McKinsey
& Company

McKinsey Analytics

McKinsey Quarterly

Confronting the risks of artificial intelligence

April 2019 | Article

By Benjamin Cheatham, Kia Javanmardian, and Hamid Samandari

With great power comes great responsibility. Organizations can mitigate the risks of applying artificial intelligence and advanced analytics by embracing three principles.

A**rtificial intelligence (AI)** is proving to be a double-edged sword. While this can be said of most new technologies, both sides of the AI blade are far sharper, and neither is well understood.

Consider first the positive. These technologies are starting to improve our lives in myriad ways, from simplifying our shopping to enhancing our healthcare experiences. Their value to businesses also has become undeniable: nearly 80 percent of executives at companies that are deploying AI recently told us that they're already seeing moderate value from it. Although the widespread use of AI in business is still in its infancy and questions remain open about the pace of progress, as well as the possibility of achieving the holy grail of "general intelligence," the potential is enormous. McKinsey Global Institute research suggests that by 2030, AI could deliver additional global economic output of $13 trillion per year.

Yet even as AI generates consumer benefits and business value, it is also giving rise to a host of unwanted, and sometimes serious, consequences. And while we're focusing on AI in this article, these knock-on effects (and the ways to prevent or mitigate them) apply equally to all advanced analytics. The most visible ones, which include privacy violations, discrimination, accidents, and manipulation of political systems, are more than enough to prompt caution. More concerning still are the consequences not yet known or experienced. Disastrous repercussions—including the loss of human life, if an AI medical algorithm goes wrong, or the compromise of national security, if an adversary feeds disinformation to a military AI system—are possible, and so are significant challenges for organizations, from reputational damage and revenue losses to regulatory backlash, criminal investigation, and diminished public trust.

Because AI is a relatively new force in business, few leaders have had the opportunity to hone their intuition about the full scope of societal, organizational, and individual risks, or to develop a working knowledge of their associated drivers, which range from the data fed into AI systems to the operation of algorithmic models and the interactions between humans and machines. As a result, executives often overlook potential perils ("We're not using AI in anything that could 'blow up,' like self-driving cars") or overestimate an organization's risk-mitigation capabilities ("We've been doing analytics for a long time, so we already have the right controls in place, and our practices are in line with those of our industry peers"). It's also common for leaders to lump in AI risks with others owned by specialists in the IT and analytics organizations ("I trust my technical team; they're doing everything possible to protect our customers and our company").

Leaders hoping to avoid, or at least mitigate, unintended consequences need both to build their pattern-recognition skills with respect to AI risks and to engage the entire organization so that it is ready to embrace the power and the responsibility associated with AI. The level of effort required to identify and control for all key risks dramatically exceeds prevailing norms in most organizations. Making real progress demands a multidisciplinary approach involving leaders in the C-suite and across the company; experts in areas ranging from legal and risk to IT, security, and analytics; and managers who can ensure vigilance at the front lines.

This article seeks to help by first illustrating a range of easy-to-overlook pitfalls. It then presents frameworks that will assist leaders in identifying their greatest risks and implementing the breadth and depth of nuanced controls required to sidestep them. Finally, it provides an early glimpse of some real-world efforts that are currently under way to tackle AI risks through the application of these approaches.

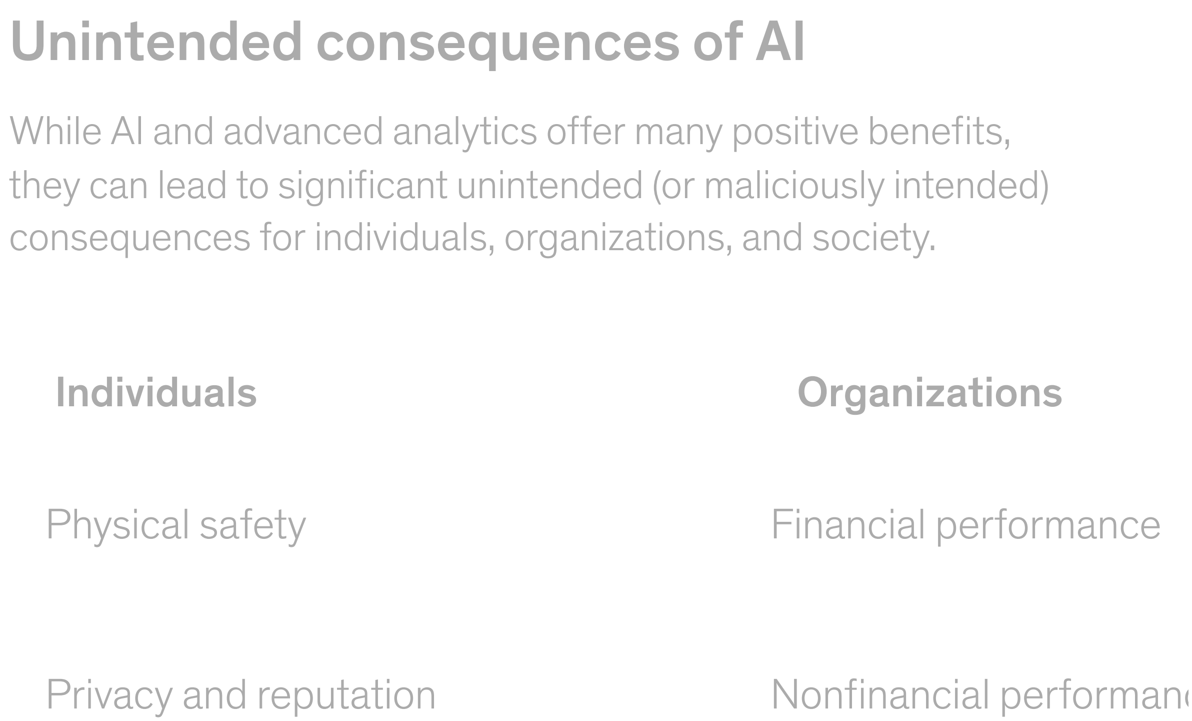
Before continuing, we want to underscore that our focus here is on first-order consequences that arise directly from the development of AI solutions, from their inadvertent or intentional misapplication, or from the mishandling of the data inputs that fuel them. There are other important consequences, among which is the much-discussed potential for widespread job losses in some industries due to AI-driven workplace automation. There also are second-order effects, such as the atrophy of skills (for example, the diagnostic skills of medical professionals) as AI systems grow in importance. These consequences will continue receiving attention as they grow in perceived importance but are beyond our scope here.

