

Derisking AI by design: How to build risk management into AI development

The compliance and reputational risks of artificial intelligence pose a challenge to traditional risk-management functions. Derisking by design can help.

by Juan Aristi Baquero, Roger Burkhardt, Arvind Govindarajan, and Thomas Wallace



Artificial intelligence (AI) is poised to redefine how businesses work. Already it is unleashing the power of data across a range of crucial functions, such as customer service, marketing, training, pricing, security, and operations. To remain competitive, firms in nearly every industry will need to adopt AI and the agile development approaches that enable building it efficiently to keep pace with existing peers and digitally native market entrants. But they must do so while managing the new and varied risks posed by AI and its rapid development.

The reports of AI models gone awry due to the COVID-19 crisis have only served as a reminder that using AI can create significant risks. The reliance of these models on historical data, which the pandemic rendered near useless in

some cases by driving sweeping changes in human behaviors, make them far from perfect.

In a previous article, we described the challenges posed by new uses of data and innovative applications of AI. Since then, we've seen rapid change in formal regulation and societal expectations around the use of AI and the personal data that are AI's essential raw material. This is creating compliance pressures and reputational risk for companies in industries that have not typically experienced such challenges. Even within regulated industries, the pace of change is unprecedented.

In this complex and fast-moving environment, traditional approaches to risk management may not be the answer (see sidebar "Why traditional model risk management is insufficient"). Risk management

Why traditional model risk management is insufficient

Model risk management (MRM) in regulated industries such as banking is currently performed by dedicated and independent teams reporting to the chief risk officer. While these firms have developed a robust MRM approach to improve the governance and control of their critical models determining capital requirements and lending decisions, this approach is usually not ideal for firms with different requirements or in less heavily regulated industries, for the following reasons:

- MRM is typically based on a point-in-time model assessment (for example, once every one to five years), which assumes that the models are largely static between reviews. AI models learn from data, and their logic changes when they are retrained to learn from new data. For example, a fraud model is retrained weekly in order to adapt to new scams.
- Traditional MRM workflows are often sequential and require six to 12 weeks of review time after the model development is complete, which delays deployment. These workflows are not easily adapted to the agile and iterative development cycles frequently used in AI model development.
- MRM is often focused more on traditional risk types (primarily financial risks, such as capital adequacy and credit risk) and may not fully cover the new and more diverse risks arising from widespread use of AI such as reputational risk, consumer and conduct risk, and employee risk.
- Some applications and use cases, such as chatbots, natural-language processing, and HR analytics, can qualify as "models" under regulatory definitions used in banking. But these applications are very different from the traditional model types (for example, capital models, stress-testing models, and credit-risk models), and traditional MRM approaches are not easily applied.
- AI and machine-learning algorithms are often embedded in larger AI application systems, such as software-as-a-service (SaaS) offerings from vendors, in ways that are significantly more complex and more opaque than traditional models. This greatly complicates coordination between those who review the model and those who assess the application and platform (IT risk) or the vendor (third-party risk).

cannot be an afterthought or addressed only by model-validation functions such as those that currently exist in financial services. Companies need to build risk management directly into their AI initiatives, so that oversight is constant and concurrent with internal development and external provisioning of AI across the enterprise. We call this approach “derisking AI by design.”

Why managing AI risks presents new challenges

While all companies deal with many kinds of risks, managing risks associated with AI can be particularly challenging, due to a confluence of three factors.

AI poses unfamiliar risks and creates new responsibilities

Over the past two years, AI has increasingly affected a wide range of risk types, including model, compliance, operational, legal, reputational, and regulatory risks. Many of these risks are new and unfamiliar in industries without a history of widespread analytics use and established model management. And even in industries that have a history of managing these risks, AI makes the risks manifest in new and challenging ways. For example, banks have long worried about bias among individual employees when providing consumer advice. But when employees are delivering advice based on AI recommendations, the risk is not that one piece of individual advice is biased but that, if the AI recommendations are biased, the institution is actually systematizing bias into the decision-making process. How the organization controls bias is very different in these two cases.

