.

While the Act is in force since August 2024, subsets of obligations will apply from different dates:

- February 2025: prohibited AI systems, AI literacy requirements
- August 2025: obligations for providers of general purpose AI models
- August 2026: obligations for high-risk AI systems
- August 2027: obligations for high-risk AI systems "*intended to be used as a safety component of a product, or the AI system is itself a product, covered by the Union harmonisation legislation*[5]" (for example toys or medical devices)

Other AI-specific legislation includes, for example:

- Act on the Development of Artificial Intelligence and Establishment of Trust (AI Basic Act) (South Korea)
- Local Law 144 regarding automated employment decision tools (US, New York City)
- Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI (US)
- Artificial Intelligence and Data Act (Canada)
- AI regulation: a pro-innovation approach (UK)

1.4.2 General regulations that apply to AI

Besides the evolving AI-specific regulations there are also many general regulations that apply to AI use cases and systems in areas such as:

- Non-discrimination
- Privacy protection
- Data protection
- Product liability
- Fair advertising
- Industry-specific regulations, such as rules around financial advice

1.4.3 Technical standards for AI governance

In addition to regulations, there are various initiatives underway to define technical standards for AI systems. The standards define the state-of-the-art tools and methods that can be applied when creating and using AI systems.

From a governance perspective, these standards will provide a baseline of "generally accepted" practices that organizations are encouraged to adopt.

Technical standards supporting the EU AI Act

The European Commission has requested the development of a set of technical standards covering the following requirements:

- Risk management system for AI systems.
- Governance and quality of datasets used to build AI systems.
- Record keeping through logging capabilities by AI systems.
- Transparency and information provisions for users of AI systems.
- Human oversight of AI systems.
- Accuracy specifications for AI systems.
- Robustness specifications for AI systems.

[5] European Union, *Article 6: Classification Rules for High-Risk AI Systems* (Section 1, Paragraph (a))

- Cybersecurity specifications for AI systems.
- Quality management systems for providers of AI systems, including post-market monitoring processes.
- Conformity assessment for AI systems.

Two European standardization bodies (CEN/CENELEC) have accepted that request and have formed a joint technical committee (“JTC21”) for the development and adoption of standards for AI and related data.

These technical standards will become important compliance tools since they “will grant a legal presumption of conformity to AI systems developed in accordance with them.”
In other words: if you build your AI systems to the specifications in these technical standards, the EU will assume your system is in conformity with the AI Act.

The following list of organizations provides useful information related to AI and AI governance standards.

- NIST RFM

  National Institute of Standards and Technology - AI Risk Management Framework

- ISO

  International Organization for Standardization - Artificial Intelligence

- OWASP Top 10 for Large Language Model Applications

  Open Web Application Security Project (OWASP)

Note: Given the evolving nature of AI governance standards, a flexible approach is crucial. Regularly review and adopt new or updated standards as they emerge to ensure your AI systems align with the latest best practices.

2

Introduction to IBM watsonx.governance

This chapter provides an overview of the watsonx platform, its core components, and features.

This chapter includes the following topics:

2.1 Introduction to the IBM watsonx platform and its core components

Organizations today face a common challenge: accessing data that is trapped in silos across various systems. Business applications typically store their data in application-specific sources, which often leads to duplication across multiple systems to meet different application needs. This fragmentation results in disconnected datasets that are not easily accessible to the broader enterprise ecosystem while still maintaining appropriate access rights. As a result, this limitation hinders opportunities for gaining insights and fostering innovation.

IBM AI and data platforms enable organizations to access and integrate isolated datasets for analytics purposes, ensuring that the data is of high quality and trustworthy. This capability allows companies to develop responsible and ethical solutions, fostering the creation of new products, offerings, and opportunities for business growth.

The IBM watsonx platform, along with IBM Cloud Pak for Data, is built on the robust foundation of Red Hat OpenShift. This next-generation AI and data platform enables enterprises to develop and deploy AI applications, access data stored in legacy systems, catalog data while implementing governance practices, and improve their understanding of data quality. It also allows organizations to establish policies and rules related to data privacy, ensuring that end-to-end AI solutions are ethical, responsible, and trustworthy.

Accessing high-quality datasets can be challenging in organizations when the data assets are locked in silos. The key for organizations to gain insights and bring innovation to their business is leveraging these data assets but requires a platform that enables access to data assets residing in silos, integrate the data assets, and build and deploy AI assets in an ethical and responsible way. The IBM watsonx platform consists of three core components

Figure 2-1 on page 17 illustrates the components of the watsonx platform.

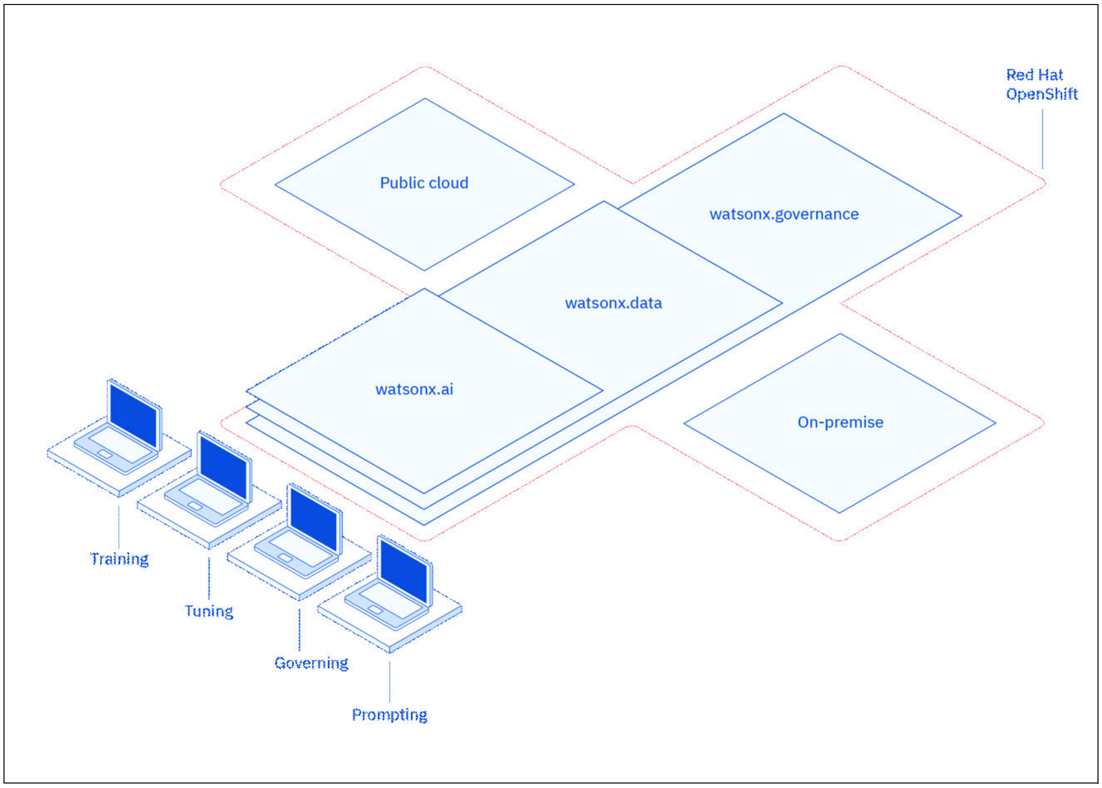


Figure 2-1 IBM watsonx platform and its core components -.ai, .data, .governance

This platform has three components:

- IBM watsonx.ai: An enterprise-grade AI studio that helps operationalize and scale the development of AI applications by bringing together traditional machine learning and generative AI capabilities with high-quality data across the AI lifecycle. With watsonx.ai, AI developers can build, train, adapt, and tune models with your enterprise data and operationalize the models to generate insights, support tasks, and automate business workflows.
- IBM watsonx.data: A fit-for-purpose data lakehouse service, that makes it possible for enterprises to scale AI workloads using all their data optimized for governed data and AI workloads. It serves as a data source for AI, enabling the enterprise ecosystem to access trusted and quality data while enforcing policies and rules for data privacy and security. watsonx.data supports querying, governance, and open data formats to access and share data for different AI use cases such as Retrieval Augmented Generation (RAG). It is based on open-source technologies, including Presto, Iceberg, and Milvus.
- IBM watsonx.governance: An end-to-end solution for AI governance to enable responsible, transparent, and explainable AI workflows in addition to monitoring and evaluation capabilities that allow you to keep track of your entire AI landscape. IBM watsonx.governance helps business analysts understand the trustworthiness of their AI solutions.

2.2 Introduction to IBM watsonx.ai

IBM watsonx.ai is a cutting-edge AI platform that empowers organizations to scale and accelerate their AI initiatives.

Offering a comprehensive suite of tools including a foundation model library, enterprise grade studio, machine learning frameworks, and a runtime serving environment. watsonx.ai enables enterprises to build, train, and deploy models with ease. As a core component of the larger IBM watsonx platform, it drives AI-driven transformation across multiple industries.

Key components of watsonx.ai

The following list describes the key components of watsonx.ai:

- Foundation models: IBM watsonx.ai provides access to large-scale, pre-trained foundation models designed for various AI tasks. These models, such as generative AI and large language models (LLMs), can understand and generate human-like text, making them suitable for applications like customer service automation, content generation, data analysis and more.
- Generative AI: Generative AI in watsonx.ai enables businesses to create new content, generate insights, automate workflows and etc. With the capability to train custom models, watsonx.ai empowers organizations to build personalized AI-driven solutions.
- Machine Learning and ModelOps: watsonx.ai supports the entire AI model development lifecycle, from data preparation and training to deployment and monitoring. Through ModelOps, watsonx.ai ensures that models are optimized, accountable, and compliant with industry standards.
- Data Science and Analytics: With advanced data science tools, watsonx.ai allows organizations to analyze and use data for training AI models. These capabilities provide a strong foundation for data-driven decision-making.

IBM watsonx.ai in enterprise workflows

IBM watsonx.ai integrates seamlessly with existing enterprise workflows. By offering pre-built connectors and APIs, the platform allows businesses to embed AI models into their operational workflows quickly, tailoring solutions to meet specific requirements.

IBM watsonx.ai as a developer and data scientist toolkit

IBM watsonx.ai provides a flexible toolkit for developers, AI engineers, and data scientists, supporting coding and no/low-coding environments. Developers can leverage SDKs, APIs, and workflows in their preferred programming languages, while non-coders benefit from intuitive natural language processing tools. watsonx.ai also supports IDEs like RStudio and Python Notebooks, accelerating AI model production.

Domains where watsonx.ai is useful

IBM watsonx.ai has been widely adopted across industries, supporting use cases such as natural language processing, image recognition, fraud detection, and predictive analytics. Key industry applications include:

- Banking and Financial Services: Automating processes such as credit scoring, fraud detection, and risk assessment, which improves accuracy and efficiency.
- Healthcare: Enhancing diagnostics, personalizing treatment plans, and optimizing workflows using large AI models.
- Retail: Improving customer experience through personalized recommendations (next best offer), basket analysis, inventory management, and automated support systems.
- Manufacturing: Streamlining production processes, predictive maintenance, and quality control to improve efficiency and reduce downtime.
- Transport: Enhancing route optimization, fleet management, and safety systems for improved logistics and customer service.

- Leisure and Luxury: Personalizing customer experiences, optimizing inventory, and enhancing marketing strategies to meet customer preferences.

Advantages of watsonx.ai

The advantages of watsonx.ai include the following items:

- Dependability: All models offered within the watsonx.ai platform are thoroughly tested, resulting in highly robust models that deliver consistent results.
- Transparent Accountability: All Granite models (IBM's proprietary family of models) are trained exclusively on open-source data, ensuring no hidden liabilities or legal complexities when used in AI solutions.
- Scalability: Supports large AI models and integrates seamlessly with data platforms to scale AI operations.
- Flexibility: Offers customization options, allowing AI models to be tailored to specific business needs. Bring your own foundation model and upload to watsonx.ai to accomplish a range of industry and domain-specific generative AI use cases.
- Enhanced Governance: Ensures responsible development and deployment of AI solutions.
- Faster ROI realization: Accelerates AI solution development and deployment, enabling businesses to quickly realize benefits.

Conclusion

IBM watsonx.ai represents a significant advancement in AI, combining the creativity of generative AI with the precision of traditional machine learning. Its comprehensive suite of tools and strong focus on governance enable businesses to accelerate AI initiatives responsibly and at scale. As AI continues to evolve, watsonx.ai stands at the forefront, helping organizations unlock new possibilities and drive impactful innovations across many sectors.

> **Note:** For more information on watsonx.ai refer to IBM Redbooks *Simplify Your AI Journey: Unleashing the Power of AI with IBM watsonx.ai*, SG24-8574.

2.3 Introduction to IBM watsonx.data

IBM watsonx.data provides a modern, open data lakehouse architecture that integrates seamlessly across on-premises and multi-cloud environments. Its design enables organizations to manage all types of data workloads, from traditional analytics to AI training.

Core features of watsonx.data

The following list highlights the core features of watsonx.data:

- High-performance Data Querying and Analytics: Provides rapid querying capabilities for large datasets, empowering businesses to extract insights efficiently.
- Built-in Governance and Security: Ensures compliance with data privacy and security regulations across multi-cloud environments.
- Cost Optimization: Reduces data warehousing costs by up to 50%, offering a more economical solution for data storage and processing.
- Shared Metadata Layer: Facilitates seamless data access and operations, streamlining workflows and improving data consistency.

Key features

IBM watsonx.data key features include the following items:

- Scalable Data Lakehouse: watsonx.data blends the flexibility of data lakes with the high performance of data warehouses. It is designed for hybrid cloud environments and supports open data formats like Parquet and ORC.
- SQL-based Querying: Enables users to execute high-speed analytics with SQL, allowing data engineers and analysts to handle extensive datasets effectively.
- Machine Learning Integration: Seamlessly integrates with machine learning workflows, enabling the development and deployment of AI models using real-time data streams. This integration enhances decision-making and automates processes within organizations.

> **Note:** For more information on watsonx.data refer to IBM Redbooks *Simplify Your AI Journey: Hybrid, Open Data Lakehouse with IBM watsonx.data,* SG24-8570.

2.4 Introduction to IBM watsonx.governance

IBM watsonx.governance is a dedicated application for governance of AI to help organizations ensure that their use of AI is profitable, compliant, secure and fair.

2.4.1 Key capabilities

The key capabilities of watsonx.governance include the following items:

- Maintaining an inventory of all AI inside an organization, including self-built solutions, AI embedded in your enterprise application and AI embedded in your products and services.
- Automated AI discovery - through integration with IBM Guardium AI Security, you can identify AI deployments that are not yet registered and apply the appropriate level of governance.
- Role-based dashboards.
- Identify legal obligations for your use cases.
- Identify AI risks for your use cases, leveraging the IBM AI Risk Atlas created by IBM Research and the IBM AI Ethics Board.
- Automate your use case review processes.
- Automate your third-party foundation model onboarding processes.
- Automate your lifecycle governance processes such as model reviews and model change requests.
- Automate your regulatory compliance management processes such as managing regulatory chance and managing regulator interactions.
- Manage legal obligations across use cases (for example, AI literacy).
- Automate your operational risk management processes such as managing loss events, loss impacts and loss recoveries.
- AI documentation through capture of metadata about the (versions of) AI assets you build.
- Quantitative evaluations of your AI assets - dozens of pre-built metrics to measure model health, drift, quality, toxic language, PII, fairness, adversarial robustness of both machine learning and generative AI use cases.
- Global and local explainability of machine learning models.

- Standard and ad-hoc reporting.
- Integration with AI platforms such as IBM watsonx.ai, Amazon SageMaker and Bedrock, Google Vertex AI, and Microsoft Azure.
- Configure the solution to fit your specific requirements.

2.4.2 Use cases

IBM watsonx.governance is a comprehensive governance solution. Depending on a customer's situation, they could use the solution for one or more of the following use cases:

- Comply with the EU AI Act or other legislation.
- Mitigate reputational and operational risks from their use of AI.
- Govern the onboarding of AI-enabled enterprise applications.
- Govern the development of AI-enabled products and services.
- Integrate AI risks into the operational risk management of AI-enabled business processes.
- Set up regular monitoring of an already deployed AI solution.

2.4.3 Benefits of watsonx.governance

By centralizing and automating governance of AI, customers achieve benefits such as:

- Bring together all technical and non-technical stakeholders into a common governance framework.
- Decrease the risk of being fined or damaging your brand reputation.
- Reduce the cost of compliance.
- Free up time to deploy more models.
- Capture model benefits earlier by reducing the time to deployment.
- Recapture model benefits lost due to model drift.

As mentioned in 1.2.3, “Concern 3: Governance does not contribute to value generation” on page 6, using AI governance software also contributes to setting up an organization for the value generation benefits mentioned in that paragraph.

2.4.4 Synergy between watsonx.data and watsonx.governance

IBM watsonx.data and IBM watsonx.governance provide a holistic solution for organizations. While watsonx.data focuses on data analysis and AI workloads, watsonx.governance ensures that all data is managed responsibly, securely, and in compliance with relevant regulations.

By combining the strengths of both components, organizations can develop a unified data strategy that balances innovation with accountability.

2.4.5 Synergy between watsonx.ai and watsonx.governance

The synergy between IBM watsonx.ai and IBM watsonx.governance makes it easy for AI developers to contribute to AI governance processes. The integration between the two components automates even more of the work, for example:

- Model metadata is automatically captured.
- Model evaluations can be easily set up from their development workspaces (UI or programmatic).

- Inferencing payload data can be auto-logged for fully automated runtime model monitoring.

This comprehensive approach allows organizations to balance rapid innovation with the responsibility of managing AI risks effectively. The combined capabilities help businesses streamline their workflows and maintain an audit trail, enabling a seamless interplay between AI development and AI governance.

2.5 Reference architecture

Figure 2-2 illustrates a reference architecture for an end-to-end AI and data solution that leverages the IBM watsonx platform and IBM Cloud Pak for Data. This architecture is designed to build, test, deploy, manage, govern, and consume AI solutions across the enterprise.

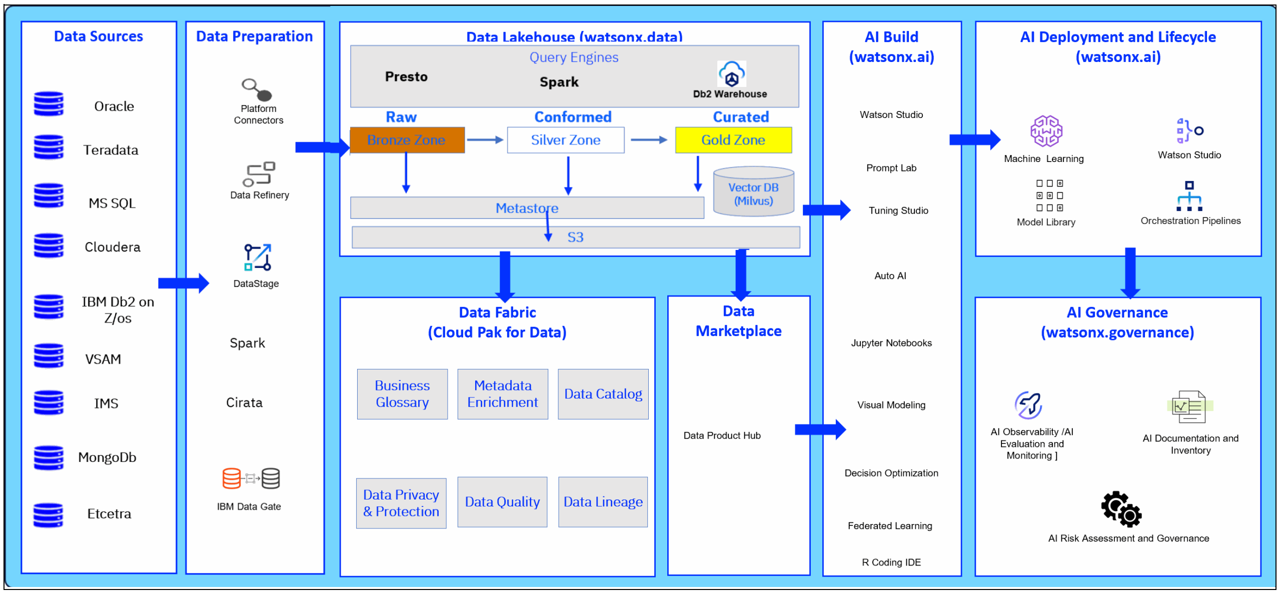


Figure 2-2 Reference architecture for integrated AI and data platform for an enterprise using IBM watsonx

2.5.1 Data Onboarding

The left side of Figure 2-2 illustrates the existing and legacy data sources that drive the organization's day-to-day operations. These data sources are often designed and developed for specific purposes and are accessible only by a select group of users or applications. Typically, these existing and legacy data sources are built in silos, lacking a clear understanding of the relationships between the data assets they contain. To resolve this, organizations can use watsonx.data to store structured, semi-structured, and unstructured data and make it directly accessible for AI and business intelligence (BI).