These additional risks also stand to tax risk-management teams that are already being stretched thin. For example, as companies grow more concerned about reputational risk, leaders are asking risk-management teams to govern a broader range of models and tools, supporting anything from marketing and internal business decisions to customer service. In industries

with less defined risk governance, leaders will have to grapple with figuring out who should be responsible for identifying and managing AI risks.

AI is difficult to track across the enterprise

As AI has become more critical to driving performance and as user-friendly machine-learning software has become increasingly viable, AI use is becoming widespread and, in many institutions, decentralized across the enterprise, making it difficult for risk managers to track. Also, AI solutions are increasingly embedded in vendor-provided software, hardware, and software-enabled services deployed by individual business units, potentially introducing new, unchecked risks. A global product-sales organization, for example, might choose to take advantage of a new AI feature offered in a monthly update to their vendor-provided customer-relationship-management (CRM) package without realizing that it raises new and diverse data-privacy and compliance risks in several of their geographies.

Compounding the challenge is the fact that AI risks cut across traditional control areas—model, legal, data privacy, compliance, and reputational—that are often siloed and not well coordinated.

AI risk management involves many design choices for firms without an established risk-management function

Building capabilities in AI risk management from the ground up has its advantages but also poses challenges. Without a legacy structure to build upon, companies must make numerous design choices without a lot of internal expertise, while trying to build the capability rapidly. What level of MRM investment is appropriate, given the AI risk assessments across the portfolio of AI applications? Should reputational risk management for a global organization be governed at headquarters or on a national basis? How should we combine AI risk management with the management of other risks, such as data privacy, cybersecurity, and data ethics? These are just a few of the many choices that organizations must make.

Baking risk management into AI development

To tackle these challenges without constraining AI innovation and disrupting the agile ways of working that enable it, we believe companies need to adopt a new approach to risk management: derisking AI by design.

Risk management by design allows developers and their business stakeholders to build AI models that are consistent with the company's values and risk appetite. Tools such as model interpretability, bias detection, and performance monitoring are built in so that oversight is constant and concurrent with AI development activities and consistent across the enterprise. In this approach, standards, testing, and controls are embedded into various stages of the analytics model's life cycle, from development to deployment and use (Exhibit 1).

Typically, controls to manage analytics risk are applied after development is complete. For example, in financial services, model review and validation often begin when the model is ready for implementation. In a best-case scenario, the control function finds no problems, and the deployment is delayed only as long as the time to perform

those checks. But in a worst-case scenario, the checks turn up problems that require another full development cycle to resolve. This obviously hurts efficiency and puts the company at a disadvantage relative to nimbler firms (see sidebar "Learning the value of derisking by design the hard way").

Similar issues can occur when organizations source AI solutions from vendors. It is critical for control teams to engage with business teams and vendors early in the solution-ideation process, so they understand the potential risks and the controls to mitigate them. Once the solution is in production, it is also important for organizations to understand when updates to the solution are being pushed through the platform and to have automated processes in place for identifying and monitoring changes to the models.

It's possible to reduce costly delays by embedding risk identification and assessment, together with associated control requirements, directly into the development and procurement cycles. This approach also speeds up pre-implementation checks, since the majority of risks have already been accounted for and mitigated. In practice, creating a detailed control framework that sufficiently

Learning the value of derisking by design the hard way

A large food manufacturer developed an analytics solution to forecast demand for each of its products across geographies in order to optimize manufacturing, logistics, and the overall supply chain. The new model showed higher accuracy compared with the company's existing expert-based approach.

But before the model was deployed, the manufacturer initiated an independent

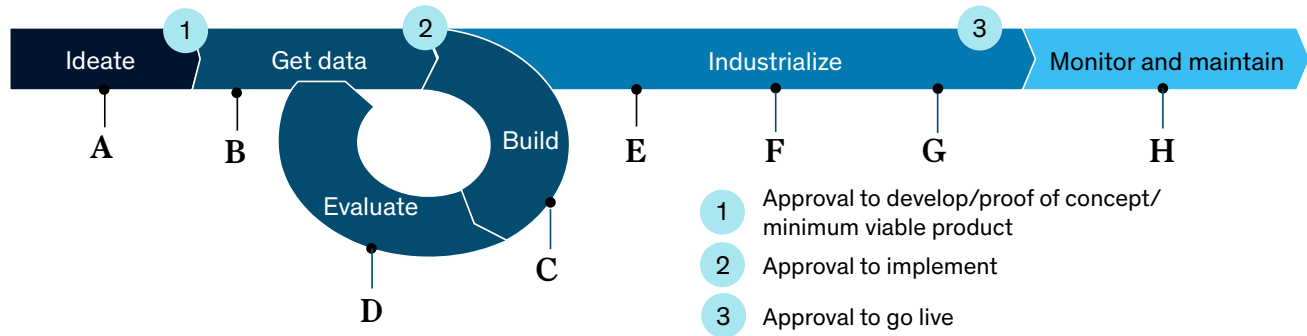
third-party review of the model, which uncovered several problems with the model, including a critical data leakage. The model had accidentally included a feature that captured the actual demand. Once the feature was removed, the model accuracy dropped below the existing expert-based approach.

This revelation led to a complete redesign of the model architecture and the

realization that the company needed to undertake a broader initiative to embed risk management into model development to prevent this and other issues from recurring. The manufacturer began the effort by creating new roles within the group to perform model review, defining roles and responsibilities for model checks throughout the modeling pipeline, and implementing standards for development and documentation of analytics.

Exhibit 1

Risk management by design embeds controls across the algorithmic model's life cycle.



A Designing the solution

Controls examples: scoping review, evaluation metrics, assessment of environment including available data

B Obtaining reliable data required to build and train model

Controls examples: data-pipeline testing, data-sourcing analysis, statistical-data checks, process and data-usage fairness, automated documentation generation

C Building a model that achieves good performance in solving the problem specified during ideation

Controls examples: model-robustness review, business-context metrics testing, data-leakage controls, label-quality assessment, data availability in production

D Evaluating performance of model and engaging business regularly to ensure business fit

Controls examples: standardized performance testing, feature-set review, rule-based threshold setting, model-output review by subject-matter expert, business requirements, business restrictions, risk assessment, automated document generation, predictive-outcome fairness

E Moving model to production environment

Controls examples: nonfunctional-requirements checklist, data-source revalidation, full data-pipeline test, operational-performance thresholds, external-interface warnings

F Deploying model where it starts being used by the business

Controls examples: colleague responsibility assignment and training, escalation mechanisms, workflow management, audit-trail generation

G Inventory management of all models

Controls examples: search tool, automated inventory statistical assessment and risk overview by department

H Live monitoring in production

Controls examples: degradation flagging, retraining scheduler, periodic testing such as Bayesian hypothesis testing, automated logging, and audit-trail generation

Review and approval for continued use

Controls example: verification that algorithm continues to work as intended and its use continues to be appropriate in current environment

covers all these different risks is a granular exercise. For example, enhancing our own internal model-validation framework to accommodate AI-related risks results in a matrix of 35 individual control elements covering eight separate dimensions of model governance.

Embedding appropriate controls directly into the development and provisioning routines of business and data-science teams is especially helpful in industries without well-established analytics development teams and risk managers who conduct independent review of analytics or manage

associated risk. They can move toward a safe and agile approach to analytics much faster than if they had to create a stand-alone control function for review and validation for models and analytics solutions (see sidebar “An energy company takes steps toward derisking by design”).

As an example, one of the most relevant risks of AI and machine learning is bias in data and analytics methodologies that might lead to unfair decisions for consumers or employees. To mitigate this category of risk, leading firms are embedding several types of controls into their analytics-development processes (Exhibit 2):

- **Ideation.** They first work to understand the business use case and its regulatory and reputational context. An AI-driven decision engine for consumer credit, for example, poses a much higher bias risk than an AI-driven chatbot that provides information to the same customers. An early understanding of the risks of the use case will help define the appropriate requirements around the data and methodologies. All the stakeholders ask, “What

could go wrong?” and use their answers to create appropriate controls at the design phase.