To integrate these disjointed data sources, we can onboard them onto IBM's open architecture lakehouse, watsonx.data, which combines elements of a data warehouse and data lakes. It offers a unified platform where users can store data or connect data sources to manage and analyze enterprise data.

IBM watsonx.data allows two approaches for onboarding data into the platform:

- Accessing the data in place.

- Replicating the data onto the platform in Iceberg open data format using various extract, transform, and load (ETL) options.

IBM watsonx.data allows users to access data in existing data warehouses and data lakes through predefined platform connectors, including Teradata, Snowflake, SingleStore, SQL Server, PostgreSQL, MySQL, MongoDB, IBM Db2, and IBM Netezza. This solution reduces data duplication, and the costs associated with storing data in multiple locations.

Suppose existing data is stored in an external storage system such as IBM Cloud Object Store, Amazon S3, IBM Storage Ceph, MinIO, HDFS (in Hadoop/Cloudera), Google Cloud Storage, or Azure Data Lake Storage. In that case, it can be accessed directly if the data is in Iceberg or Delta Format. If the data is stored in other common formats like Parquet or CSV, it can be accessed in its native format, followed by running ETL jobs to convert it to the open format like Iceberg and brought into the platform.

Cirata is an IBM partner that developed a cloud migration solution that automates the seamless transfer of continuous HDFS data and Hive metadata to watsonx.data.

IBM watsonx.data supports loading data from existing on-premises data lakes using ETL jobs developed with DataStage or Spark. Additionally, data loading can be done through a web console or command line interface.

Existing data sources like Db2 for z/OS, IMS, and VSAM on the mainframe can be replicated using IBM Data Gate in the Iceberg open data format within watsonx.data. Data Gate is a replication technology that synchronizes data from IBM Z to various hybrid-cloud targets.

The architecture of watsonx.data enforces schema and data integrity, facilitating the implementation of robust data security and governance mechanisms. Integrating watsonx.data with IBM Knowledge Catalog on Cloud Pak for Data provides knowledge workers with self-service access to data assets, allowing them to utilize these assets to gain insights. Once the data from these sources is onboarded into the platform in its raw format, the data assets from watsonx.data are imported into a governed catalog in IBM Knowledge Catalog, where they can be enriched with business semantics by mapping business terms to technical data assets (such as database tables and columns). Data quality can be assessed through profiling and running data quality analyses. Data protection rules are established to control access to sensitive data in governed catalogs, which can include denying access, redacting columns, obfuscating columns, substituting columns, or filtering rows.

In addition, semi-structured and unstructured data and documents can be processed, split, and stored in the watsonx.data vector database, Milvus, along with their metadata and vector embeddings in the same datastore. Milvus is designed to store, index, and manage embedding vectors used for similarity search and retrieval-augmented generation, empowering embedding similarity search and AI applications.

2.5.2 Data Preparation

Businesses can derive a greater value by integrating data from various sources and domains into higher layers, such as the silver and gold layers of a medallion data architecture. Additionally, watsonx.data provides multiple query engines, including Presto and Spark, allowing users to select the most suitable engine based on the characteristics of their workload. IBM watsonx.data seamlessly integrates with Db2 Warehouse and Netezza Performance Service, facilitating data sharing across these products and enabling users to leverage the most appropriate engine for each specific task.

Once structured and unstructured data is onboarded to watsonx.data in the watsonx platform, users can access the data assets as long as they have the necessary access rights. They can build, train, tune, and deploy traditional ML models or utilize generative AI models to enhance their business use cases with AI.

2.5.3 AI Building and Deployment

IBM watsonx.ai provides low-code and no-code tools, including Auto AI for creating traditional machine learning models, Prompt Lab for prompt engineering, and Prompt Studio for adapting and fine-tuning generative AI models. Additionally, watsonx.ai offers a Studio and a Python Software Development Kit (SDK) designed for data scientists and AI engineers to build, train, adapt, fine-tune, and deploy both traditional machine learning and generative AI assets.

2.5.4 AI Lifecycle Management and Governance

The IBM watsonx platform enables end-to-end AI lifecycle management and governance. IBM recommends following an end-to-end process to build AI solutions that are ethical and responsible - secure, safe, transparent, and trustworthy, as shown in Figure 2-3.

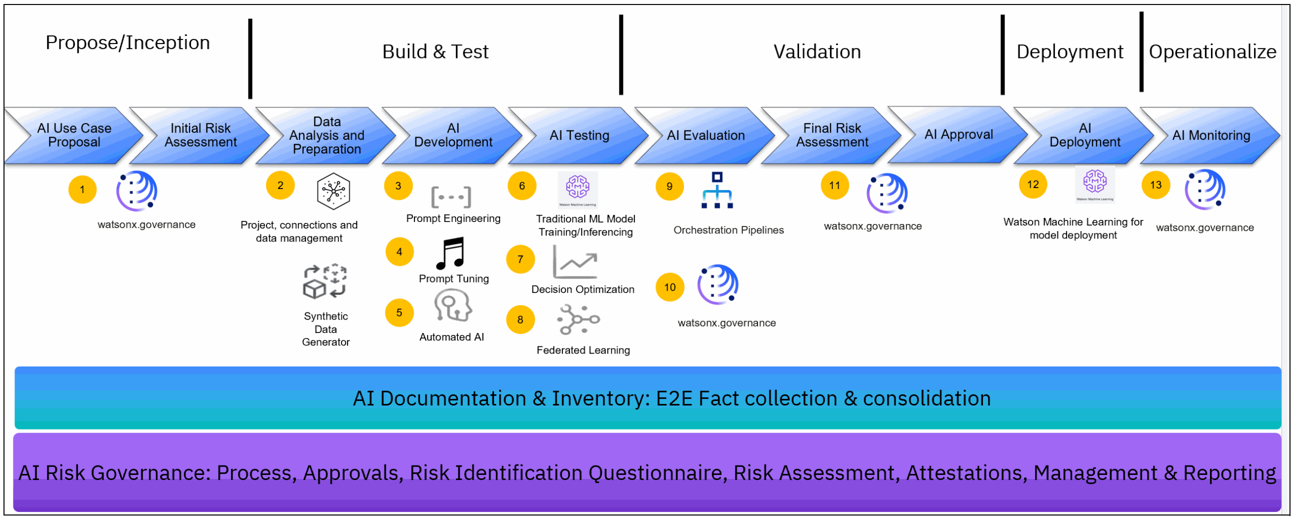


Figure 2-3 End-to-end flow for AI lifecycle and governance powered by IBM watsonx

In this approach, an AI Use Case Owner or Requestor starts by creating an AI use case in watsonx.governance. This involves outlining the business purpose of the use case and detailing how AI will be utilized to achieve the intended outcomes.

Next, the AI Use Case Owner identifies potential risks associated with the AI use case by answering a set of questions and conducts an initial risk assessment within watsonx.governance. Once this risk identification and assessment are completed, input from legal, HR, the AI Ethics Council, operations, security, and finance is gathered. Based on the risk analysis and risk profiles associated with the use case, a decision is made to approve or reject the development of the AI use case and its related assets.

If the use case is approved for development, the AI Developer, which can be an AI Engineer or Data Scientist, collaborates with a Data Engineer to shape and prepare the dataset in watsonx.data for either training traditional ML models or tuning large language models.

Following the data preparation, the AI Developer can build and train traditional ML models or adapt or tune for generative AI models specific to the use case. The AI Developer then validates these traditional ML models and AI assets leveraging generative AI models and test data to evaluate performance metrics such as F1 score and ROUGE, as well as fairness, explainability, and model health (including latency, number of transactions, and token count).

Subsequently, a Model Validator-an independent Data Scientist or AI Engineer- evaluates the traditional ML model or Generative AI assets in a pre-production environment using production-like data within watsonx.governance.

An AI Risk Reviewer then examines the evaluation performance metrics, risk scorecard, and model health metrics. Based on this final risk assessment, the reviewer either approves or rejects the AI asset for deployment in production within watsonx.governance.

Finally, the ModelOps Engineer deploys the approved model into production using CI/CD pipelines and activates ongoing AI monitoring to track key metrics, such as runtime quality, drift, fairness, explainability, and other relevant indicators specific to the use case. The ongoing monitoring metrics are published in the Governance Console in watsonx.governance. Each metric has associated thresholds, and alerts can be configured in the Governance Console to notify relevant stakeholders, such as the AI Use Case Owner or AI Risk Reviewer, when these thresholds are breached. Once the appropriate parties receive notifications about threshold breaches, they can initiate the investigation process or follow the procedures for issue and change management, which may involve redeveloping a new version or completely different AI asset.

Facts, documentation, and evidence regarding AI assets are collected and stored in watsonx.governance for future reviews and audits by regulatory agencies, internal stakeholders, and external parties. Additionally, specific reports related to the AI lifecycle and its stages can be generated in the Governance Console of watsonx.governance.

3

Implementing AI governance strategy

Artificial Intelligence (AI) governance is a multifaceted process that ensures AI systems are developed and deployed responsibly and ethically. This chapter provides a comprehensive overview of the end-to-end AI lifecycle governance process, exploring the various steps involved in achieving this goal.

This chapter contains the following sections:

- “Understanding the end-to-end AI lifecycle governance process” on page 28
- “Elements of model risk governance” on page 29
- “Considerations to implement AI governance strategy” on page 37

3.1 Understanding the end-to-end AI lifecycle governance process

This section describes the governance process which is divided into two primary levels: macro and micro as shown in Figure 3-1.

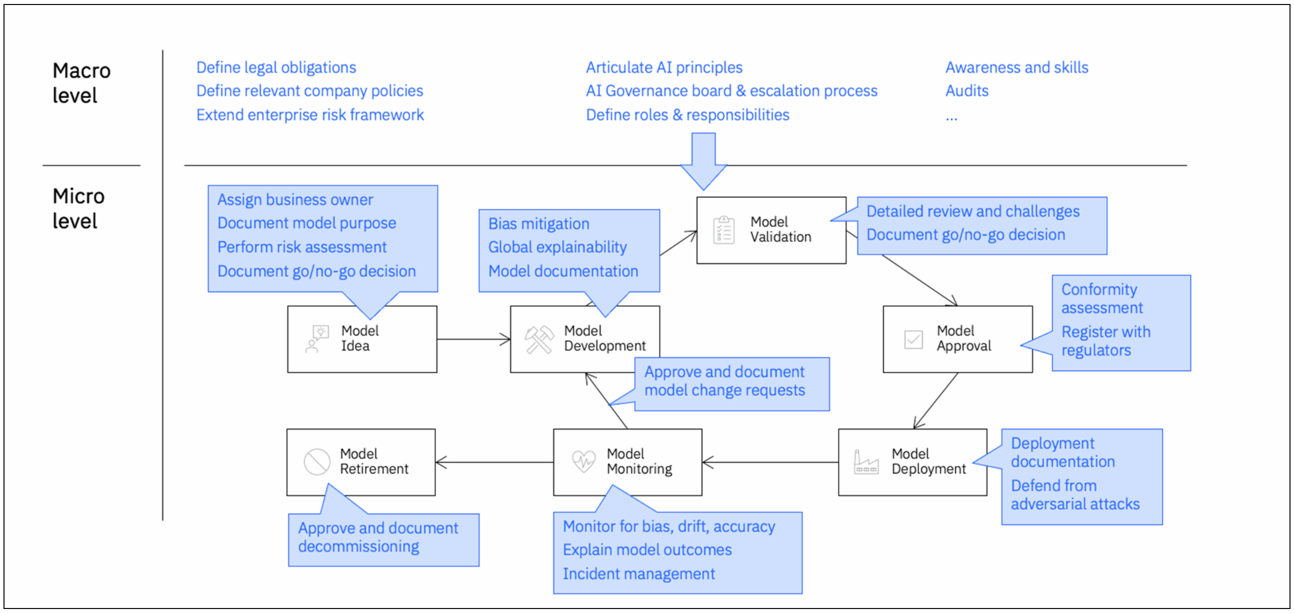


Figure 3-1 End-to-end AI lifecycle governance process

Macro level

The macro level encompasses high-level strategic steps such as defining legal obligations, articulating AI principles, and extending enterprise risk frameworks. These steps ensure that the overarching governance structure aligns with organizational values and regulatory requirements.

Micro level

The micro level involves more granular tasks, including assigning business owners, performing risk assessments, and documenting go/no-go decisions. These steps ensure that day-to-day operations are conducted in a manner that supports the macro-level objectives.

The intersection of macro and micro-level specifics dictates actions related to AI models, such as model approval, deployment, and monitoring. This level ensures that each AI model is rigorously evaluated and managed throughout its lifecycle.

The process begins with model proposal approval, where a model entry is created in the Model Inventory and continuously updated with new information. The data scientist then uses a tool of their choice to develop the model, with training data and metrics from popular open-source frameworks automatically captured and saved to the model entry. Custom information can also be saved.

Next, the pre-production model is evaluated for accuracy, drift, and bias, with performance metadata captured and synced. The model is then reviewed and approved for production, and deployed in the preferred platform, with relevant metadata captured and synced. Finally, the production model is continuously monitored, with performance data captured and synced, and a dashboard provides a comprehensive view of the performance metrics for all models, allowing stakeholders to proactively identify and react to any issues.

This chapter will delve into the end-to-end AI lifecycle governance process, highlighting the key personas, and components of each level and how they interconnect to ensure that AI systems are developed and deployed in a manner consistent with organizational values and goals. We will also discuss the challenges and opportunities associated with implementing this process, emphasizing the benefits of improved transparency, accountability, and trust.

Throughout this chapter, we will examine the key components of each governance level and discuss how they work together to ensure responsible and ethical AI development and deployment. By understanding these components, readers will gain a comprehensive view of the governance process and its importance.

By the end of this chapter, readers will have a thorough understanding of the end-to-end AI lifecycle governance process. They will be able to identify the key components of each level and understand how they work together to ensure that AI systems are developed and deployed responsibly and ethically. This knowledge will be invaluable for applying AI governance strategies within their own organizations.

3.2 Elements of model risk governance

To convert macro-level requirements into micro-level ones, a framework is needed to map these requirements. This is achieved through the use of personas that define the high-level requirements, while watsonx.governance objects are essential for effective management and governance of AI models throughout their lifecycle. This chapter will explore the elements required to implement the AI governance strategy

3.2.1 Personas

Figure 3-2 on page 30 shows a typical governance flow which begins with defining an AI use case to solve a business problem, followed by requesting an AI asset, such as a model or prompt template, to address the issue. The process involves various roles, starting with the model owner, who defines the problem and identifies the need for an AI solution. The developer then builds the AI asset, which is subsequently tested by the validator to ensure it meets the required standards. Next, the risk officer reviews and approves the solution, taking into account organizational risk management policies. Once approved, the ModelOps engineer deploys the AI asset into production, and finally, the application developer monitors its performance, identifying areas for improvement. Throughout this process, some organizations may choose to combine certain roles or responsibilities, tailoring the governance flow to their specific needs.

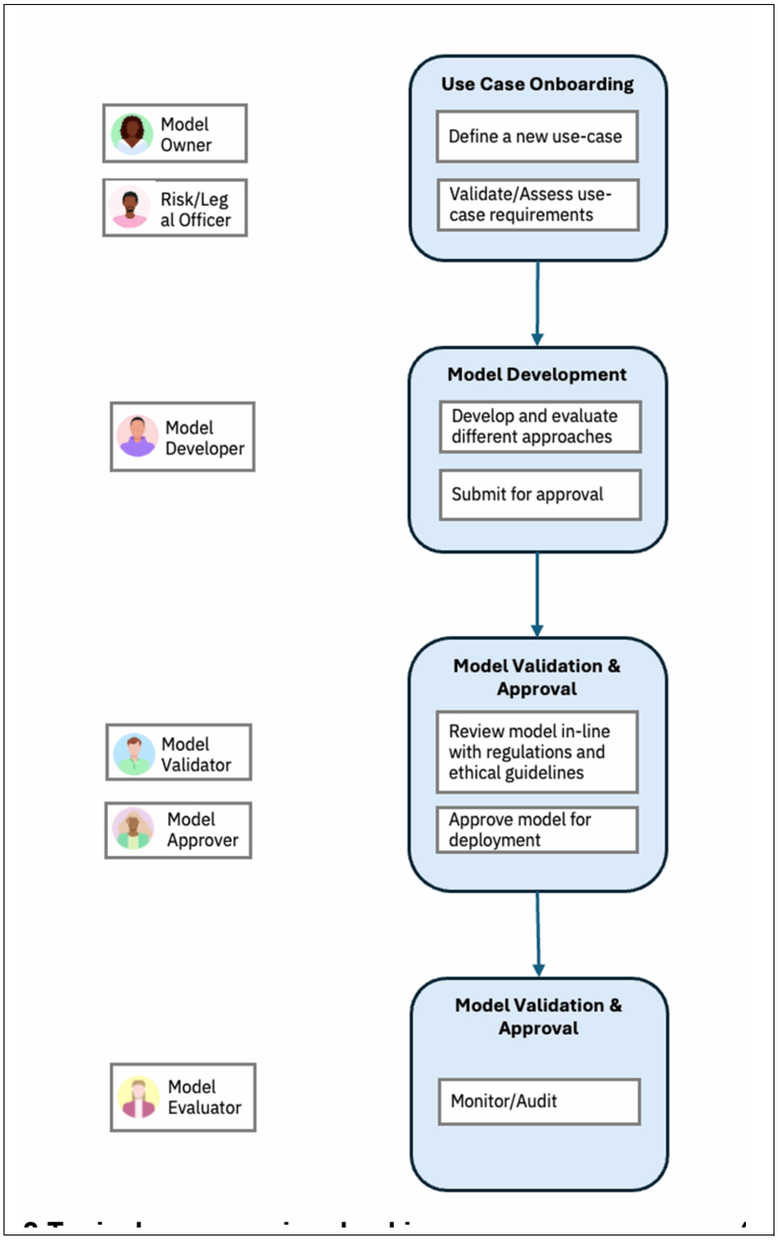

Model Owner
Risk/Leg al Officer
Use Case Onboarding
Define a new use-case
Validate/Assess use-case requirements
Model Developer
Model Development
Develop and evaluate different approaches
Submit for approval
Model Validator
Model Approver
Model Validation & Approval
Review model in-line with regulations and ethical guidelines
Approve model for deployment
Model Evaluator
Model Validation & Approval
Monitor/Audit

Figure 3-2 Typical personas involved in a governance process (source)

Consider the expertise required for your governance team. A typical governance plan may include the following roles, which can sometimes be filled by the same person or, in other cases, represent a team of people.

- Model Owner: This individual creates an AI use case to address a business need, requests the model or prompt template, manages the approval process, and tracks the solution through the AI lifecycle.
- Risk and Compliance Manager/ Legal Team: This person determines the policies and compliance thresholds for the AI use case, such as rules for testing fairness or screening output for hateful and abusive speech.
- Model Developer or Data Scientist: This role involves working with the data in a dataset or a large language model (LLM) to create the machine learning model or LLM prompt template.

- Model Validator: The validator tests the solution to ensure it meets the goals outlined in the AI use case.
- Model Evaluator: After deployment, the app developer evaluates the deployment to monitor performance against the metric thresholds set by the risk and compliance manager. If performance falls below specified thresholds, the app developer collaborates with other stakeholders to address issues and update the model or prompt template.

3.2.2 Objects

Figure 3-3 illustrates the key components of AI governance, which is a critical aspect of ensuring responsible and ethical AI development and deployment. The three main elements of AI governance are model risk governance, model inventory, and lifecycle tracking, and evaluation and monitoring. Model risk governance involves identifying and mitigating potential risks associated with AI models, such as bias, accuracy, and security. Model inventory and lifecycle tracking involves maintaining a comprehensive record of all AI models in use, including their development, deployment, and maintenance. Evaluation and monitoring involves continuously assessing the performance and impact of AI models, identifying areas for improvement, and making necessary adjustments. By implementing these components, organizations can ensure that their AI systems are transparent, accountable, and aligned with their values and goals.