- **Data sourcing.** An early risk assessment helps define which data sets are “off-limits” (for example, because of personal-privacy considerations) and which bias tests are required. In many instances, the data sets that capture past behaviors from employees and customers will incorporate biases. These biases can become systemic if they are incorporated into the algorithm of an automated process.
- **Model development.** The transparency and interpretability of analytical methods strongly influence bias risk. Leading firms decide which methodologies are appropriate for each use case (for example, some black-box methods will not be allowed in high-risk use cases) and what post hoc explainability techniques can increase the transparency of model decisions.
- **Monitoring and maintenance.** Leading firms define the performance-monitoring requirements, including types of tests and

An energy company takes steps toward derisking by design

Companies in industries that have been running analytical models for decades under the scrutiny of regulators, such as financial services, often have a foundation for moving to a derisk-by-design model. Organizations in industries that have adopted analytics more recently and are less regulated (at least in the area of model outputs) will need to build their capabilities nearly from scratch.

One large North American energy company initiated a multiyear analytics transformation in order to improve the efficiency of current assets—for example,

to produce higher-quality coal. The company set up an analytics center of excellence (CoE), which discovered that thousands of analytics use cases had been developed and deployed across the organization without any clear oversight, creating risks for human health and safety, financial performance, and company reputation.

In response, the CoE appointed a model manager to oversee the model-governance rollout across the organization. The manager’s team identified six key priorities: implementing a process to identify






















models as they are developed; creating a centralized inventory for all analytics use cases and related information (such as developer and owners); establishing a tiering system to identify the most material models; creating standards for model development and documentation; defining and implementing requirements for model review and monitoring for all models; and defining model-governance processes, roles, and responsibilities for all stakeholders across the modeling pipeline. These changes helped the organization take a giant step toward embedding risk management into the end-to-end process of model development.

Exhibit 2

Bias is one important risk that can be mitigated by embedding controls into the model-development process.

Illustrative example

 Guidance and checklists  Analytical methods and data tools

Ideation	Data sourcing	Model development	Industrialization, monitoring, and maintenance
<p>Determine the level of bias risk, given model use and context.</p>	<p>Detect and mitigate bias risk in data.</p>	<p>Find and reduce bias through modeling.</p>	<p>Continuously monitor and manage bias risk in production.</p>
<p> Bias and explainability risk assessment</p>	<p> Bias-detection techniques</p>	<p> Explainable AI techniques to explain root cause</p>	<p> Context monitoring</p> <ul style="list-style-type: none"> - Regulatory changes - Legal changes - Company-policy changes - Usage appropriateness
<p>How can we set up a team to reduce or mitigate risk of bias?</p>	<p> Fair-representation techniques</p>	<p> Counterfactual analysis</p>	
<p> Guidance on convening a diverse team</p>	<p> Evaluation of risk from the choice of data sets and collection methods</p>	<p> Review of underlying hypotheses</p>	<p> Model monitoring</p> <ul style="list-style-type: none"> - Data drift - Model metrics - Bias metrics in outcomes
<p>What legal and reputational constraints should we take into account?</p>	<p> Mitigation of risk in feature selection and engineering</p>	<p> Fairness-aware algorithms</p>	
<p> Scoping and regulatory guidance</p>	<p> Documentation with data sheets for data sets</p>	<p> Remediation with post-processing techniques on output</p>	<p> Model maintenance</p> <ul style="list-style-type: none"> - Database of metrics and trend tracking - Updates of documentation
<p>How will we measure bias for this use case in this usage context?</p>		<p> Documentation with model cards</p>	
<p> Creation of bias risk metrics</p>			
<p>What is the level of our analytics capabilities?</p>	<p> Execute development checks and controls to manage the risk of bias</p>		
<p> Capability context assessment</p>	<p> Monitor model for bias metrics</p>		

frequency. These requirements will depend on the risk of the use case, the frequency with which the model is used, and the frequency with which the model is updated or recalibrated. As more dynamic models become available (for

example, reinforced learning, self-learning), leading firms use technology platforms that can specify and execute monitoring tests automatically.

Putting risk managers in a position to succeed—and providing a supporting cast

To deploy AI at scale, companies need to tap an array of external and unstructured data sources, connect to a range of new third-party applications, decentralize the development analytics (although common tooling, standards, and other centralized capabilities help speed the development process), and work in agile teams that rapidly develop and update analytics in production.