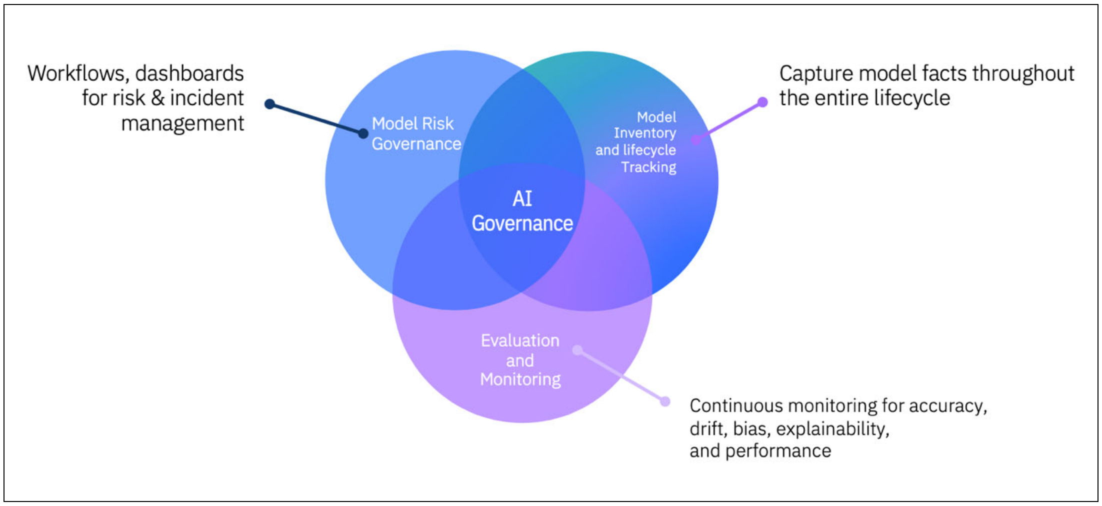

Workflows, dashboards for risk & incident management
Model Risk Governance
Model Inventory and lifecycle Tracking
Capture model facts throughout the entire lifecycle
AI Governance
Evaluation and Monitoring
Continuous monitoring for accuracy, drift, bias, explainability, and performance

Figure 3-3 A governed, trusted AI lifecycle

To implement the framework, several objects play a crucial role in ensuring the effective management and governance of AI models throughout the lifecycle. These objects interact with different personas, including Model Developers, Model Owners, Model Validators, and Model Risk Reviewers.

Figure 3-4 on page 32 illustrates a hierarchical framework for implementing AI governance strategies, comprising three key components: Evaluation and Monitoring, Model Validation and Review, and Risk and Compliance. Such a framework enables the creation of multiple hierarchical structures, each with its own unique architecture, which can be used to organize and capture Models. One possible organizational structure could include divisions into groups such as geography, business unit, line of business etc.

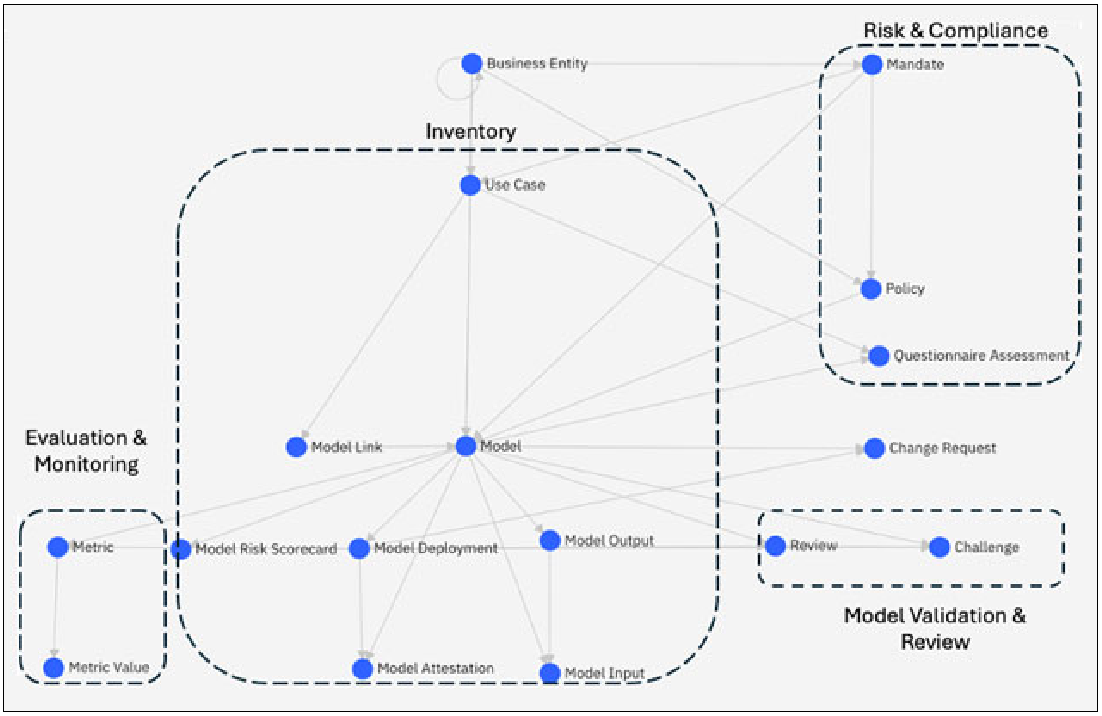


Figure 3-4 Hierarchical framework for implementing AI governance

Business Entity

Business Entities are abstract representations of your business structure. A business entity can contain sub-entities (such as departments, business units, or geographic locations). This structure is used within the system to organize access rights and simplify corporate reporting needs.

For example: A bank's retail lending department is a Business Entity.

Persona interactions:

- Model Developers: Create models that meet the specific needs of the Business Entity. For example, a Model Developer may create a credit risk model for the retail lending department.
- Model Owners: Ensure that models are aligned with the Business Entity's goals and objectives. For example, the Model Owner for the retail lending department may ensure that the credit risk model is aligned with the department's risk appetite.
- Model Validators: Validate models to ensure they meet the Business Entity's requirements. For example, a Model Validator may validate the credit risk model to ensure it meets the retail lending department's requirements for accuracy and fairness.

Inventory

A centralized repository that organizes, documents, and maintains an enterprise-wide collection of models or AI assets, including their usage, issues, and governance activities.

Persona interaction:

- Model Owner: Interacts with the inventory to document model changes and updates, track issues and resolve model-related problems

- Risk and Compliance Manager: Interacts with the inventory to schedule and track model reviews, assign and track model risk assessments, monitor model performance and identify potential compliance risks associated with models

Inventory is primarily comprised of the following objects:

1. **Use Case** - The Use Case object is a subclass of Entity and a superclass of the Model object. Its main function is to serve as a repository for models during development. The use case encapsulates both the qualitative and quantitative requirements of the AI application to be deployed. It helps define and demonstrate how a specific model can solve a particular business problem or achieve an objective. It describes practical scenarios and contexts in which the model will be implemented, offering comprehensive insight into its intended uses and expected results. Additionally, use cases synchronize the watsonx.governance Evaluation and Monitoring Component (performance metrics and quality control) with the Model Inventory and Lifecycle Tracking Component (model reports). Changes in one are automatically reflected in the other. This ensures that users always have up-to-date information about AI models.
2. **Model** - The Model object represents a quantitative method, system, or approach used within an organization to transform input data into quantitative estimates. This is achieved through the application of statistical, economic, financial, or mathematical theories, techniques, and assumptions.

   The Model object captures essential information, including:

   - **Model Description**: A detailed description of the model
   - **Model Ownership**: Information about the model's owner
   - **Model Status**: The current status of the model
   - **Development Lifecycle Dates**: Important dates related to the model's development
   - **Model Type and Category:** Classification of the model
   - **Model Risk Assessment data**: Data related to the model's risk assessment
3. **Model Deployment** - The Model Deployment object is a child entity of the Model object, serving as a crucial component for tracking the deployment of one or more models. The Model Deployment object is designed to:
   - Govern individual usages of a Model in a production ecosystem
   - Record-specific Model Versions and their deployment locations
   - Inform risk tiering of Models by highlighting the number of areas or functions supported by each Model
4. **Model Attestation** - Model Attestation is a process that enables organizations to request regular sign-offs or attestations for their models. The MRG administrator initiates this process by creating a set of blank model attestations, which are then assigned to the respective model owners. These owners are required to answer a series of questions about their models and submit their completed attestations. Typically, model attestations are conducted on an annual or quarterly basis, serving as a way to verify the completeness and accuracy of a model's information, as well as the overall model inventory.
5. **Model Output** - For organizations seeking a more detailed approach to model documentation, the model output object offers a solution. This object enables the recording of a model's outputs, with a focus on capturing the description and overview of each output from a governance perspective. By utilizing the model output object, organizations can maintain a more granular and comprehensive record of their models' outputs.

6. **Model Input** - For organizations seeking a more detailed approach to model documentation, the Model Input object provides a means to capture and record the inputs of a model. The object includes key fields such as:
   - **Input Owner**: The individual or team responsible for the input
   - **Type**: The classification of the input
   - **Status**: The current state of the input
   - **Description**: A detailed explanation of the input

   Additionally, a model input object can be linked to a model output object, allowing for the creation of model chains at a granular level. This provides an alternative to the model link approach, offering a more detailed and nuanced understanding of model relationships.
7. **Model Risk Scorecard** - Model risk assessments are a critical component of the model development and documentation process, as well as an ongoing requirement for models in production. To facilitate this process, the Model Risk Scorecard object is utilized to conduct thorough risk assessments. This process involves answering a series of questions about the model, which triggers a calculation of a risk score. This score, in turn, determines the model's tier, providing a clear indication of the model's risk level.

Model validation and review

During this phase, the model use case and facts collected during the development phase are utilized to validate the model. The goal is to ensure the model is functioning as intended and meets the required standards.

Persona interaction:

- Model Validator reviews the model documentation and use-case. In case of any shortcomings, they challenge the inconsistence and clarify any unclear or missing information

To implement a model validation strategy, the following objects are primarily utilized:

1. **Review** - The review object is a critical component of model governance, serving as a record of all model review activities. As a child of both the model deployment and model objects, it provides a comprehensive view of review outcomes. The review object is designed to capture the results of various types of reviews (for example, pre- and post-implementation), and reviews consumed by independent teams to ensure model integrity and effectiveness.

   By utilizing the review object, organizations can maintain a centralized record of all model review activities, facilitating informed decision-making and effective model governance.
2. **Challenge** - The challenge object serves as a repository for documenting and evidencing concerns or issues related to any part of the Model Inventory. When a challenge is raised, the response is recorded, providing a clear audit trail. As a child of both the model and model deployment objects, the challenge object ensures that all relevant information is linked and easily accessible.

   There could be many factors prompting different personas to challenge existing deployments, such as:
   - Regulatory non-compliance: Models that fail to meet new or updated regulatory requirements
   - Data relevancy issues: Models that rely on outdated, incomplete, or inaccurate data
   - Performance deterioration: Models that experience a decline in performance over time

When a challenge is identified, stakeholders will review it in accordance with established processes to address the concerns. This may involve remediation efforts to bring the model into compliance, update the data, or improve the model's performance.

Evaluation and monitoring

To ensure the model operates within acceptable parameters, performance metrics are continuously monitored to circumvent reputational and organizational risks. During the evaluation and monitoring phase, developers, model validators and ModelOps engineers interact with the platform to:

- Monitor performance metrics and thresholds.
- Analyze model outputs and detect potential issues.
- Receive alerts and notifications for threshold breaches.
- Investigate and respond to potential issues or anomalies.
- Collaborate to identify root causes and develop solutions.

Persona interaction:

- Model Validator interacts with the platform to: Monitor performance metrics and thresholds, analyze model outputs and detect potential issues, and identify data quality problems or biases
- ModelOps Engineer interacts with the platform to Interacts with the platform to receive alerts and notifications for threshold breaches, investigate and respond to potential issues or anomalies and collaborate with Model Evaluators to identify root causes

For evaluation and monitoring, the following objects are primarily utilized:

1. **Metric** - The Metric object is used to record the definition of a performance measurement that an organization wants to track. This involves setting key parameters, including:
   - **Metric Type:** The type of performance metric being tracked
   - **Threshold:** The threshold at which the metric is considered cautionary, critical, etc.
   - **Collection information:** Additional details about the metric's collection and calculation
2. **Metric Value** - The Metric Value object records the result of the metric performance measurement. It is designed to behave in a way to allow the organization to store time series results of measurement.
3. **Change Request** - The change request object is a critical component of model governance, providing a structured process for requesting, justifying, and approving changes to models after they have gone live. This object's workflows enable organizations to:
   - **Request and justify changes:** Clearly articulate the need for changes and provide supporting rationale
   - **Route changes for approval:** Direct changes to the relevant stakeholders and approvers, based on the type and impact of the change
   - **Obtain auditable approvals:** Ensure that all changes are properly approved and documented, with a clear audit trail

The Change Request object allows for various approval paths and levels of approval, depending on the nature and scope of the change. This flexibility enables organizations to balance the need for control and oversight with the need for agility and responsiveness.

Risk and compliance

The increasing use of AI models in various industries has raised concerns about risk and compliance. AI models can pose significant risks if not properly designed, trained, and deployed, including biases, errors, and unintended consequences. Moreover, AI models must comply with various regulations and standards, such as the European Union AI Act, United States AI Executive Order, CCPA, and HIPAA, to ensure the protection of sensitive data and prevent harm to individuals. To mitigate these risks, organizations must implement robust risk management and compliance frameworks that include model risk assessments, data quality checks, and ongoing monitoring and evaluation.

Since risk and compliance management involves identifying, assessing, and mitigating potential risks associated with non-compliance with external mandates and internal policies. The following components are essential to a robust risk and compliance framework:

1. **Mandate** - Mandates represent external requirements that organizations must comply with, such as laws, regulations, and standards. When necessary, mandates can be broken down into Sub-Mandate objects, which provide a more detailed understanding of the mandate's sub-sections. This hierarchical structure enables organizations to effectively manage and comply with complex regulatory requirements.
2. **Requirement** - Requirements represent specific obligations that an organization must fulfill to comply with related mandates or sub-mandates. These requirements are detailed in the requirement object, which provides a clear understanding of what the organization needs to do to meet regulatory obligations. By detailing these requirements, organizations can ensure they are meeting their regulatory obligations and maintaining compliance.
3. **Policy** - Policies represent internal guidelines adopted by an organization's Board of Directors or senior governance body. These guidelines are designed to provide direction and oversight for the organization's operations and decision-making processes. The Policy object is used to store and manage policy information, including the policy text, which can be stored in standardized fields or as an attachment to the object.

   Policies are managed through a review and approval process, which ensures that they are properly vetted and authorized before being implemented.
4. **Questionnaire Assessment** - Questionnaires are a powerful tool for assessing risk and compliance, as well as collecting information for specific processes and asset risks. watsonx.governance streamlines, standardizes, and centralizes the collection of questionnaire-based assessment information, making it easier to gather insights from across the organization.

   With Questionnaire Assessment object, information from business users within the organization can be gathered. Respondents complete the questions and submit the finished questionnaire assessment, providing valuable insights into risk and compliance.

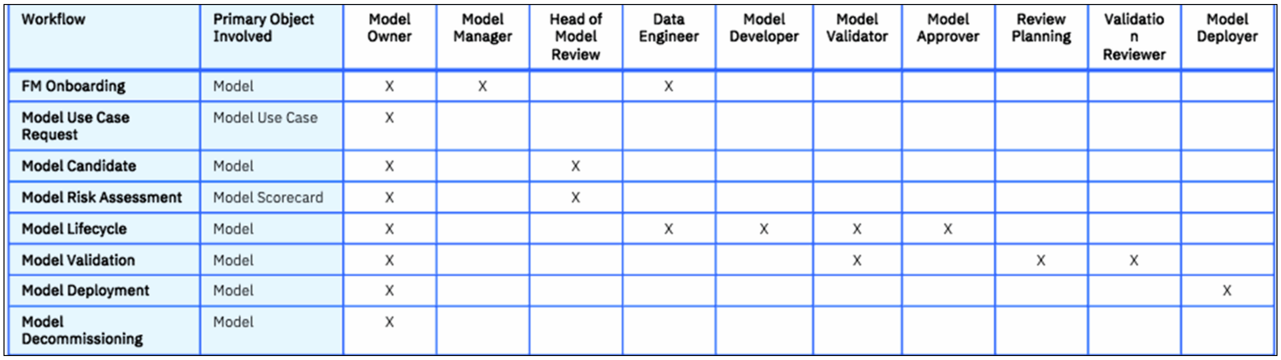
3.2.3 Workflows

The implementation of AI governance requires structured workflows to manage the lifecycle, risk, and compliance of AI models and their associated use cases. This section outlines key workflows provided by watsonx.governance Model Risk Governance (MRG), integrating them into the end-to-end governance process discussed in this chapter. These workflows facilitate responsible AI development, deployment, and monitoring, ensuring adherence to organizational processes and regulatory requirements. These workflows can be automated and involve multiple user interactions and feedback/approval capturing mechanisms effectively reducing the time and effort to manage the entire process of AI governance. All actions taken by users with workflows are auditable ensuring confidence in the system results. Predefined workflows are provided as part of watsonx.governance, but more can be built or existing workflows can be customized through the UI designer.

These available workflows can be categorized into:

- **Models**: Model Candidate, Model Validation, Model Deployment, Model Risk Assessment, Model Attestation, Model Decommission, and Model Change Request
- **Use Cases**: Use Case Request, Use Case Stakeholder Review, Use Case Development and Validation, and Use Case Deployment Approval
- **Metrics**: Metric Value and Metric Value Creation
- **Questionnaires**: AI Assessment and Questionnaire Assessment
- **Other**: Challenges and Model Risk Assessment

Figure 3-5 illustrates a matrix of various objects and personas involved in AI governance workflows. Each row and column represent different components and stakeholders discussed in 3.2, "Elements of model risk governance" on page 29, showcasing their interconnected roles in executing standardized processes. This visualization underscores the necessity of collaboration and coordination among diverse stakeholders to ensure the effective implementation of AI governance workflows.



Workflow	Primary Object Involved	Model Owner	Model Manager	Head of Model Review	Data Engineer	Model Developer	Model Validator	Model Approver	Review Planning	Validatio n Reviewer	Model Deployer
FM Onboarding	Model	X	X		X						
Model Use Case Request	Model Use Case	X									
Model Candidate	Model	X		X							
Model Risk Assessment	Model Scorecard	X		X							
Model Lifecycle	Model	X			X	X	X	X			
Model Validation	Model	X					X		X	X	
Model Deployment	Model	X									X
Model Decommissioning	Model	X									

Figure 3-5 Workflows bridging governance and personas for standardized processes

Benefits of structured workflows

By integrating these workflows into the AI governance process, organizations can:

- Ensure transparency, accountability, and ethical compliance throughout the AI lifecycle.
- Mitigate risks effectively while adhering to regulatory requirements.
- Foster trust in AI systems through robust and scalable governance practices.
- Minimize efforts to govern the AI use cases in their entirety.
- Standardize the approval and data capture process for all the use cases.

These workflows are critical in enabling organizations to manage their AI models and use cases with confidence and precision, aligned with the macro and micro-level governance strategies discussed in this chapter.

3.3 Considerations to implement AI governance strategy

Implementing an effective AI governance strategy requires careful consideration of an organization's unique characteristics and needs. This section outlines the key factors to consider when configuring an AI governance solution to meet the specific requirements of an individual organization.

3.3.1 Understanding organizational characteristics

When implementing an AI governance strategy, it is essential to consider the following factors that are specific to each organization:

- **Geographies**: Different regions have distinct regulatory requirements, cultural nuances, and market conditions that impact AI governance.
- **Market sectors**: Various industries, such as banking, telecommunications, and public sector, have unique challenges and requirements for AI governance.
- **AI use cases**: Organizations may employ different types of AI, such as machine learning or generative AI or general purpose AI, for internal or external purposes, which affects governance needs.
- **Organizational structure**: Centralized or decentralized structures influence how AI governance is implemented and managed.
- **Tech stack**: The use of specific platforms, AI in enterprise apps, or open-source technologies impacts AI governance requirements.

3.3.2 Configuring AI governance

Based on these organizational characteristics, the following steps can be taken to configure an AI governance solution:

- **Implementing legal obligations**: Develop a library of mandates and assessment templates that reflect the organization's specific legal requirements.
- **Establishing a risk framework**: Create a library of risks and controls, and assessment templates that align with the organization's risk management policies.
- **Defining organizational structure**: Configure business entities, roles, and responsibilities to match the organization's structure.
- **Manage collaboration and access control**: Use roles and access control features to ensure that team members have appropriate access to meet governance goals
- **Developing policies and procedures**: Create workflows for lifecycle governance, risk management, and compliance management that reflect the organization's policies and procedures.
- **Develop a communication plan**: Establish a plan for communication and decision-making, including the use of email, messaging tools, or other collaboration platforms
- **Implement a simple governance solution**: Start with a basic implementation and build incrementally to a more comprehensive solution
- **Plan for more complex solutions:** Consider extending the AI governance implementation to include external models, custom properties, and tailored reports

3.3.3 Leveraging out-of-the-box product content

While configuration is necessary, it is essential to note that out-of-the-box watsonx.governance capabilities and product content can provide a solid foundation for AI governance. This content can be used to illustrate best practices and provide a starting point for customization, rather than requiring organizations to start from scratch.

By considering these factors and configuring an AI governance solution accordingly, organizations can establish a strong foundation for effective AI governance and set the stage for successful implementation of subsequent chapters' topics.

3.3.4 Example use case

When implementing an AI governance strategy, an organization should start by analyzing various factors including: regions where it operates and utilizes AI models, the industry it serves, the specific AI tools and applications it relies on, its organizational structure, and its existing technology stack (for example, IBM Watson Studio). This evaluation will help the organization understand its operational landscape and tailor its AI governance strategy to address its specific needs. A well-designed governance strategy enables the organization to manage cross-team collaboration efforts, control access to sensitive data, and ensure compliance with regional regulations. Additionally, it helps support the implementation of a clear communication, aligning stakeholders and providing a strong foundation for continuous monitoring, auditing, transparency, and accountability over time.

4

Onboarding a new foundation model

The pace of innovation in artificial intelligence (AI) is high, especially in the development of newer and better foundation models. New versions of major models are regularly released, and numerous smaller models are created for specific languages, such as Danish; data types like geospatial; or business domains, such as IT programming. Additionally, a vibrant open-source community exists, with over one million models available on Hugging Face, an online marketplace for open-source AI models.

That increasing choice in models means there is an increasing need for organizations to govern those choices.

- Organizations should define the minimal standards a foundation model needs to meet before they will allow it to be applied in their use cases. This chapter describes the key considerations from the point of view of different stakeholders.
- These considerations might require a trade-off decision when the minimal standards are in conflict (for example, would you accept a new model that provides better performance on a certain business task, but it creates a copyright infringement risk if it's not clear where the training data was sourced?). Organizations should define who gets to make the decision and how it gets documented.
- Lastly, organizations should define the worfklow to automate the evaluation of a new model candidate by the various stakeholders.

This chapter has the following sections:

4.1 Key considerations to onboard a foundation model

When onboarding a foundation model, several key aspects must be considered to ensure a smooth and effective process. The following subsections outline the crucial considerations for data acquisition and preparation, data processing and filtering, model evaluation and validation, model security and robustness, and model performance monitoring.

4.1.1 Data transparency

Transparency is highly desirable as it makes information available, shareable, legible, and verifiable. In the context of training a foundation model, which involves multiple stages, transparency efforts are often targeted at different parts of the pipeline. Documentation is particularly crucial for the data pile used for training the foundation model, where gathering information on data acquisition and preparation methodologies for foundation model training is essential.

There are frameworks, such as The Foundation Model Transparency Index, which evaluate the transparency of foundation models across several composite indexes, including data, data labor, data access, and others. These indexes collectively aim to assess the transparency of various aspects of model development, such as data usage, labor practices, computational resources, methodologies, and strategies to mitigate privacy and copyright concerns (Bommasani, 2024[1]).

Data scientists or AI Center of Excellence or Enterprise AI team onboarding a foundation model should leverage such assessments to ensure compliance with internal and external regulations and policies.

4.1.2 Model evaluation and validation

To ensure the reliability and safety of large language models, a comprehensive evaluation and validation strategy is required, incorporating the following key elements

- FMEval framework: Leverage the Foundation Model Evaluation Framework (FMEval) for systematic, reproducible, and consistent validation and evaluation of new large language models.
- Evaluation modes: Support both fine-tuning and prompting (in-context learning) evaluation, with readily available academic and business benchmarks.
- Content filtering: Implement robust content filtering mechanisms to detect and mitigate the generation of hate speech, abuse, profanity (HAP), pirated content, malware, and other undesirable outputs.

To evaluate foundation models, a data scientist should compare the foundation model to be onboarded with current state-of-the-art models using a wide range of standard benchmark evaluations across top-level categories, such as:

- human exams (MMLU, MMLU-Pro (Wang, 2024))
- common sense (OBQA, SIQA),
- reading comprehension (BoolQ, SQuAD 2.0),
- reasoning (ARC-C, GPQA),
- code (HumanEval),
- math (GSM8K),

[1] Bommasani, R. K. (2024). The Foundation Model Transparency Index v1.1 May2024. arXiv preprint.

- Hugging Face's Open LLM leaderboards

Additionally, it is important to evaluate for different functions such as tool calling, RAG patterns, and other target domains specific to organizational use-cases, as discovered and discussed by data scientists or AI Center of Excellence or Enterprise AI team.

4.1.3 Model security and robustness

Foundation models can be vulnerable to various security risks, including:

- Data poisoning: Attackers can manipulate training data to compromise model performance or inject malicious behavior.
- Model stealing (or extraction): Attackers can steal the model itself or its weights, allowing them to use it for malicious purposes or create competing products.
- Adversarial attacks: Attackers craft specific input data (adversarial examples) designed to mislead the model, causing incorrect or malicious outputs. This includes techniques like prompt injection, where carefully crafted prompts can manipulate the model's behavior.

> **Note:** AIR-Bench is a benchmark for evaluating robustness against adversarial attacks. It provides a standardized way to measure how well a model resists these attacks.

Red teaming (IBM, n.d.) is crucial for data scientists to identify model vulnerabilities. This involves ethical hackers attempting to elicit unintended and potentially harmful behavior from the model, such as generating undesirable content through adversarial attacks and prompt injection. Data scientists responsible for onboarding foundation models should employ a comprehensive approach, using a combination of internal and external, automated and manual red teaming techniques to thoroughly identify and mitigate weaknesses. Traditional security measures, such as data encryption (at rest and in transit), user access controls, and firewalls, are also essential.

4.1.4 Ensuring model health and performance

To guarantee that a model meets the desired output and performance expectations of end-users, data scientists should evaluate model health metrics relative to the use cases expected to be delivered with this foundation model. Two key metrics to track are latency and throughput.

Latency

Latency measures the time it takes for a model to process a scoring request. It is calculated by tracking the time elapsed between receiving a request and generating a response, typically measured in milliseconds (ms). This metric helps identify any delays or bottlenecks in the model's processing pipeline.

Throughput

Throughput measures the number of scoring requests and transaction records that a model can process per second. This metric indicates the model's ability to handle a high volume of requests efficiently.

The following are additional references:

- Bommasani, R. K. (2024). The Foundation Model Transparency Index v1.1 May2024. arXiv preprint.
- IBM. (n.d.). Responsible Use Guide.

- Wang, Y. X. (2024). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574.

4.2 Considerations for legal team for approving a new foundation model

The Legal Team must have special considerations when approving a new Foundation Model. What kind of license is attached to the foundation model? Who owns the model? Does indemnification apply to the new Foundation Model in review? This section will cover how watsonx.governance can be used to answer these questions and more.

4.2.1 Model licensing

Standard software licensing applies to all foundation models and is a great place for the legal team to begin. Once foundation models are published, they are packaged with a license which provides a great depth of detail as to how the model can and should be used. These licenses can be found at multiple places. In watsonx.governance, the model license can be found attached to Model cards. You can access the model card by using the hamburger menu and clicking through **Inventory** → **Models** → **Choose a specific model**. The license will be listed under Development/Training. The model license can also be found on the model card in watsonx.ai, under terms. Finally, before the acceptance of a foundation model onto the watsonx.governance platform, a legal team may obtain the model License through the official publication page. As of this writing, any model which is publicly available can have its license checked through the huggingface.co hub. Table 4-1 highlights several common model licenses.

Table 4-1 Some common Model Licenses

License Name	Summarization
Apache 2.0 License	A permissive free software license that allows users to use, modify, and distribute the licensed software, including in commercial products, without paying royalties to the original authors.
Llama Community License	Similar to Apache 2.0 but in addition, if the monthly active users of the products or services made available by or for the licensee is greater than 700 million monthly active users in the preceding calendar month, the licensee must request a license from Meta, which Meta may grant to the licensee in its sole discretion.
MIT License	Similar to the Apache 2.0 license. This license is succinct and permissive for ensuring open ownership of software.

Terms and conditions attached to AI assets

Not all AI assets come with a simple model license. For example, much of the popular OpenAI models are not covered under these licenses. In cases like these, additional questions must be answered. A legal team may want to sit directly with model providers and other teammates such as the data science team to gain a thorough understanding of the foundation model. Some of the questions to ask may be:

- What sources of data are used to train the model?
- What are the terms and conditions under which this model will be made available?
- Are there contractual limitations on the use cases that the model can be used for (facial recognition, military, any non-research use)?

- Are there indemnification limitations (such as, no indemnification for customer-facing use)?
- What is the carbon footprint of the training for this model?

By answering these and other questions, a legal team may make the educated decision as to whether or not to onboard the foundation model.

Intellectual property ownership in generative AI deployments

Beyond the ownership of the foundation models themselves, several critical intellectual property considerations arise when deploying generative AI. These concerns extend to the software used to package and deploy these models, as well as the input data provided and the output generated. Therefore, thorough discussions with legal counsel and the model/software provider are essential. Key questions to address include:

- Training Data Ownership and Licensing: Does the model provider possess clear legal rights (including ownership or valid licenses) to use all data incorporated into the model's training dataset? This is crucial for avoiding potential copyright infringement claims.
- Copyrighted Material Exclusion: What specific measures has the model provider implemented to identify and exclude copyrighted material from its training data? Understanding their methodology is vital for assessing the risk of IP issues.
- Customer Data Usage for Model Training: Does the model provider use customer data to further train or refine its models? If so, what mechanisms are in place to allow customers to opt out of such usage? Transparency and control over data usage are paramount.
- Fine-Tuned Model Ownership: If the model provider offers fine-tuning services, who retains ownership of the intellectual property rights in the resulting fine-tuned model: the provider or the customer? This needs to be clearly defined in contractual agreements.

These points should be carefully reviewed to ensure compliance with intellectual property law and protect the interests of all parties involved. More on this will be covered in section “Legal indemnification” on page 46.

Local and national law implications

AI legislation has advanced at different rates throughout the world. The only thing that is common across all government bodies is that they are all putting forth some effort to regulate AI usage. It is not in the scope of this book (and would be challenging) to document all of the world's efforts to regulate AI. This section will provide a brief introduction to two different methods that have been used to implement AI Regulation - State level, and National level. At the state level, we will quickly introduce how the United States has been implementing AI regulation. At a national level, we mention the EU AI act in Western Europe.

This section has two following subsections - Implications in the United States, and Implications in Western Europe. These two examples are meant to serve as a comparison to how a government may enforce its AI policies. Legal teams should consult their respective Local and National laws for specifics.

Implications in the United States

Before approving any foundation model, it is critical to review local and national regulations as it applies to the industry. In the US, many state laws are beginning to form which work to govern the usage of AI. The legal team should be aware of any relevant law and perform the appropriate questioning to cover these. Some examples of passed legislation in the US include:

- NYC Local Law 144 - Implemented in 2021, the first law in the United States requiring bias auditing against AI tools. New York has proposed more than 30 additional AI laws since then, with almost all still in Proposal or Failed as of this writing.
- Tennessee ELVIS Act - Signed into law on March 21, 2024. This law protects the voices of Artists from all disciplines against AI-generated deepfakes created without their permission.
- Colorado Consumer Protections for Artificial Intelligence Act - Signed into law May 17, 2024. The act requires high-risk AI systems to use reasonable safeguards to protect consumers from *algorithmic discrimination*.

In October of 2022, The United States Federal Government also published guidance which could serve for legal conversations under the name of “Blueprint for an AI Bill of Rights.” This document outlines important topics which should be considered by the larger team when onboarding a Foundation Model. As of this writing, the blueprint serves only as guidance and does not enforce any official Federal law.

Implications in Western Europe (The European Union)

For completeness of consideration, the legal team must review any candidate Foundation Model and how it relates to the EU AI Act. The Act defines a specific set of obligations for providers of general purpose AI models (as foundation models are called in its legal terminology). In addition, the European Union will provide an associated Code of Practice which is planned to be approved in or before May 2025. The obligations and the Code will give buyers some concrete expectation towards a provider of a model that's brought onto the market after the Act has gone into force (August 2024).

Note that foundation models already on the market before entry into force, need to be compliant by August 2nd, 2027. Vendors might decide to withdraw their model from the market before that date to avoid the compliance implications. Don't wait too long to engage your existing model providers to understand their intent and plan for compliance, and make sure you have a plan yourself to switch models when needed.

Users of foundation models have no specific obligations, beyond what might result from the AI system(s) that use that foundation model.

4.2.2 Legal obligations on the part of the vendor

Any vendor of AI is tied to the same local and national laws as mentioned earlier. In addition, most of the earlier section around model licensing also applies to the vendor. One more important topic left to cover is the topic of Legal Indemnification. In law, indemnification (or an indemnity) is any undertaking by one party to protect another party of some financial burden.

In AI, the usage of foundation models can expose parties to novel financial burdens. As such, a major effort has been put forth around the idea of legal indemnification from AI models. This next subsection will cover this point.

Legal indemnification

AI has been controversial in its effects and repercussions. Countless lawsuits have risen against AI companies, because of this, a growing need of consumers of AI is to be indemnified against any legal repercussions. Every AI producer has a different position on indemnification and therefore, specifics must be explored on a per-foundation model basis. By approaching the legal review of Foundation Models in a systemic way, the legal team can ensure that an AI program will be implemented with proper control and in a safe and legal way.

4.2.3 A final note on legal considerations

As was covered in the earlier section of the US Blueprint for AI Bill of Rights, not all measures of AI safety are required by law. Legal considerations typically follow the letter of the law, but this is not the end of the story for what dangers an un-controlled AI system could pose. Even with all relevant legal details being covered, the ethics behind any AI effort must still be questioned.

It is also important to note that many laws and regulations are vague in their definitions and guidance. Similar to other types of technology-focused requirements, those for AI require broad statements as to not grow stale shortly after going into effect given the current speed of evolution in the industry. For these reasons it is important to approach any requirement that is new or without established precedent with a conservative and ethical mindset.

The next section 4.3, “Ethical considerations for approving a new foundation model” on page 47 will focus on this topic of AI ethics.

4.3 Ethical considerations for approving a new foundation model

With AI usage gaining momentum, there is an increased focus on the ethical aspect of understanding how a foundation model is trained and any ethical risk it can expose if it is in used as part of an enterprise use case. When considering a new foundation model, an ethical stakeholder will want to consider at least the following dimensions.

4.3.1 Fairness

An AI system or a model should be fair and free of any direct or indirect bias in its prediction. With the AI system using a foundation model, the foundation model itself needs to be evaluated for fairness.

Bias can be detected by evaluating the foundation model for differences in accuracy and performance by using social demographic-protected attributes.

A model can exhibit fairness issues in various ways. If the model is looked at in isolation and without sufficient context, it might not appear to exhibit a favorable or unfavorable outcome. In other cases, the model output can decide the prioritization level and can introduce bias in the entire end-to-end process where the source of bias is tied to the model behavior.

It is critical that the business owners, stakeholders, and designers look at the model in the context of the overall ecosystem where the training, testing, or the production of data is generated, how the model is developed and evaluated, and where the model is being used to decide different ways in which the model might lead to a biased outcome. This is a key step in the overall methodology relating to bias detection.

It is important to note that bias in some attributes of a model might be reasonable. For example, a foundation model that is fine-tuned to pre-screen loan applications might be biased against people with poor credit. This is deemed as reasonable. Since most model bias is injected unwittingly through skewed data or constrained training methods, it cannot be effectively detected and resolved exclusively through manual testing or checklist validations.

4.3.2 Transparency

With AI usage gaining momentum, there is an increased focus on the transparency aspect of the foundation models for audit purposes. There are various governance, risk, compliance, or regulatory needs for information covering the nature and intended uses of the foundation model. Transparency about the model's overall accuracy, its ability to explain particular decisions, its fairness regarding protected classes, and information about the provenance of training data and assurances that suitable privacy protections have been maintained, all should be properly documented and available for audit purposes.

Transparency builds the trust in the AI system by increasing the understanding of how the model was created and deployed and enabling the ability to control how AI is created and deployed. This can prevent undesirable situations, such as a model training with unapproved data sets, models having biases, or models having unexpected performance variations.

This documentation of facts about the foundation model (for example model cards) can have the following properties:

- It can vary in content and are tailored to the particular foundation model being documented.
- It is tailored to the needs of their target audience or consumer documenting the tasks the foundation model can be used for.
- It captures the details about the data the foundation model has been trained on.
- It includes the information about training algorithms, parameters, fairness constraints or other applied approaches, and features.
- It shows the benchmark accuracy documenting the performance of the model.
- It contains information about the model owner (organization), model version, as well as the license model can be used with.

4.3.3 Privacy

An important aspect to build trust in an AI system or model is to take measures to manage and safeguard foundation models and its data that is trained on Personal Information (PI). If a model is trained on PI without applying any specific privacy techniques, and that model is made publicly available or shared with a nontrusted third party, the model might reach the wrong hands, and potentially violate the privacy of the people whose information was used in training it. PI must be properly handled and safeguarded wherever it is stored or used in the organization.

The same safeguards must be applied for models trained on proprietary, intellectual, copyrighted and confidential information.

Model providers must implement measures to prevent data leakage during inference. This includes:

- Data minimization: Process only the absolute minimum amount of user data necessary for the model to function.
- Differential privacy: Apply noise to user data or model outputs to enhance privacy and make it difficult to identify or isolate individual contributions.
- Federated learning: Train models collaboratively across multiple decentralized data sources using techniques like federated learning, thereby avoiding the risks associated with centralizing sensitive user data

- Model monitoring: Continuously monitor model behavior for signs of unexpected data leakage or privacy violations.
- Transparency and user control: Provide users with clear information about how their data is used by the model and offer options for controlling data access and sharing.

4.3.4 Explainability

AI systems are increasingly used to inform high stakes decisions. Explainability and interpretability of these systems and the models within the system are becoming essential. There are many ways to explain these models and systems, the appropriate choice depends on the usage context and type of explanation that is needed by the consumer of the explanation.

4.3.5 Robustness

An AI system is considered robust if it can continue to perform well and reliably even when faced with difficult or unexpected situations. These situations can include anything from slight changes in the data it receives to deliberate attempts to trick or manipulate the system. A robust AI system is designed to handle these challenges effectively, minimizing mistakes and providing consistent and trustworthy results.

4.3.6 Third-party help

There are third-party evaluations to help you understand various ethical characteristics of a foundation model you are considering using. Use these tools to speed up your model onboarding process.

Two examples from the Standard Center for Research on Foundation Models:

- Foundation Model Transparency Index - This index scores foundation models on 100 transparency indicators across several dimensions.
- AIR-Bench (short for “AI risk benchmark” - This benchmarks scores foundation models on an extensive safety taxonomy that covers content safety risks, societal risks, legal and rights-related risks, and system and operational risks.

There are many other options available. A quick web search will give you plenty of options to consider.

The IBM AI Ethics Board has published an extensive overview of AI risks, called the AI Risk Atlas. It lists several specific risks in the ethics dimensions listed in this paragraph. Each risk comes with a description, categorizations, and links to examples of that risk in third-party publications whenever possible.

4.4 Considerations for financial stakeholders for approving a new foundation model

A stakeholder from a Finance function will want to consider at least the following dimensions when considering a new foundation model.

4.4.1 Total cost of ownership

First, what is the sum total of the costs associated with acquiring and using another foundation model? Many models are available under a “free” license, but there's more to the total cost of ownership than that.

What is the initial investment?

- What are the initial license costs to acquire the foundation model? This might take the form of a license for a model specifically, but could (also) include licenses for a platform that the model is hosted on. Based on existing vendor relationships, discounts might be available. As the market for generative AI evolves, new pricing models might appear.
- What are the costs to onboard this new model? Consider procurement, IT, legal, security and other cost components. The assessment that is the topic of this chapter is part of such an onboarding process and comes with a cost.
- What are the costs of skills development to use this new model effectively? Consider training for administrators, data scientists, risk management and other relevant roles.

What are the run costs?

- If the model is deployed as SaaS, what are the charges for inferencing? Foundation model inferencing is typically charged by the number of calls to the model and/or the number of tokens processed/generated by the model. Pricing models might differ, as does the amount charged for comparable numbers of calls/tokens.
- If the model is deployed in-house, what are the total costs over the foreseeable lifetime of this investment? Consider hardware, electricity, cooling, personnel and all other relevant cost components. These costs will depend on the technical specifications of the model, such as the number of parameters or the model architecture.

4.4.2 Return on investment

Second, how will the investment in another foundation model pay back for itself?

Existing use cases

Will this foundation model allow us to execute existing use cases better/cheaper/faster? How? How much? How well does it align with our strategic imperatives? Consider the following factors, and the interplay between them:

- Improve the accuracy of an existing application.
- Improve legal indemnification for an existing application.
- Improve the ease of governing one or more existing applications.
- Reduce AI ethics risk for an existing application.
- Reduce inferencing costs for an existing application.
- Reduce energy consumption for an existing application.