These requirements make large-scale and rapid deployment incredibly difficult for traditional risk managers to support. To adjust, they will need to integrate their review and approvals into agile or sprint-based development approaches, relying more on developer testing and input from analytics teams, so they can focus on review

rather than taking responsibility for the majority of testing and quality control. Additionally, they will need to reduce one-off “static” exercises and build in the capability to monitor AI on a dynamic, ongoing basis and support iterative development processes.

But monitoring AI risk cannot fall solely on risk managers. Different teams affected by analytics risk need to coordinate oversight to ensure end-to-end coverage without overlap, support agile ways of working, and reduce the time from analytics concept to value (Exhibit 3).

AI risk management requires that each team expand its skills and capabilities, so that skill sets in different functions overlap more than they do in historical siloed approaches. Someone with a core skill—in this case, risk management, compliance, vendor risk—needs enough analytics know-how

Exhibit 3

The responsibilities for enabling safe and ethical innovation with artificial intelligence span multiple parts of the organization.

Business		
Front line Confirm soundness of predictive drivers, modeling approach, and results based on business experience	Operations Validate insights against business experience; ensure appropriate use-case calibration (eg, clarity on modeling objectives)	Business-unit control Ensure tests required by second-line-of-defense functions are performed, including ongoing monitoring and testing of models in use
Analytics		
Data scientists, developers Develop best-in-class models in line with second-line-of-defense standards; provide transparency into model behavior (ie, explainability)	Data Data engineers/strategists Maintain data quality; ensure applicability of new features (ie, feature engineering) to modeling objectives	Technology IT (software and hardware) Mitigate implementation risks by ensuring adequacy of production environment (eg, scalability, preventing data leakage)
Risk and control functions		
Model risk management Develop standards providing guard-rails on AI/ML model development; assess AI/ML model risk	Compliance and legal Provide guidance on compliance risks (eg, prevent bias arising from use of certain restricted customer characteristics)	Cloud risk, vendor risk, etc Provide guidance on mitigating key nonfinancial risks (eg, reputational damage, third-party) linked to AI/ML models

to engage with the data scientists. Similarly, data scientists need to understand the risks in analytics, so they are aware of these risks as they do their work.

In practice, analytics teams need to manage model risk and understand the impact of these models on business results, even as the teams adapt to an influx of talent from less traditional modeling backgrounds, who may not have a grounding in existing model-management techniques. Meanwhile, risk managers need to build expertise—through either training or hiring—in data concepts, methodologies, and AI and machine-learning risks, to ensure they can coordinate and interact with analytics teams (Exhibit 4).

This integration and coordination between analytics teams and risk managers across the model life cycle requires a shared technology platform that includes the following elements:

- an agreed-upon documentation standard that satisfies the needs of all stakeholders (including developers, risk, compliance, and validation)
- a single workflow tool to coordinate and document the entire life cycle from initial concept through iterative development stages, releases into production, and ultimately model retirement
- access to the same data, development environment, and technology stack to streamline testing and review

Exhibit 4

Both analytics and risk professionals will need to complement their traditional skill sets with sufficient knowledge of the others’ function.

	Data and analytics professionals	Risk and control officers
Core competencies	<ul style="list-style-type: none"> • Math, statistics, machine learning, deep learning • Building algorithmic models • Collecting, cleansing, structuring data • Creating data visualizations and dashboards • Explaining model drivers 	<ul style="list-style-type: none"> • Knowledge of applicable regulations • Identification and analysis of risks • Credible and independent review of business activities
New complementary skills	<ul style="list-style-type: none"> • Awareness of analytics risks, including bias, fairness, and instability • Understanding of where risks can arise in analytics-development life cycle • Ability to use risk-management tools as part of analytics-development process (eg, explainability and bias testing, model-performance-monitoring dashboards) • Understanding of risk-control team’s role and responsibilities and ability to engage with them effectively 	<ul style="list-style-type: none"> • General understanding of analytics techniques and their implications, including performance vs interpretability trade-offs • Awareness of best practices in testing for bias, fairness, and stability and ability to understand results from risk-management tools such as explainability reports • Understanding of data/feature-selection practices and their effect on risks (eg, bias) • Understanding of analytics teams’ roles and responsibilities and ability to engage with data and analytics professionals