New use cases

Will this foundation model allow us to enable new use cases? How? How much? How well does it align with our strategic imperatives? For example:

- The new model has been tuned to handle IT optimization use cases better than our existing models, so we can now enable our IT developers to use generative AI and enhance their productivity by X%.

- The new model has been tuned to handle the Dutch language better than our existing models, so we can now enable a conversational assistant for our Dutch customers and improve first-time resolution by Y%.

4.4.3 Build or buy

Instead of creating a solution ourselves with a new foundation model, can we buy tooling that already does that? As more applications become AI-enabled, the benefits of buy rather than build could be:

- Achieve the projected benefits faster with a turn-key application.
- Enhanced functionality from a specialist vendor compared to what we could deliver in a first stage ourselves.
- Focus scarce internal AI resources on the most strategic AI projects.

4.4.4 Exit strategy

The speed of innovation in AI is very high, how easily can we change course if something better comes along? Consider aspects such as:

- Contractual cancellation periods.
- Non-recoverable license, support or services costs already committed.
- Early termination fees.
- Charges to extract our data/IP/solutions from a vendor's platform.
- Portability of assets and skills to a new solution.

4.4.5 Other factors

Lastly, there might be factors not yet considered by other stakeholders that would fall to the finance team in an organization. For example:

- Risks - many risks will already have been addressed by other stakeholders as described in the previous paragraphs in this chapter, but there might be specific ones to be addressed by the Finance function.
- Sustainability - how does a new foundation model impact the organization's ESG (environmental, social, and governance) posture? How does it impact our direct and indirect emissions or freshwater usage?
- Security - How does the model vendor protect the organization against adversarial attacks, data leaks and other security risks?

5

Assessing a new use case

A use case in watsonx.governance is the starting point to solve a business problem using an Artificial Intelligence (AI) asset, such as a model or prompt template as part of the solution. The process of assessing a new use case for an organization involves following a business process facilitated by a workflow engine. This process identifies risks, assesses applicable regulations and policies, and decides whether to approve or reject the use case from development through production to decommissioning.

This chapter covers the following topics:

- “Business process workflow” on page 54
- “Approval workflow” on page 55
- “Risk identification assessment” on page 56
- “Applicability assessment” on page 58

5.1 Business process workflow

The typical business process to assess a new use case, shown in Figure 5-1, combines automated workflows in watsonx.governance with manual checks and balances to ultimately approve or reject a use case and record the transparent process and findings along the way. This process can always be customized to fit an organization's specific needs.

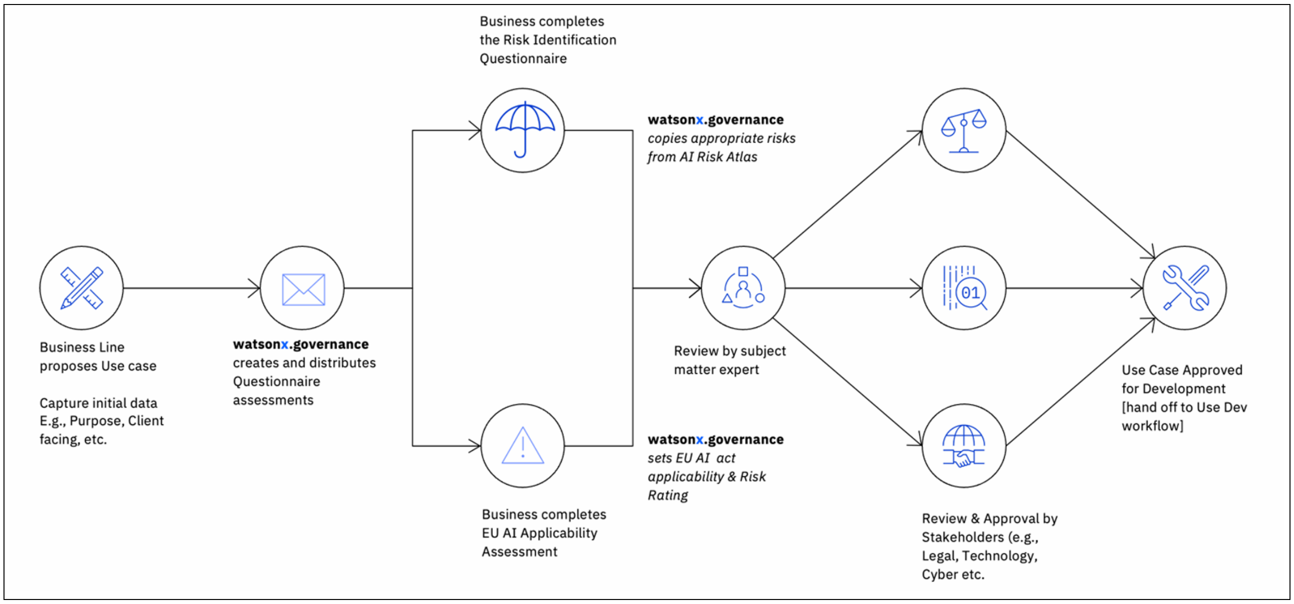


Figure 5-1 Typical use case assessment process workflow leveraging watsonx.governance

The typical process flows as follows:

- Propose the use case: The process starts with a member of the business proposing a use case within the watsonx.governance console. The creator could be any member of the business who originates the problem to be solved. Typically, this would be a member of a business line, but could also be technical depending on the problem to be solved.
- Questionnaire Assessments: Once the use case is created, watsonx.governance creates assessment questionnaires to be completed by the business line and necessary technical teams. The answers to these questions allow watsonx.governance to auto-populate the use case with applicable risks from the IBM AI Risk Atlas that will need to be addressed during the life of the use case.
- Applicability Assessments: It is important to ensure your use case will remain compliant with external regulations, such as the EU AI Act in Europe. Once a use case is created, watsonx.governance will create an EU AI Act applicability assessment to determine if the use case is complaint or at risk of violating the mandates of this regulation. While the EU AI Act Assessment is installed with watsonx.governance, additional assessments are available from IBM partners for regulations that may be applicable in other regions.
- SME Review: Once the use case and the appropriate assessments are completed, a subject matter expert (SME) will review the auto-populated risks and further enrich the use case with mandates, processes, and policies that could be affected by the identified risks. The SME can also setup controls to address issues that arise after the use case is approved for the next steps in the development lifecycle. watsonx.governance can help to automate the assignments of the mandates, processes, policies and controls through custom workflow created to accommodate the specific needs of the organization.

- Stakeholder Review & Approval: Once SME review is complete, required business stakeholder will be automatically notified by watsonx.governance to review the use case and approve or reject the case. They will also be asked to provide comments to explain the review decision. The number of stakeholders notified can be automated via a watsonx.governance workflow or pre-defined in a use case template depending on organization needs.
- Use Case Approved for Development: If all stakeholders approve of the use case, watsonx.governance marks it as “Approved for Development”. This provides audited authorization for development of the solution to begin and marks the end of the initial use case assessment.

5.2 Approval workflow

Once a use case is created in the watsonx.governance console, a use case approval workflow is triggered by the watsonx.governance console, as shown in Figure 5-2.

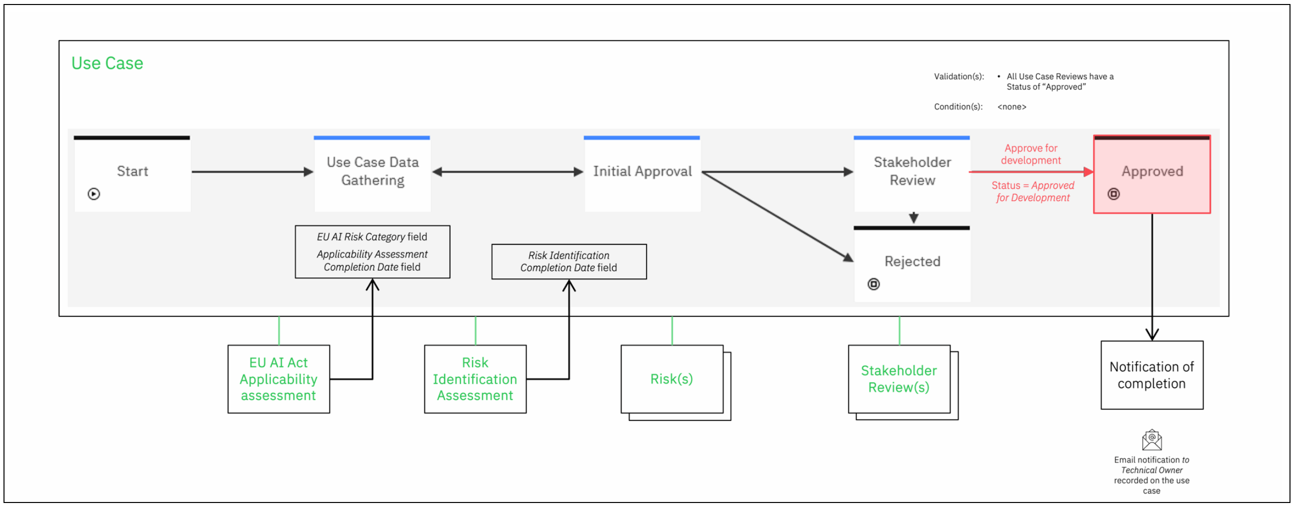

Use Case
Validation(s): • All Use Case Reviews have a Status of “Approved”
Condition(s): <none>
Start
Use Case Data Gathering
Initial Approval
Stakeholder Review
Approve for development
Status = Approved for Development
Approved
Rejected
EU AI Risk Category field
Applicability Assessment Completion Date field
Risk Identification Completion Date field
EU AI Act Applicability assessment
Risk Identification Assessment
Risk(s)
Stakeholder Review(s)
Notification of completion
Email notification to Technical Owner recorded on the use case

Figure 5-2   The default use case approval workflow executed in the watsonx.governance console

The use case approval workflow and a review and approval process that collects key information about the use case and captures the approval or rejection of the use case from the identified stakeholders, guiding those users through all required actions along the way.

While any workflow can be customized to the needs of any organization, the steps in the default use case approval workflow are:

- Start: This step in the flow begins the moment a new use case is created. The creator is required to provide a name for the use case, and use case owner, a description of the use case, and a primary Business Entity. Optionally, the creator may also provide the purpose of the use case and its type. There are two default use case types: AI or Non-AI.
- Use Case Data Gathering: Once the use case is created, the approval workflow moves to gather more detailed information about the use case to be governed. At this stage, the use case owner or risk assessor can provide the initial assumed risk level, identify the stakeholders who will provide final approval, and identify the technical owner of the use case. It must also be determined if the use case will leverage foundation models during this stage in the workflow. Once all the required information is complete, the user can submit the case for Initial Approval.

- Initial Approval: In this stage, the use case owner completes a Risk Identification assessment questionnaire and validates a series of risks automatically assigned based on answers provided. In addition, the use case owner can optionally complete and applicability assessment questionnaire for the EU AI Act. Completion of the Risk Identification assessment is required to move to the Stakeholder Review stage.
- Stakeholder Review: At this stage, the stakeholders identified in the Use Case Data Gathering stage are notified to review the use case details, provide comments and approve or reject the use case for development. The use case cannot be approved unless all stakeholders have provided approval. Once all stakeholders have approved the use case, it is placed in “Approved for development” status. The owner and technical owner of the use case are automatically notified when the use case has been approved for development.

5.3 Risk identification assessment

As part of the Initial Approval stage of the use case assessment workflow, the use case owner or risk assessor completes a risk identification assessment created by an assessment workflow in the watsonx governance console, as seen in Figure 5-3.

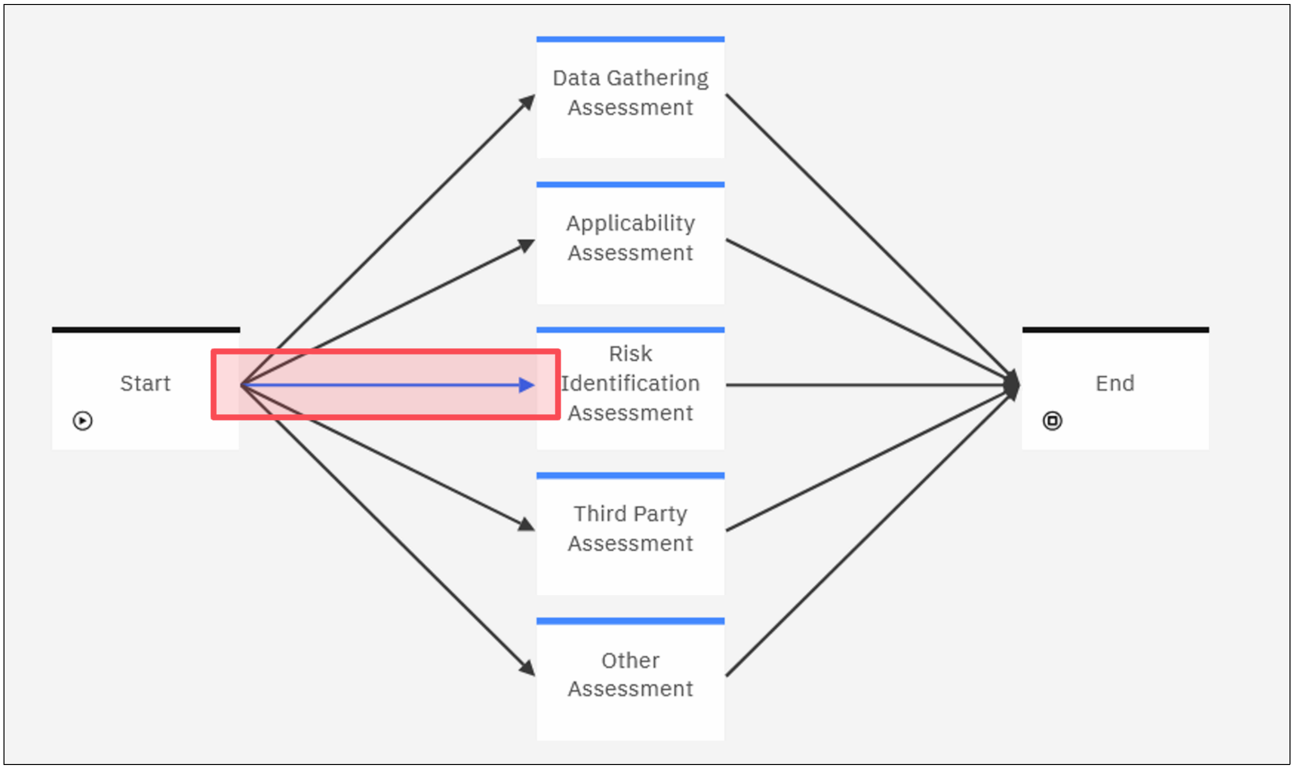
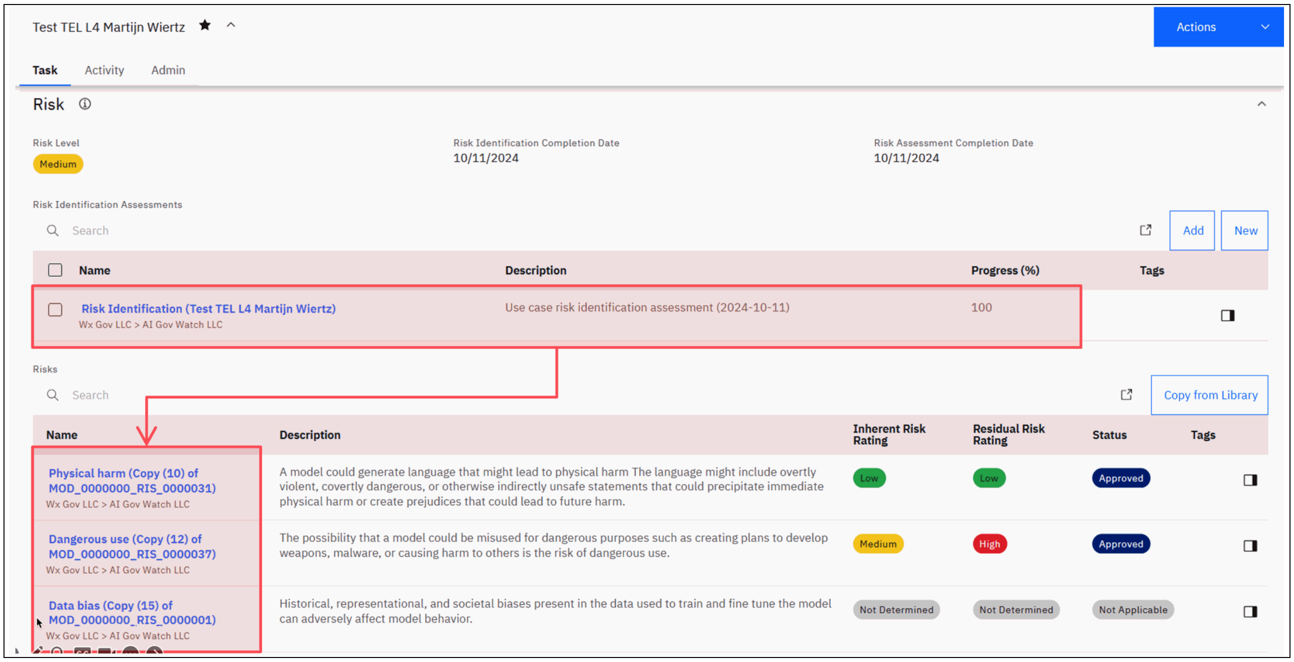


Figure 5-3 Risk Identification Workflow as seen in the watsonx.governance console

A risk identification assessment is a dynamic questionnaire that identifies applicable risks from the IBM Risk Atlas based upon answers provided as seen in Figure 5-4 on page 57.



Test TEL L4 Martijn Wiertz ★

Actions

Risk

Risk Level	Risk Identification Completion Date	Risk Assessment Completion Date
Medium	10/11/2024	10/11/2024

Risk Identification Assessments

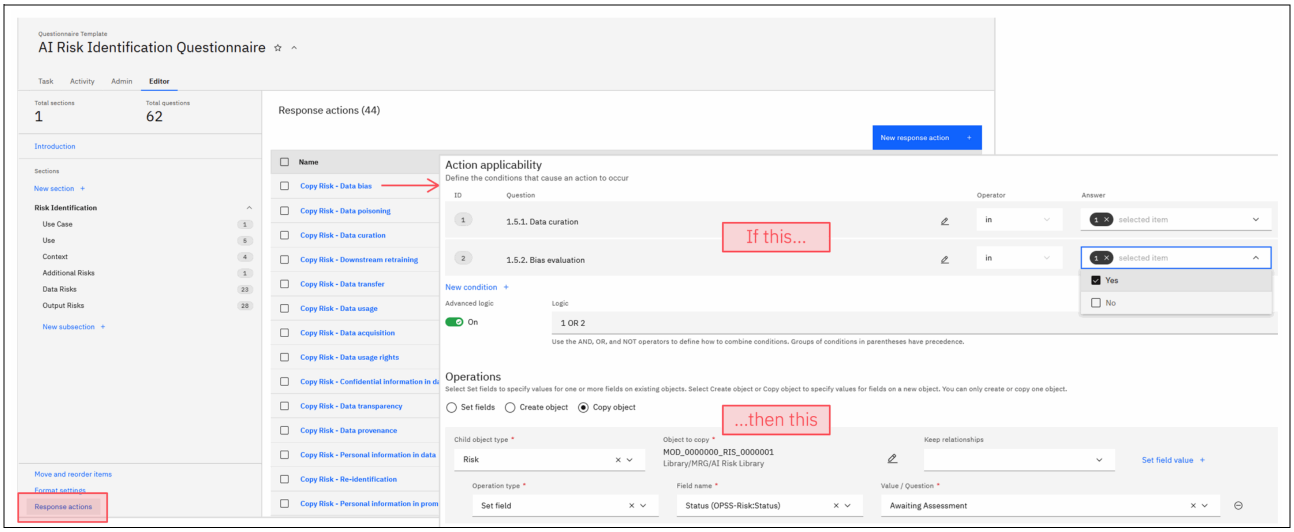
Name	Description	Progress (%)	Tags
Risk Identification (Test TEL L4 Martijn Wiertz) Wx Gov LLC > AI Gov Watch LLC	Use case risk identification assessment (2024-10-11)	100	

Risks

Name	Description	Inherent Risk Rating	Residual Risk Rating	Status	Tags
Physical harm (Copy (10) of MOD_0000000_RIS_0000031) Wx Gov LLC > AI Gov Watch LLC	A model could generate language that might lead to physical harm The language might include overtly violent, covertly dangerous, or otherwise indirectly unsafe statements that could precipitate immediate physical harm or create prejudices that could lead to future harm.	Low	Low	Approved	
Dangerous use (Copy (12) of MOD_0000000_RIS_0000037) Wx Gov LLC > AI Gov Watch LLC	The possibility that a model could be misused for dangerous purposes such as creating plans to develop weapons, malware, or causing harm to others is the risk of dangerous use.	Medium	High	Approved	
Data bias (Copy (15) of MOD_0000000_RIS_0000001) Wx Gov LLC > AI Gov Watch LLC	Historical, representational, and societal biases present in the data used to train and fine tune the model can adversely affect model behavior.	Not Determined	Not Determined	Not Applicable	

Figure 5-4 Dynamic assignment of use case risks based on risk assessment answers

Risks assigned are determined based on a “if this, then that” selection structure as seen in Figure 5-5.



Questionnaire Template

AI Risk Identification Questionnaire ☆

Total sections	Total questions
1	62

Risk Identification	
Use Case	1
Use	5
Context	4
Additional Risks	1
Data Risks	23
Output Risks	28

Response actions (44)

Name
Copy Risk - Data bias
Copy Risk - Data poisoning
Copy Risk - Data curation
Copy Risk - Downstream retraining
Copy Risk - Data transfer
Copy Risk - Data usage
Copy Risk - Data acquisition
Copy Risk - Data usage rights
Copy Risk - Confidential information in da
Copy Risk - Data transparency
Copy Risk - Data provenance
Copy Risk - Personal information in data
Copy Risk - Re-identification
Copy Risk - Personal information in prom

Action applicability

Define the conditions that cause an action to occur

ID	Question	Operator	Answer
1	1.5.1. Data curation	in	1 selected item
2	1.5.2. Bias evaluation	in	1 selected item

☑ Yes
☐ No

If this...

Advanced logic: On

Logic: 1 OR 2

Use the AND, OR, and NOT operators to define how to combine conditions. Groups of conditions in parentheses have precedence.

Operations

Select Set fields to specify values for one or more fields on existing objects. Select Create object or Copy object to specify values for fields on a new object. You can only create or copy one object.

○ Set fields ○ Create object ● Copy object

...then this

Child object type *: Risk

Object to copy *: MOD_0000000_RIS_0000001 Library/MRG/AI Risk Library

Keep relationships:

Operation type *	Field name *	Value / Question *
Set field	Status (OPSS-Risk:Status)	Awaiting Assessment

Figure 5-5 Dynamic risk applicability assignment selection structure rules

Once assigned, each risk must be evaluated and approve or rejected for applicability for the use case. These risks are the primary bases for assessing and managing risk throughout the lifecycle of your use case. While risks are assigned from the IBM Risk Atlas by default, additional risks or risk sources can be added as required by an organization.

Once all assigned risks have been approved or rejected, the use case can be approved for stakeholder review and defined in the primary use case approval workflow.

5.4 Applicability assessment

Applicability assessment, as seen in Figure 5-6, should also be completed by the use case owner as part of the Initial Approval stage of the use case assessment workflow.

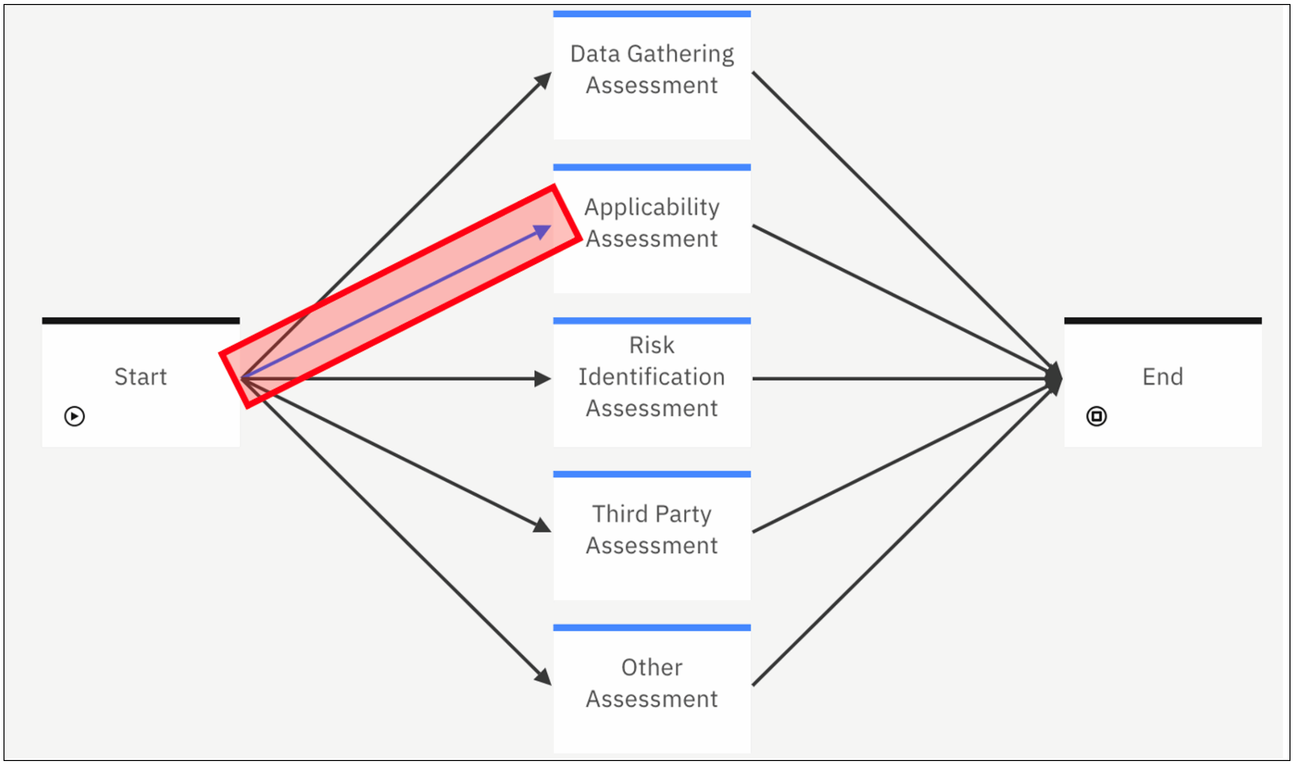
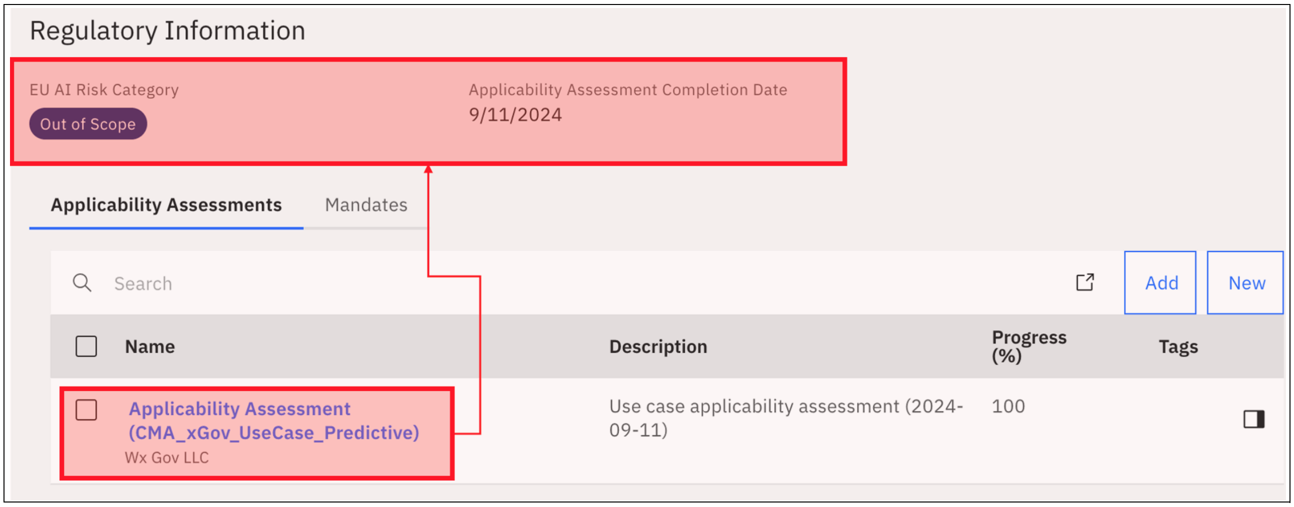


Figure 5-6 Applicability Assessment Workflow as seen in the watsonx.governance console

This assessment enables the use case owner to assess their AI use cases using a simple dynamic questionnaire that aids in determining whether a use case is in scope for government regulations (such as the EU AI Act) and which risk category the use case aligns to (Prohibited, High, Limited, Minimal, Out of Scope) as seen in Figure 5-7.



Regulatory Information

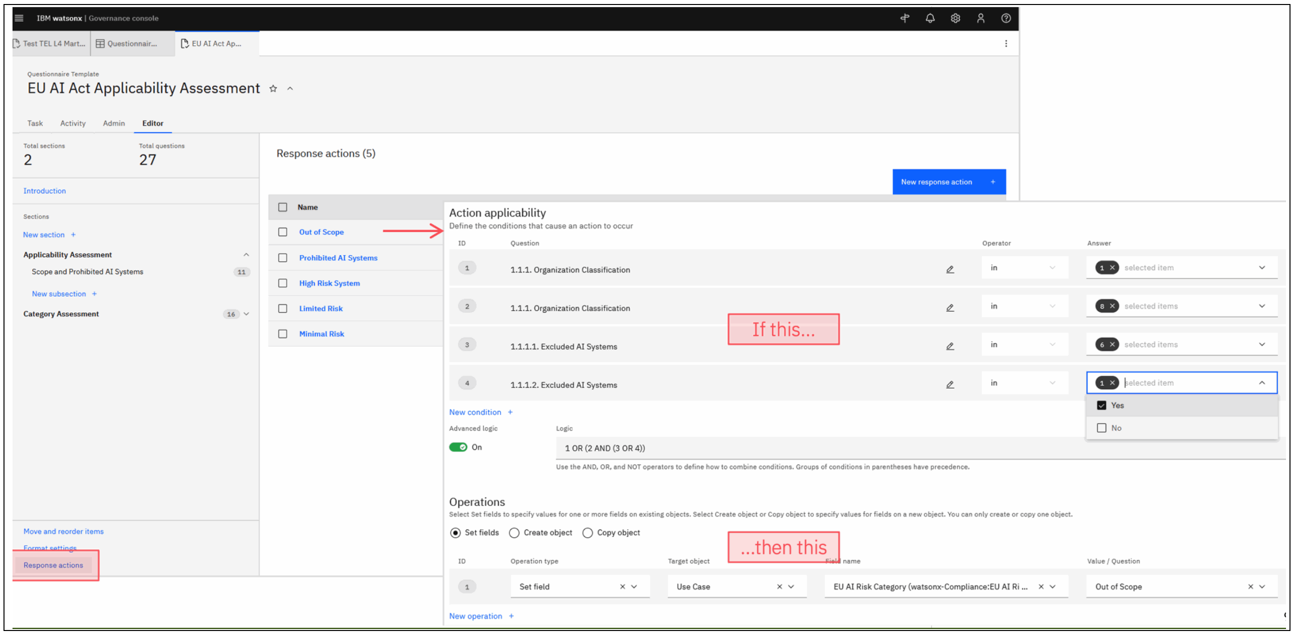
EU AI Risk Category: Out of Scope

Applicability Assessment Completion Date: 9/11/2024

□	Name	Description	Progress (%)	Tags
□	Applicability Assessment (CMA_xGov_UseCase_Predictive) Wx Gov LLC	Use case applicability assessment (2024-09-11)	100	

Figure 5-7 Dynamic assignment of risk category based on applicability assessment

The questionnaire to determine the risk category follows a “if this then that” selection structure as seen in Figure 5-8.



Questionnaire Template

EU AI Act Applicability Assessment

Total sections: 2

Total questions: 27

Response actions (5)

□	Name
□	Out of Scope
□	Prohibited AI Systems
□	High Risk System
□	Limited Risk
□	Minimal Risk

Action applicability

Define the conditions that cause an action to occur

ID	Question	Operator	Answer
1	1.1.1. Organization Classification	in	1 selected item
2	1.1.1. Organization Classification	in	8 selected items
3	1.1.1.1. Excluded AI Systems	in	6 selected items
4	1.1.1.2. Excluded AI Systems	in	1 selected item

☑ Yes
☐ No

Advanced logic: On

Logic: 1 OR (2 AND (3 OR 4))

Use the AND, OR, and NOT operators to define how to combine conditions. Groups of conditions in parentheses have precedence.

Operations

Select Set fields to specify values for one or more fields on existing objects. Select Create object or Copy object to specify values for fields on a new object. You can only create or copy one object.

● Set fields ○ Create object ○ Copy object

ID	Operation type	Target object	Field name	Value / Question
1	Set field	Use Case	EU AI Risk Category (watsonx-Compliance:EU AI Ri ...	Out of Scope

Figure 5-8 The questionnaire to determine the risk category

6

Governing the end-to-end lifecycle of an AI asset

Lifecycle governance is a major pillar of the IBM AI Governance framework. This pillar represents the idea that any AI asset should be tracked throughout the lifetime of its usage, without gaps in lineage or traceability.

Regardless of where the AI asset originates from, watsonx.governance comes with tooling to monitor performance and also to monitor the lifecycle stage that the asset is in. Lifecycle governance applies to traditional ML/AI use cases as well as generative AI.

End-to-end lifecycle governance results in increased efficiency through automated evaluations of models as well as the publishing of evaluation results and metrics. This also introduces traceability, auditability, and enhanced project management capabilities through transparent ML/AI asset lifecycle management functionality.

This chapter has the following sections:

6.1 What is the AI lifecycle?

Before reading through how watsonx.governance defines the lifecycle of an AI asset, it is useful to review the general lifecycle for data projects.

Figure 6-1 represents the CRISP-DM methodology which encapsulates an industry-standard blueprint for completing data mining or data science projects.

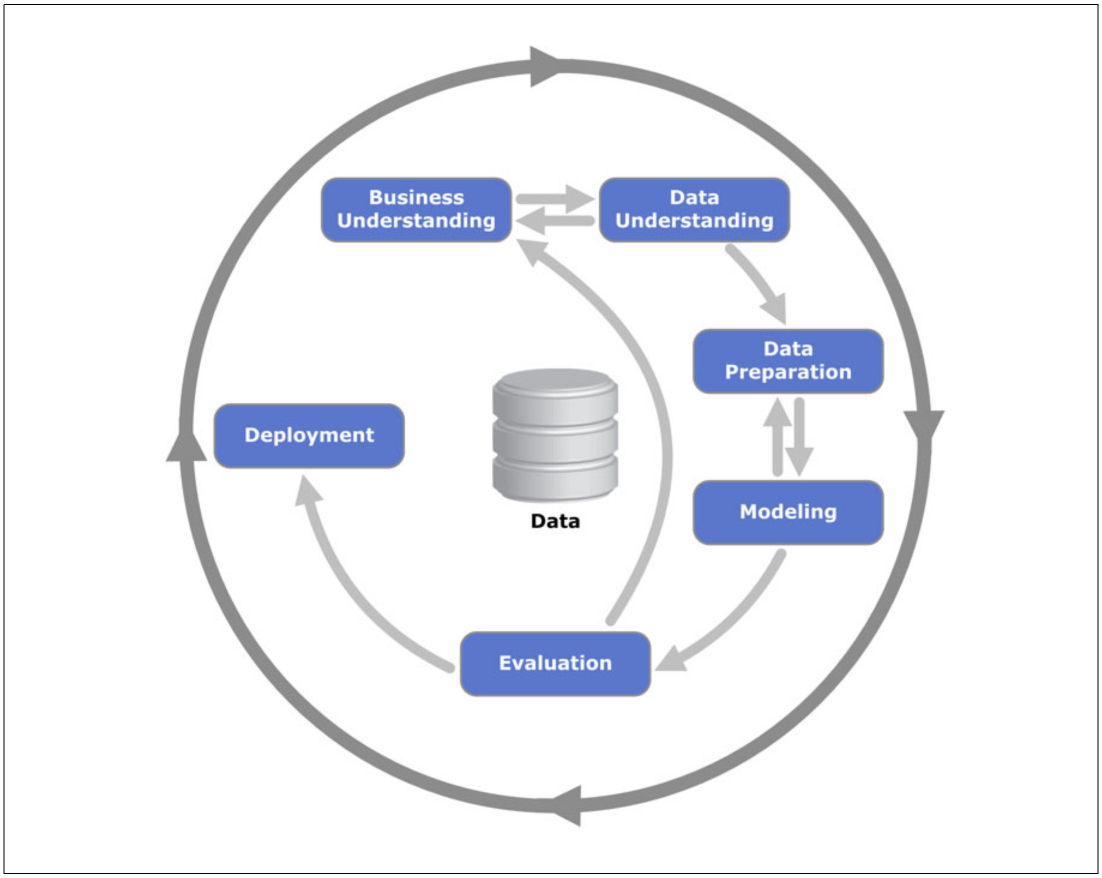
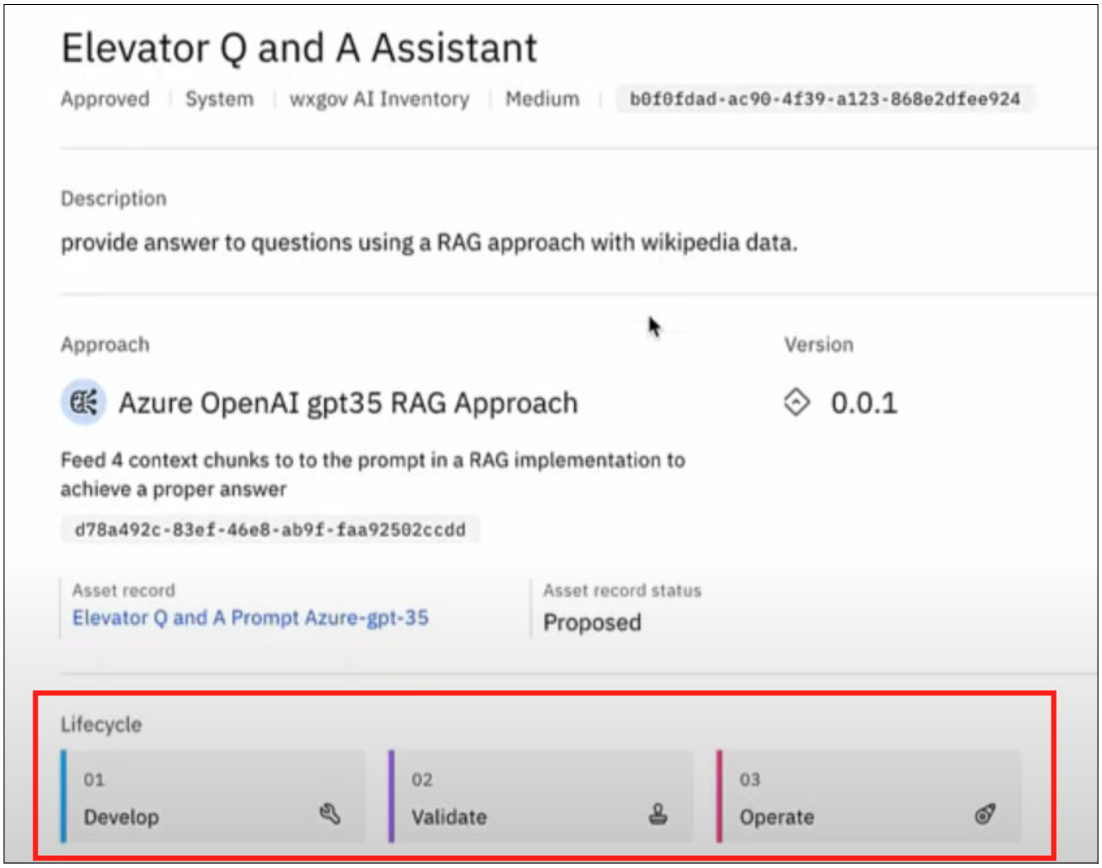


Figure 6-1 The CRISP-DM Methodology (Cross-industry Standard Process for Data Mining)

Because AI is a form of machine learning (albeit a very large and complex version), every step of the CRISP-DM applies to the AI lifecycle during initial foundation model development (commonly referred to as model pretraining.) Differences begin to appear when we consider an AI asset after it has been developed. It is unreasonable to expect every enterprise to train their own model from scratch. For this reason, in virtually every case, citizen engineers and scientists will engage with AI assets which have already been pre-trained, and therefore, a modified version of the CRISP-DM methodology should be used when considering the lifecycle of an AI asset.

The pre-trained nature of foundation models allows us to create a more focused view on the lifecycle of AI assets being used in use cases. This lifecycle can be generally depicted with three chronological stages: *Develop stage, Validate stage, and Operate stage*.

Figure 6-2 on page 63 highlights the general lifecycle stages in a red box.