- tools to support automated and frequent (even real-time) AI model monitoring, including, most critically, when in production
- a consistent and comprehensive set of explainability tools to interpret the behavior of all AI technologies, especially for technologies that are inherently opaque

Getting started

The practical challenges of altering an organization’s ingrained policies and procedures are often formidable. But whether or not an established risk function already exists, leaders can take these basic steps to begin putting into practice derisking AI by design:

- *Articulate the company’s ethical principles and vision.* Senior executives should create a top-down view of how the company will use data, analytics, and AI. This should include a clear statement of the value these tools bring to the organization, recognition of the associated risks, and clear guidelines and boundaries that can form the basis for more detailed risk-management requirements further down

in the organization (see sidebar “Building risk management into AI design requires a coordinated approach”).

- *Create the conceptual design.* Build on the overarching principles to establish the basic framework for AI risk management. Ensure this covers the full model-development life cycle outlined earlier: ideation, data sourcing, model building and evaluation, industrialization, and monitoring. Controls should be in place at each stage of the life cycle, so engage early with analytics teams to ensure that the design can be integrated into their existing development approach.
- *Establish governance and key roles.* Identify key people in analytics teams and related risk-management roles, clarify their roles within the risk-management framework, and define their mandate and responsibilities in relation to AI controls. Provide risk managers with training and guidance that ensure they develop knowledge beyond their previous experience with traditional analytics, so they are equipped to ask new questions about what could go wrong with today’s advanced AI models.

Building risk management into AI design requires a coordinated approach

While AI applications can be developed in a decentralized fashion across an organization, managing AI risk should be coordinated more centrally in order to be effective. A major North American bank learned this lesson when it set out to create a new set of AI risk-management capabilities to complement its existing risk frameworks. Initially, multiple groups began their own AI risk-management efforts. This fragmentation created a host

of challenges around key risk processes, including tracking and assessing the risks of AI embedded in vendor technologies, triaging and risk oversight of AI tools, building controls into AI model development involving multiple analytics groups, and operationalizing ethical principles on data and AI approved by the board. As a result, the bank struggled to demonstrate that all AI risks were managed through the development life cycle.

The bank alleviated these issues by establishing one multidisciplinary team to define a clear target state of AI risk management, build alignment across stakeholders, clarify AI governance requirements, and specify the engagement model and technical requirements to achieve the target state.

- *Adopt an agile engagement model.* Bring together analytics teams and risk managers to understand their mutual responsibilities and working practices, allowing them to solve conflicts and determine the most efficient way of interacting fluidly during the course of the development life cycle. Integrate review and approvals into agile or sprint-based development approaches, and push risk managers to rely on input from analytics teams, so they can focus on reviews rather than taking responsibility for the majority of testing and quality control.

- *Access transparency tools.* Adopt essential tools for gaining explainability and interpretability. Train teams to use these tools to identify the drivers of model results and to understand the outputs they need in order to make use of the results. Analytics teams, risk managers, and partners outside the company should have access to these same tools in order to work together effectively.

- *Develop the right capabilities.* Build an understanding of AI risks throughout the

organization. Awareness campaigns and basic training can build institutional knowledge of new model types. Teams with regular review responsibilities (risk, legal, and compliance) will need to become adept “translators,” capable of understanding and interpreting analytics use cases and approaches. Critical teams will need to build and hire in-depth technical capabilities to ensure risks are fully understood and appropriately managed.

AI is changing the rules of engagement across industries. The possibilities and promise are exciting, but executive teams are only beginning to grasp the scope of the new risks involved. Existing approaches to model risk-management functions may not be ready to support deployment of these new techniques at the scale and pace expected by business leaders. Derisking AI by design will give companies the oversight they need to run AI ethically, legally, and profitably.

Juan Aristi Baquero and **Roger Burkhardt** are partners in McKinsey’s New York office, **Arvind Govindarajan** is a partner in the Boston office, and **Thomas Wallace** is a partner in the London office.

The authors wish to thank Rahul Agarwal for his contributions to this article.

Copyright © 2020 McKinsey & Company. All rights reserved.