Elevator Q and A Assistant

Approved | System | wxgov AI Inventory | Medium | b0f0fdad-ac90-4f39-a123-868e2dfee924

Description

provide answer to questions using a RAG approach with wikipedia data.

Approach

Azure OpenAI gpt35 RAG Approach

Feed 4 context chunks to to the prompt in a RAG implementation to achieve a proper answer

d78a492c-83ef-46e8-ab9f-faa92502ccdd

Version

0.0.1

Asset record

Elevator Q and A Prompt Azure-gpt-35

Asset record status

Proposed

Figure 6-2 A Sample Lifecycle Diagram

Within each stage of this AI lifecycle, actions should be taken by users to progress the AI asset through its lifecycle. The following section covers the high-level activities that take place within each of the three stages.

> **Note:** In traditional ML use cases, we also split use Testing stages in addition to validation stages. This will be covered in the considerations section for ML.

Development stage

In the development stage, users must work to stand up the initial AI use case solution. As described in Chapter 4, “Onboarding a new foundation model” on page 41 and Chapter 5, “Assessing a new use case” on page 53, the AI use case will go through an approval process before being worked on by the appropriate practitioners. Once the approval process is complete, an engineer may begin to develop the technical assets for the AI use case. The engineer can achieve many things in this stage. Prompt engineering, parameter tuning, and solution experimentation all fall within the development stage. For traditional ML, this stage also includes activities such as feature engineering, exploratory data analysis, and other pre-processing tasks which go into the model development process.

Validation stage

After development efforts are complete, the validation stage of the life cycle is started. This stage represents the process of putting development efforts through testing. An independent practitioner such as an AI engineer or data scientist can use IBM's watsonx.governance to run automated Prompt Evaluations, which will return evaluation metrics for the given AI use case. A programmer can also use custom code to run evaluations.

Similarly, for traditional ML, validation is achieved through the traditional data science approach of utilizing validate/test datasets to understand model performance.

Using watsonx.governance, the validation stage represents our first opportunity at automating otherwise lengthy evaluations of ML and AI models. It is the first glance of how our models may perform against performance metrics such as readability for Q&A LLM use cases, or accuracy for classification use cases.

Operation stage

The operation stage represents any AI or ML asset which reaches production-level efforts. In this stage of the AI lifecycle, the model asset is live and consumed by users. Model monitoring and asset lineage become central at this stage. The IBM watsonx.governance platform provides tools for monitoring critical asset metrics such as drift, fairness, and bias.

Before diving into implementation, the next section will briefly review the different kinds of metrics that can be utilized during lifecycle governance with watsonx.governance.

6.2 Metrics in watsonx.governance

The metrics in this section are by no means exhaustive and watsonx.governance is capable of implementing custom monitors and metrics based on user needs. The ability to add custom monitors and metrics sets watsonx.governance apart from its competitors in the market. This section gives an overview to some of the most common and important kinds of metrics that can be applied to assets for effective lifecycle monitoring.

6.2.1 Drift detection

Users can configure Drift v2 evaluations in watsonx.governance to measure changes in their data over time to ensure consistent outcomes for the model. These evaluations can be used to identify changes in the model output, the accuracy of your predictions, the distribution of the input features, the metadata and more.

The drift in the user deployments is always detected in comparison to a baseline data. This baseline data needs to be a good representation of the ideal dataset that the user is expecting in their deployment. It can be the training data used to train the predictive model, the test data used to validate the model, or even the past production data. As part of the monitor configuration process, certain computations are learned on the baseline data to learn the data patterns. These can vary from dividing data in frequency bins to learn the density functions of your input features (*feature drift*), to training auto-encoders to learn the context represented by the embeddings (*embedding drift*), training proxy/meta models to learn user model behavior (*model quality drift*) and to look at how the metadata like character counts and word counts is changing (*input and output metadata drift*). Any change in the data is reported as a metric on different dimensions.

6.2.2 Explainability

For the predictive AI models, watsonx.governance gives users a sneak peek into the black box by giving localized explanations to understand how the different feature values are impacting the outcomes of the specific transactions. By aggregating these local explanations for a sample of such transactions, a global explanation is presented so that the user can understand the general factors that are influencing the model decisions.

To this end, watsonx.governance utilizes both open-source algorithms (*LIME and SHAP*) and IBM Research® built contrastive explanations. By generating and analyzing data points, in the vicinity or the local neighborhood of a given transaction, Local Interpretable Model-agnostic Explanations or LIME can tell which features of a structured record, words and phrases from a text paragraph, and which areas of the image are responsible for the model outcome. SHAP (SHapley Additive exPlanations) is rooted in game theory as it uses the classic Shapley values, to determine how much of each of the features has contributed to the model prediction. IBM Research built contrastive explanations that look at the neighborhood of a given data point, to determine how much of a delta change is required in the input features to flip the model outcome or to maintain the same outcome. A highly important feature in this case is a feature to which the model is least sensitive as they require a large change in the value for model to flip the outcome.

6.2.3 Model health

To understand the model health and performance of a given model deployment (for both predictive AI and generative AI AI-based models), it is imperative to know the how the said deployment is being used. To aid with that, watsonx.governance helps in calculating and visualizing the total number of *scoring requests* received by the deployment in a given time period. Across these requests, common statistical attributes like minimum, maximum, mean, and median of the *number of records* are also calculated. It is also important to know how much time is taken by the system to process a record and/or a request. Similarly, watsonx.governance can also measure throughput of the system by looking at the number of records and requests processed in a second. For generative AI-based deployments, the input and output token counts processed across the scoring requests can be visualized as well. If there are multiple users registered on the system and using the deployments, watsonx.governance can also present the real-time view to see the total number of users and the aggregated views to see the average number of users.

6.2.4 Generative AI quality

To assess the quality of the content generated by a prompt, watsonx.governance has many Generative AI Quality metrics. Some of these metrics work in the presence of a reference input and hence work off the feedback data. However, there are reference free metrics as well, that do not require any reference input and can be calculated on the production data.

IBM watsonx.governance also allows the use of widely available LLM models to evaluate the performance of the user prompts through the LLM-as-a-judge feature.

6.2.5 RAG quality metrics

By adding external sources of context to the prompt of an LLM, Retrieval Augmented Generation (RAG) systems enhance the quality of the content generated by the model. With this perspective in place, watsonx.governance can monitor both the phases of a RAG-based system.

The retrieval phase can be assessed by looking at the context pulled and seeing how relevant it is to the question asked by the user (context relevance). By looking at retrieval precision, one can tell if the retrieved information is directly addressing the user query. The system also looks at how the different retrieved contexts are ranked. If the most relevant context has the top rank, the metric reciprocal rank will be 1 else it will be much lower. The ranking quality of the retrieved information can also be measured by Normalized Discounted Cumulative Gain (NDCG) as a higher score on this metric, indicates the retrieved contexts are ranked in the correct order.

The content generation phase can be assessed not only by looking at the overall quality of the answers generated by the system, but also by analyzing the content watsonx.governance can tell how much of the context is used in the answers generated. By measuring how well the answer aligns with the context (faithfulness), or by measuring if the answer is relevant to the prompt (answer relevance) or by looking at the number of questions that were unanswered by the model (unsuccessful requests), watsonx.governance gives the quality of the answers. The prominent content analysis metrics measure the percentage of keywords in the answer that are derived from the context (coverage) and the overall sequence of words in the answer are direct extractions from the context (density).

In addition to the above metrics, along with the faithfulness, watsonx.governance also gives out the top source attributions for the generated by answer by highlighting the relevant sentences in the context. This capability tries to open the black box, and is a step in the direction of providing explainability for the foundation model-generated answers.

With a review of the common stages, we can now move onto implementation details for how to enable and complete lifecycle monitoring.

6.3 How to implement Lifecycle Governance

Lifecycle Governance begins at the use case level. In watsonx.governance, an AI use case is created and used to maintain a hierarchical organization of AI assets. This AI use case structure allows users to view their ML/AI assets as they relate to the business problem of the use case.

One use case can hold multiple assets with each asset being represented by a factsheet. These factsheets hold all details of a given asset. Consider an AI factsheet to be a sort of “nutrition label” to the underlying AI asset. This factsheet is what will be used to represent the AI asset as it moves through each stage of its lifecycle.

The following sections will show us how to set up AI use cases and how to create AI factsheets for AI assets.

6.3.1 Getting started: Setting up your AI use cases

After installation and administration of watsonx.governance, we can implement lifecycle governance beginning at the use case level. The purpose of lifecycle governance is monitoring the progress of AI assets individually. These AI assets progress through what is called an AI use case. These AI use cases are central to lifecycle monitoring because it allows us to group AI assets together which are working towards similar goals.

Before setting up lifecycle monitoring for the AI assets, we will set up our AI use case in watsonx.governance.

AI use case and AI factsheet setup

Follow these steps to set up the AI use case in watsonx.govenance:

1. Use the Options menu from the **watsonx home page** to access the **AI use cases** page (under **AI governance** → **AI use cases**). See Figure 6-3 on page 67.

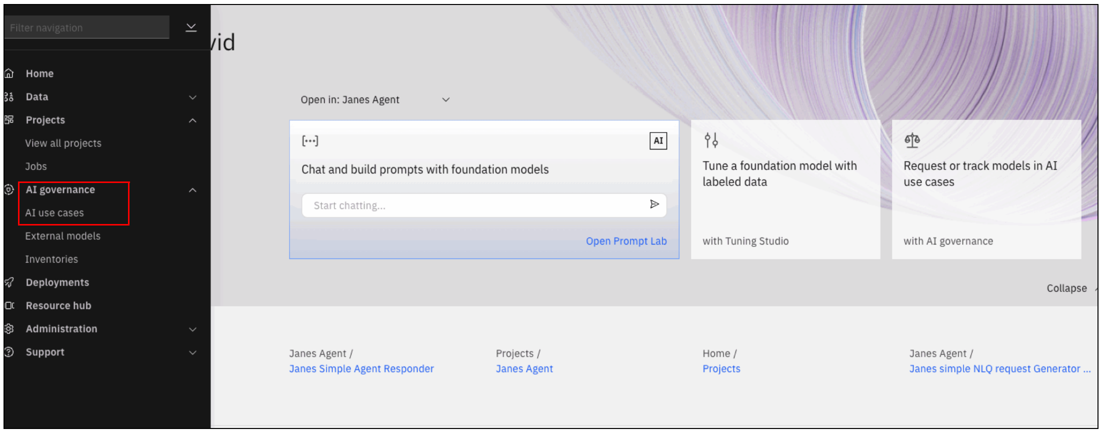


Figure 6-3 AI use cases

2. Create a new inventory using the options menu from the **AI use cases** page.

   Be sure to activate **Governance Console integration** to ensure that your AI Inventory is synced with your Governance console. This synchronization effect persists throughout the work done in this chapter.

3. Once the new inventory is created, create an **AI use case** and fill out all details for that AI use case. Any details posted to this **AI use case** will also be reflected in the Governance Console.

With an AI use case successfully created, we will move on to tracking the lifecycle of an AI asset for that AI use case. This will be achieved by attaching AI factsheets to the AI assets realizing and implementing the AI use case and using that AI factsheet to accomplish lifecycle governance. As mentioned 6.3, “How to implement Lifecycle Governance” on page 66, an AI factsheet is the organizational representation of an AI asset which gives us all asset details. The factsheet also gives us an effective vehicle for traversing between lifecycle stages.

Creating an AI factsheet for a given asset will be covered in detail in subsequent sub-sections. Once the AI factsheet is created, we can move onto lifecycle governance.

As described in 6.1, “What is the AI lifecycle?” on page 62, there are general activities in each lifecycle stage which should be completed before moving onto the next lifecycle stage. Using watsonx.governance, we can achieve full lifecycle monitoring for the AI asset. The following subsections will cover the basic steps for accomplishing lifecycle governance using watsonx.governance.

The following subsections assume that the reader is experienced with LLM experimentation and that they are ready to begin governing and monitoring their AI solutions. Special considerations exist based on implementation details such as what platform the model is hosted on, and what is the use case. These considerations will be found under each appropriate subsection.

6.4 Lifecycle implementation and considerations

This section of the chapter will assume that there is a valid AI use case configured in watsonx.governance. It will also make the assumption that a solution exists (either on watsonx.ai or another third party) which realizes that AI use case. One final major assumption is the existence of ground-truth data.

As mentioned in the explanation of the validation stage, evaluations provide first-hand performance reviews to assets in a qualitative and quantitative way. To achieve this, ground-truth data is critical and must be acquired in order to effectively implement lifecycle governance.

The goal of this section is to explain the connection between the assets (prompt template assets or models) that realize the use case, which lifecycle stage those assets belong in, and how we automate the lifecycle of those assets as they progress from development through to operation.

Before diving into specific considerations, here are the higher-level activities which must be completed to effectively govern the lifecycle of an AI asset. A Prompt Template Asset or a Model must exist that contributes to the implementation of an established use case. A Prompt Template Asset (PTA) can be either a Prompt Template from IBM's watsonx.ai, or a detached prompt template representing a prompt on a 3rd party platform. Similarly, a Model is simply any ML model which solves a traditional data science / ML use case and can be either deployed in IBM watsonx.ai runtime or a 3rd party serving platform. With the asset ready, the steps are as follows:

- Track the solution in an AI use case to place in the development stage.
- Perform an evaluation to move the asset to the validation stage.
- Promote the asset to a production space to move the asset into the operation stage.

The above three steps are repeated in some variation based on platform and use case specifics. As a reminder, everything accomplished in this chapter will be reflected across the

Governance Console for administrative and business users to be automatically updated of progress.

6.4.1 UI-driven implementation of lifecycle governance

> **Note:** This subsection covers a UI-driven implementation of lifecycle governance using watsonx.gov and watsonx.ai. For code-driven approaches, review 6.2.3, “Model health” on page 57 and 6.2.4, “Generative AI quality” on page 57. The watsonx.ai documentation will also be helpful.

After installation and administration, watsonx.ai seamlessly integrates with watsonx.governance. Once an AI use case is established, users can run through a typical watsonx.ai workflow to begin experimenting with prompts in the prompt lab, or in a jupyter notebook environment. For instructions on how to perform this experimentation, refer to the watsonx.ai documentation or the Redbooks publication: *Simplify Your AI Journey: Unleashing the Power of AI with IBM watsonx.ai*, SG24-8574.

Development

After saving your prompt lab experiment, from the project home screen of watsonx.ai, view your assets tab. Find your saved prompt lab template, and using the hamburger menu of that prompt lab template, choose **Track in AI use case**. After following the guided click-through setup, your Prompt Template Asset (PTA) will be automatically placed under the development stage, tracked in the AI use case of your choice.

Validation

With a PTA being tracked against an AI use case in the development stage, users can utilize the power of watsonx.governance to perform fast and effective evaluations of their assets. To perform this step, validation data is required.

Work with use case SMEs and other teammates in order to build up a quality dataset of at least 10 records in CSV format which you can use to validate the performance of your PTA. Based on your use case, the format of your data will vary. Explore the evaluation page for your specific use case in order to identify the correct format for your validation data. For example, a summarization use case will have the following format for the test data:

```
-Original_Text1, Ground_Truth_Summary1-
-Original_Text1, Ground_Truth_Summary2-
```

Once data is acquired, use the hamburger menu to access the options of the given PTA. Click on Evaluate to open the evaluation page. Using the Actions drop-down, click on "Evaluate Now" and follow through the options on-screen to set up the evaluation experiment. To review the available evaluation metrics, check the latest documentation for your version of watsonx.governance.

Note: There are multiple ways to access PTA evaluations. You could access this page through your AI Factsheet, on the second tab. You could also run evaluations directly from the Prompt lab. Feel free to explore the accessibility of this watsonx.governance feature.

After completing the evaluation, go back to your AI use case and scroll down. You should now see this asset as being tracked in the validation stage as a result of running the evaluation experiment. Before moving onto the operation stage, it is most common for assets to sit in the validation stage while iterations happen to improve performance. Evaluations can be constantly run against assets in the validation stage and the records of the latest evaluation can be viewed from the AI factsheet for that asset.

Operation

Once the validation stage is completed and the asset is ready to move into operation stage, head to your watsonx platform home page. Using the hamburger menu from the left, click on **Deployments** to access the watsonx.ai runtime platform (traditionally known as Watson Machine Learning or WML), and create a new deployment space with "Production" type selected. Go back to the watsonx.ai project which holds the PTA, and using the options menu attached to the PTA, choose the **Promote to Space** option. Follow the on-screen guidance to promote the PTA to the newly created production space.

Go to the home page of the deployment space on watsonx.ai runtimes. From the assets tab, the recently promoted PTA should be listed. Use the options menu attached to the PTA from this screen and choose the **Deploy** option. Follow the guided screen to fill out the relevant details, and click **OK**.

After a few moments, your new PTA will be deployed (and consumable) on the production deployment space. You can now check again from your AI use case page to observe that the asset has moved from development, to validation, and finally into the operation stage. Congratulations, you have successfully performed a lifecycle on this AI asset!

Note: Configuring monitors is an additional task beyond the concept of simply moving through a lifecycle stage. This is specifically related to the operation stage of any AI asset; once the asset is in production (operation stage), the monitors for that deployment can be configured to track specific metrics. To achieve this through the UI for a watsonx.ai prompt, use the **Actions** drop down menu on the top right section of the production deployment and choose **Activate Monitors**. Finally, follow the guided instructions to complete monitor configuration.

From this point on, considerations for each of the following sub-sections will avoid repeating information and more advanced topics will be introduced throughout each section. At the highest level, all integrations of watsonx.governance with ML/AI providers work with the same concept of having an AI use case set up at watsonx.governance. The following sections will become more technical with various links to resources in the form of SDKs, notebooks, and code commands to help programmers accelerate implementation.

6.4.2 Considerations for lifecycle governance for traditional ML hosted on watsonx.ai

The same seamless integration applies for traditional ML when considering the watsonx.ai and watsonx.governance platforms. Assuming the initial ML model has been saved as an asset to the watsonx.ai project space, it can be associated to a use case through the UI in the same manner as explained in section 6.4.1. In addition to the UI-based approach, lifecycle governance can be implemented via code.

This section will focus on useful APIs that can be used to achieve full lifecycle governance this as well as code examples. It will make an assumption of intermediate Python programming skills on behalf of the implementer.

> **Note:** This subsection will utilize code-driven approaches. For more information, see Tracking a machine learning model.

API and SDK for watsonx.governance

This section assumes an API or Python-driven SDK approach to programming is being undertaken. Two main API and SDK can be used to work with watsonx.governance. Other APIs also exist which can help, and they will be mentioned as necessary throughout the chapter.

- AI Factsheets API and SDK - The watsonx.governance AI Factsheet API and SDK is used to control the factsheets component of watsonx.governance. It will be used to set up factsheet objects and associate them with given assets.
- OpenScale API and SDK - The watsonx.governance OpenScale API and SDK controls the computational layer of watsonx.governance. The brunt of activity comes from this API. Users can accomplish a variety of activities including:
  - Creating subscriptions to monitor deployments.
  - Performing model evaluations.
  - Associating existing factsheets to AI use cases.

Additionally, Governance Console APIs are also available for interacting with Governance Console within watsonx.governance.

Metrics

The software comes out-of-box with a plethora of traditional metrics to monitor against. For more information, see Quality evaluations. Additionally, users can implement their own custom evaluation metrics.

For custom metrics implementation, users can use this sample notebook from the IBM GitHub.

Note: The notebook linked above also covers how to code the configuration of quality monitors, fairness, drift, and explanations. It is highly recommended to consult sample notebooks for thorough and up-to-date instructions. For users who are looking to create data configurations to specific data science problems, see the IBM directory for different options on creating data configurations.

Inventory and AI use case setup

AI Inventories are used to hold AI use cases. These Inventories are agnostic to asset type - a generative AI use case can be held in the same inventory as a traditional ML/AI use case.

Inventory setup

This notebook from IBM's github showcases how to create an AI Inventory. This sample notebook demonstrates a variety of functionality including:

- Open a new AI inventory
- Add collaborators
- Delete inventory
- Modify existing inventories

Because the AI Inventory concept is native to watsonx.governance, it is always implemented regardless of where the AI assets live which are being tracked.

AI use case setup

This notebook from the IBM GitHubshowcases how to create an AI use case. The sample notebook demonstrates a variety of functionality including:

- Storing a model as an asset into watsonx.ai (this creates the AI factsheet)
- Create a new AI use case
- Associate the AI use case with a watsonx.ai project
- Track/Untrack a model under an AI use case

Development

In “Inventory and AI use case setup” on page 71, the linked notebook holds all relevant code for placing a model into the development stage of its lifecycle. Here it is again in condensed form.

Note: The following pseudo-code requires additional code to function properly. Users should consult the notebook and the API documentation linked in the previous sections for a full walkthrough.

```
# create an AI use case

ai_usecase = 
facts_client.assets.create_ai_usecase(catalog_id=ai_usecase_inventory_id, 
name=ai_usecase_name,description=ai_usecase_desc)

# track the model in the use case

watsonx_ai_model.track(usecase=ai_usecase,approach=decisiontree_approach, 
version_number="major",version_comment="major update to previous version")
```

Validation

Once the model is successfully tracked against an AI use case, it is placed into development stage. If the model is already deployed into a production space, it will instead show in the operation stage. This section will assume we have a model in the development stage of the lifecycle.

There are multiple ways to move a model from the development stage into the validation stage. One way is to use the direct API call:

```
model.set_environment_type(from_container="develop", to_container="validate").
```

For more information, see Managing the Lifecycle Phases of a Model.

Additionally, you can deploy the model to a development deployment space or a validation deployment space and this will trigger the lifecycle move. For more information, see Tracking prompt templates.

This functionality applies to all kinds of implementation, not just traditional ML.

Users can set the environment directly if the asset is in a preceding stage, or users can promote models to a validation space in watsonx.ai runtimes.

Operation

The operation stage is a uniquely important stage of the AI/ML lifecycle. Section 6.4.1, “UI-driven implementation of lifecycle governance” on page 68 covered basic steps to move an asset to the operation, but many more activities can be accomplished for a thorough implementation of a production monitor. As mentioned in section 6.2, “Metrics in watsonx.governance” on page 64 and again in 6.4.2, “Considerations for lifecycle governance for traditional ML hosted on watsonx.ai” on page 70, a variety of metrics are available for configuration through watsonx.governance. When considering a model monitor in the operation stage, these metrics can be configured using the watsonx.governance UI or using notebooks. Review and follow the notebooks and directions provided in section 6.4.2, “Considerations for lifecycle governance for traditional ML hosted on watsonx.ai” on page 70 to establish baseline data configurations and respective monitors.

Additionally, code commands also exist to directly move assets into the production (operation) stage. Here is a snippet on how to move a model into a production deployment space to see the lifecycle stage progress into operation stage:

```
model.set_environment_type(from_container="validate", to_container="production")
```

watsonx.governance comes with an expansive toolkit for handling the lifecycle of traditional ML/AI assets. Considerations for third parties do exist, but the tools that we have covered up to this point will be adapted to cover those additional considerations.

6.4.3 Considerations for prompt templates from another platform

IBM watsonx.governance empowers organizations to evaluate and monitor prompt templates for a variety of externally-hosted LLMs without the need to conduct inference on those models. This flexibility allows AI and Data Science practitioners to work with models hosted on platforms such as Google Vertex AI, Azure OpenAI, or AWS Bedrock, where all inference is performed remotely.

The platform offers a method known as the *detached prompt template* for evaluating and monitoring prompt templates for externally-hosted LLMs without requiring model inference.

This approach involves programmatically creating a detached prompt template asset, which provides a high level of control. Evaluations are then conducted on the generated prompt output, with the results logged into watsonx.governance against the detached prompt template.

Additionally, one can evaluate a detached prompt template within a deployment space by creating a detached deployment. This setup offers several benefits and capabilities:

- Evaluating a prompt template within a project or space enhances the experience of reviewing evaluation results.
- One can utilize access control for projects and spaces to invite collaborators or restrict access as needed.
- The results of the evaluations can be tracked in factsheets related to AI use cases as part of the governance solution.

Note: The following pseudo-code requires additional code to function properly. For more information, see the notebook and the API documentation.

The sample code to create a detached prompt template is shown in Example 6-1.

Example 6-1 Sample code to create a detached prompt template

```
from ibm_aigov_facts_client import DetachedPromptTemplate, PromptTemplate

   detached_information = DetachedPromptTemplate(
   prompt_id=prompt_id,
   model_id=model_id,
       model_provider=model_proivder,
       model_name=model_name,
       model_url=model_url,
       prompt_url=prompt_url,
       prompt_additional_info=prompt_additional_info
   )

   prompt_template = PromptTemplate(
       input=input_text,
       prompt_variables=prompt_variables,
       input_prefix=input_prefix,
       output_prefix=output_prefix,
       model_parameters = model_parameter
   )

   external_prompt_template_details = facts_client.assets.create_detached_prompt(
       name=prompt_name,
       description=prompt_description,
       model_id=model_id,
       task_id=task_id,
       prompt_details=prompt_template,
       detached_information=detached_prompt_template
   )
   project_pta_id = external_prompt_template_details.to_dict()["asset_id"]
```

Once the detached prompt template is created, it follows the same lifecycle as described in 6.4.2, “Considerations for lifecycle governance for traditional ML hosted on watsonx.ai” on page 70.

6.4.4 Considerations for traditional ML from another platform

Just like prompt templates from other platforms, ML assets and deployments can exist outside of watsonx.governance and still be monitored and tracked by watsonx.governance. Lifecycle governance of third-party ML models deployed on other platforms can be implemented through python code. IBM watsonx.governance provides all of the tools and methods necessary to achieve lifecycle governance regardless of where the asset exists. Section 6.4.2, “Considerations for lifecycle governance for traditional ML hosted on watsonx.ai” on page 70 introduces the tools which will be used to achieve these goals.

Many of the same steps apply when implementing lifecycle governance for ML assets deployed on other platforms. Review the setup steps covered in section 6.4.2, “Considerations for lifecycle governance for traditional ML hosted on watsonx.ai” on page 70 to configure the AI Inventory and AI use case and review the introductions to the packages along with the data statistics configurations notebooks in that section. Once those activities are complete, new assets can be tracked against use cases in their appropriate lifecycle stages.

It is worth noting that many ML processes which exist on other platforms are typically considered to be in a production state. Because of this, model tracking for third-party assets is most typically an administrative and organizational task while that asset is being developed and validated in its respective environment. Once the asset arrives into the production or operation stage, monitors can be configured for thorough and effective ML governance. For full instructions on how to configure headless subscriptions for ML models which are not hosted through watsonx.governance, see this notebook.

Batch processing can also be achieved with watsonx.governance. For step-by-step instructions on achieving batch processing, see this notebook, the documentation.

6.4.5 Governing AI embedded in a business application

If an AI/ML asset which is consumed in a business application is directly owned and operated by the governance team (or a department from the same organization as the governance team), any of the above techniques can be used to apply AI governance. Depending on the kind of AI/ML being used in the business application, a user may have to configure a detached prompt template for LLM (as described in section 6.4.3, “Considerations for prompt templates from another platform” on page 72) or they may have to configure a headless subscription for ML (as described in section 6.4.4, “Considerations for traditional ML from another platform” on page 74) If the ML/AI process is being provided by an outside vendor, various considerations should be made when looking to apply governance to that process. This section will shed light on some of the most common and important things to consider when looking to govern an existing business application where the AI/ML process is being provided through a vendor / service from outside the organization.

Consider accessibility

If the governance team is not the direct owner of an AI process, the accessibility of that AI process must be explored and evaluated. Contact the AI vendor and learn about the accessibility to the model. Can a headless subscription or a detached prompt template be configured to create a live monitor of that model or process? If not, what are the alternatives that the vendor can provide? Do those alternatives meet your organization's requirements for AI/ML governance?

Considering requirements

Think about what the requirements are for governing the business application or process. Is live monitoring the only way to achieve those requirements? Can the requirements be achieved through some other means such as a report card of the AI process, or some other status update?

Ultimately, implementing end-to-end lifecycle monitoring of AI embedded in a business application is unique. Because of the fact that the AI already exists in a business application, some assumption may be made about the stage of that AI asset (it would be considered to be in production if it is live and being consumed by users.) This would make the process shorter than tracking a model from initial development; A user may initiate lifecycle governance on a business application's AI model beginning with the production or operation stage.

Governing an AI/ML model from a business application must be accomplished with the same attention to detail regardless of who is the vendor / provider. As the process continues to evolve for AI governance, rules and requirements should be established with regarding which AI/ML providers and vendors are to be sbe utilized by the organization. Before configuring lifecycle monitoring for those AI/ML processes, verifying the vendor's ability to meet any legal requirements is an important starting point. Review Chapter 4, “Onboarding a new foundation model” on page 41 and Chapter 5, “Assessing a new use case” on page 53 for organizational processes and considerations to make prior to setting up lifecycle governance.

7

Use cases

This chapter highlights various implementations of watsonx.governance, focusing on fairness, drift, regulatory compliance, and accountability. These implementations span multiple sectors such as healthcare, banking, and finance, leveraging robust data tracking and model auditing mechanisms.

This chapter has the following sections:

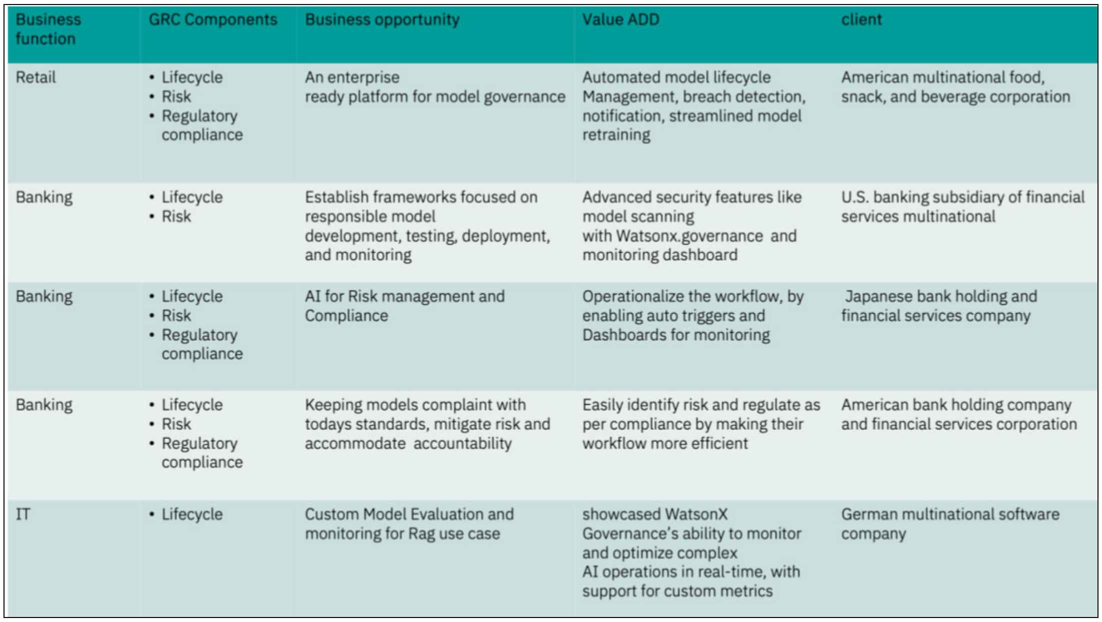
7.1 Overview of use case 1- Banking credit risk management

This section explains the concept of credit risk management, emphasizing fair, transparent approval processes, proactive monitoring with alerts, and automation of model metadata documentation to support credit decisions.

It highlights of the importance of a successful credit risk management system include the following:

- Fair and transparent approval processes.
- Proactive risk detection with alerts to avoid biased decisions.
- Automating the capture and documentation of model metadata with fact sheets to support credit decisions.

Figure 7-1 shows how watsonx.governance can improve businesses credit risk management system. Improvements are shown based on business function to help illustrate the cross-departmental value that proper governance has in any organization.



Business function	GRC Components	Business opportunity	Value ADD	client
Retail	• Lifecycle • Risk • Regulatory compliance	An enterprise ready platform for model governance	Automated model lifecycle Management, breach detection, notification, streamlined model retraining	American multinational food, snack, and beverage corporation
Banking	• Lifecycle • Risk	Establish frameworks focused on responsible model development, testing, deployment, and monitoring	Advanced security features like model scanning with Watsonx.governance and monitoring dashboard	U.S. banking subsidiary of financial services multinational
Banking	• Lifecycle • Risk • Regulatory compliance	AI for Risk management and Compliance	Operationalize the workflow, by enabling auto triggers and Dashboards for monitoring	Japanese bank holding and financial services company
Banking	• Lifecycle • Risk • Regulatory compliance	Keeping models complaint with todays standards, mitigate risk and accommodate accountability	Easily identify risk and regulate as per compliance by making their workflow more efficient	American bank holding company and financial services corporation
IT	• Lifecycle	Custom Model Evaluation and monitoring for Rag use case	showcased WatsonX Governance's ability to monitor and optimize complex AI operations in real-time, with support for custom metrics	German multinational software company

Figure 7-1 Credit risk management

7.1.1 Banking credit risk management use case

Credit risk management is the practice of mitigating losses by assessing borrowers' credit risk - including payment behavior and affordability. This process has been a longstanding challenge for financial institutions requiring continuous adaptations by businesses to better track borrower behavior.

7.1.2 Business context

A French cooperative bank provides banking products and services, focusing on risk management and regulatory compliance. However, regulatory monitoring was done manually, leading to inefficiencies and challenges in identifying emerging regulatory risks resulting from regulatory changes and new requirements impacting customer services. The French cooperative bank caters to individuals and businesses by offering a range of banking products and services.

However, the bank faced challenges due to:

- A lack of centralized tools for regulatory monitoring.
- Manual processes using Excel, resulting in inefficiencies in identifying applicable regulatory risks and regulatory changes.

7.1.3 Client need

The client required a centralized tool to aggregate legal texts from multiple data sources, enabling lawyers to track regulatory changes and link them to specific business units and banking product offerings, such as banking cards and insurance.

The client required a governance tool to:

- Aggregate legal texts from multiple data sources.
- Maintain a history of regulatory changes and link this data to impacted products and business units.

7.1.4 Client challenges

The client wants to address the challenges in the following areas:

- The client faced significant hurdles in streamlining their governance, risk, and compliance (GRC) processes:
  - **Fragmented data systems**: GRC data was scattered across multiple disparate systems, creating silos and making it difficult to obtain a unified view of risks.
  - **Manual dependency**: Reliance on manual processes for managing GRC was inefficient and error-prone, especially in the context of the monitoring and management of the rapid number of regulatory changes.
- The handling of large volumes of data, which overwhelmed manual workflows and reduced accuracy.

7.1.5 Business benefits

Here are several business benefits and improved risk management benefits:

- By addressing these challenges, the solution delivered substantial business benefits:
  - **Unified Governance Framework**: Consolidating GRC processes into a single, integrated platform provided a comprehensive view of risks, enabling better decision-making and streamlined operations.
  - **AI-driven automation**: Leveraging AI technologies accelerated GRC processes, reducing manual effort and improving efficiency. This automation not only minimized human errors but also ensured faster compliance with evolving regulatory requirements.
- **Improved risk management**: Enhanced visibility and control over GRC processes empowered the organization to proactively identify and address potential risks, fostering resilience and compliance across operations.

7.1.6 Pilot solution

This pilot solution uses watsonx.governance Governance Console (OpenPages).

In a prior Minimum Viable Product (MVP), IBM OpenPages demonstrated value by importing consumer code and GDPR laws via standard REST API such as Légifrance REST API in France. Initially, Excel was used for linking with the eventual goal being direct integration within IBM OpenPages.

List of key steps:

- **Develop a Connector**: Connect the Légifrance REST API and IBM OpenPages REST API to fetch and properly format regulatory data.
- **Automate Hyperlinks**: Enable referencing between articles for better navigation and usability.

By streamlining legal monitoring, the solution enhances risk identification and regulatory compliance, reduces prior manual steps, and improves the quality of formatting legal data.

7.2 Overview of use case 2 - Automated governance for universal bank's AI chatbot

A prominent British universal bank and financial services group, provides a broad range of offerings such as savings accounts, loans, insurance, and investment options. A key focus of the bank is continuous improvements on efficiencies around managing risk and ensuring regulatory compliance.

Their AI-powered *chatbot* is designed to provide AI-driven solutions and must adhere to strict standards of ethics, explainability, and expected performance. To support this, the bank seeks to govern its AI systems with robust, automated, and integrated platforms as data and AI technologies evolve.

7.2.1 Business context

They require a governance framework that ensures:

- Ethical and explainable AI behavior.
- Reliable results in alignment with business objectives.
- Adequate control, testing, and audit mechanisms to manage evolving data and AI models.

The solution will leverage IBM watsonx.governance, an automated and integrated AI governance platform, to manage the lifecycle and compliance of the AI application.

7.2.2 Client need

The client aims to automate the governance lifecycle of a *chatbot* across various metrics, lifecycle stages, and compliance requirements.

7.2.3 Client challenges

The client focused on plans to address two main challenges:

- The current AI implementation did not have an adequate monitoring system to maintain a stable solution.
- Need to find a way to improve mechanisms to address the diverse aspects such as bias detection, ethical compliance, handling of highly autonomous processes (HAP), and protection of personally identifiable information (PII).

7.2.4 Business benefits

The following benefits support addressing the client's challenges:

- **Consolidated governance**: Provide a unified platform for an aggregated view of risks.
- **Automation of GRC processes:** Leverage AI to streamline governance, risk, and compliance (GRC) processes, significantly accelerating time to value.

7.2.5 Pilot solution

The pilot addressed these requirements with the following key features:

- **Quantified Quality Metrics**: Faithfulness, answer relevance, handling of unsuccessful requests, keyword inclusion, answer coverage, and spelling robustness were all measured to ensure high performance.
- **Governance Dashboard**: Developed a comprehensive dashboard to simplify the oversight of their AI governance lifecycle.
- **External Model Governance**: Implemented governance for external models using detached prompt templates.

This structured solution ensures bank's *chatbot* AI product operates within ethical, regulatory, and performance parameters while automating the lifecycle governance for enhanced efficiency.

7.3 Overview of use case 3 - Belgian biopharmaceutical company

A Belgian biopharmaceutical company, is leveraging watsonx solutions to address challenges in website development, including technical documentation and module reusability, for their Drupal-based global web ecosystem.

7.3.1 Business context

IBM is developing and managing over 150 global websites for the pharma company using Drupal based technology. Client faces challenges in maintaining high-quality technical documentation, which limits clarity on the functionality of existing Drupal modules and creates barriers for local developers seeking to reuse available modules.

7.3.2 Client need

Client seeks a solution that improves the website development process by:

- Enhancing technical documentation.
- Simplifying module discovery and reuse.

- Supporting local developers with efficient tools and governance mechanisms.

7.3.3 Client challenges

The following challenges need to be addressed:

- Lack of clear, comprehensive technical documentation for existing Drupal modules.
- Inefficiencies in reusing modules across local teams due to limited information.
- Fragmented governance, making it difficult to maintain consistency and compliance across websites.

7.3.4 Business benefits

Several key benefits are:

- Streamlined documentation creation and maintenance with AI-driven tools.
- Improved developer productivity through better access to reusable modules.
- A unified governance framework for consistent and efficient website development.

7.3.5 Pilot solution

The following steps highlight the implementation of the pilot solution:

1. **Objective:** Demonstrate how watsonx solutions can enhance website development by addressing documentation challenges, improving module reusability, and supporting governance.
2. **Steps implemented**:

   The following steps show how the client addressed the challenges:

   a. **Understanding requirements**:

      Conducted workshops with pharma company stakeholders to identify pain points and gather insights into their Drupal-based ecosystem.

   b. **Developing solutions**:

      The client used the following products to develop the solution:

      - **IBM watsonx.ai:** Used for generating and maintaining AI-driven technical documentation.
      - **IBM watsonx.governance**: Implemented to centralize and streamline governance for Drupal modules.

   c. **Knowledge base creation**:

      Built a searchable repository with detailed descriptions and usage guidelines for Drupal modules.

   d. **Prototype and demonstration**:

      The client developed the prototype to demonstrate improved productivity:

      - Created a prototype showing how watsonx tools improve documentation and module discovery.
      - Demonstrated how developers can leverage these tools to enhance productivity.

   e. **Feedback loop**:

      The client implemented a feedback loop to improve the solution:

   - Collected feedback during the pilot from client's technical teams.
   - Iteratively refined the solution to align with client's specific workflows and requirements.

3. **Outcome**:

   The outcome of the following demonstrated the solution improvement:

   – Enhanced module documentation quality and accessibility.

   – Increased developer efficiency by enabling effective module reuse.

   – Established a robust governance framework for Drupal website development.

4. **Future scope**:

   The following items were identified for future additional improvement to the solution;

   – Scale the pilot solution across global websites.

   – Expand the use of watsonx tools to other areas of the pharmaceutical company's digital ecosystem for broader impact.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Simplify Your AI Journey: Hybrid, Open Data Lakehouse with IBM watsonx.data,* SG24-8570
- *Simplify Your AI Journey: Unleashing the Power of AI with IBM watsonx.ai*, SG24-8574

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM watsonx documentation
- IBM watsonx.governance
- IBM watsonx product portfolio
- IBM AI risk atlas

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

SG24-8573-00

ISBN 0738461970

Printed in U.S.A.