Understanding the risks and their drivers

When something goes wrong with AI, and the root cause of the problem comes to light, there is often a great deal of head shaking. With the benefit of hindsight, it seems unimaginable that no one saw it coming. But if you take a poll of well-placed executives about the *next* AI risk likely to appear, you're unlikely to get any sort of a consensus.

Leaders hoping to shift their posture from hindsight to foresight need to better understand the types of risks they are taking on, their interdependencies, and their underlying causes. To help build that missing intuition, we describe below five pain points that can give rise to AI risks. The first three—data difficulties, technology troubles, and security snags—are related to what might be termed enablers of AI. The final two are linked with the algorithms and human–machine interactions that are central to the operation of the AI itself. Clearly, we are still in the early days of understanding what lies behind the risks we are taking on, whose nature and range we've also sought to catalog in Exhibit 1.

Exhibit 1



Unintended consequences of AI

While AI and advanced analytics offer many positive benefits, they can lead to significant unintended (or maliciously intended) consequences for individuals, organizations, and society.

Data difficulties. Ingesting, sorting, linking, and properly using data has become increasingly difficult as the amount of unstructured data being ingested from sources such as the web, social media, mobile devices, sensors, and the Internet of Things has increased. As a result, it's easy to fall prey to pitfalls such as inadvertently using or revealing sensitive information hidden among anonymized data. For example, while a patient's name might be redacted from one section of a medical record that is used by an AI system, it could be present in the doctor's notes section of the record. Such considerations are important for leaders to be aware of as they work to stay in line with privacy rules, such as the European Union's General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA), and otherwise manage reputation risk.

Technology troubles. Technology and process issues across the entire operating landscape can negatively impact the performance of AI systems. For example, one major financial institution ran into trouble after its compliance software failed to spot trading issues because the data feeds no longer included all customer trades.

Security snags. Another emerging issue is the potential for fraudsters to exploit seemingly nonsensitive marketing, health, and financial data that companies collect to fuel AI systems. If security precautions are insufficient, it's possible to stitch these threads together to create false identities. Although target companies (that may otherwise be highly effective at safeguarding personally identifiable information) are unwitting accomplices, they still could experience consumer backlash and regulatory repercussions.

Models misbehaving. AI models themselves can create problems when they deliver biased results (which can happen, for example, if a population is underrepresented in the data used to train the model), become unstable, or yield conclusions for which there is no actionable recourse for those affected by its decisions (such as someone denied a loan with no knowledge of what they could do to reverse the decision). Consider, for example, the potential for AI models to discriminate unintentionally against protected classes and other groups by weaving together zip code and income data to create targeted offerings. Harder to spot are instances when AI models are lurking in software-as-a-service (SaaS) offerings. When vendors introduce new, intelligent features—often with little fanfare—they are also introducing models that could interact with data in the user's system to create unexpected risks, including giving rise to hidden vulnerabilities that hackers might exploit. The implication is that leaders who believe they are in the clear if their organization has not purchased or built AI systems, or is only experimenting with their deployment, could well be mistaken.

Interaction issues. The interface between people and machines is another key risk area. Among the most visible are challenges in automated transportation, manufacturing, and infrastructure systems. Accidents and injuries are possibilities if operators of heavy equipment, vehicles, or other machinery don't recognize when systems should be overruled or are slow to override them because the operator's attention is elsewhere—a distinct possibility in applications such as self-driving cars. Conversely, human judgment can also prove faulty in overriding system results. Behind the scenes, in the data-analytics organization, scripting errors, lapses in data management, and misjudgments in model-training data easily can compromise fairness, privacy, security, and compliance. Frontline personnel also can unintentionally contribute, as when a sales force more adept at selling to certain demographics inadvertently trains an AI-driven sales tool to exclude certain segments of customers. And these are just the *unintended* consequences. Without rigorous safeguards, disgruntled employees or external foes may be able to corrupt algorithms or use an AI application in malfeasant ways.

AI risk management: Three core principles

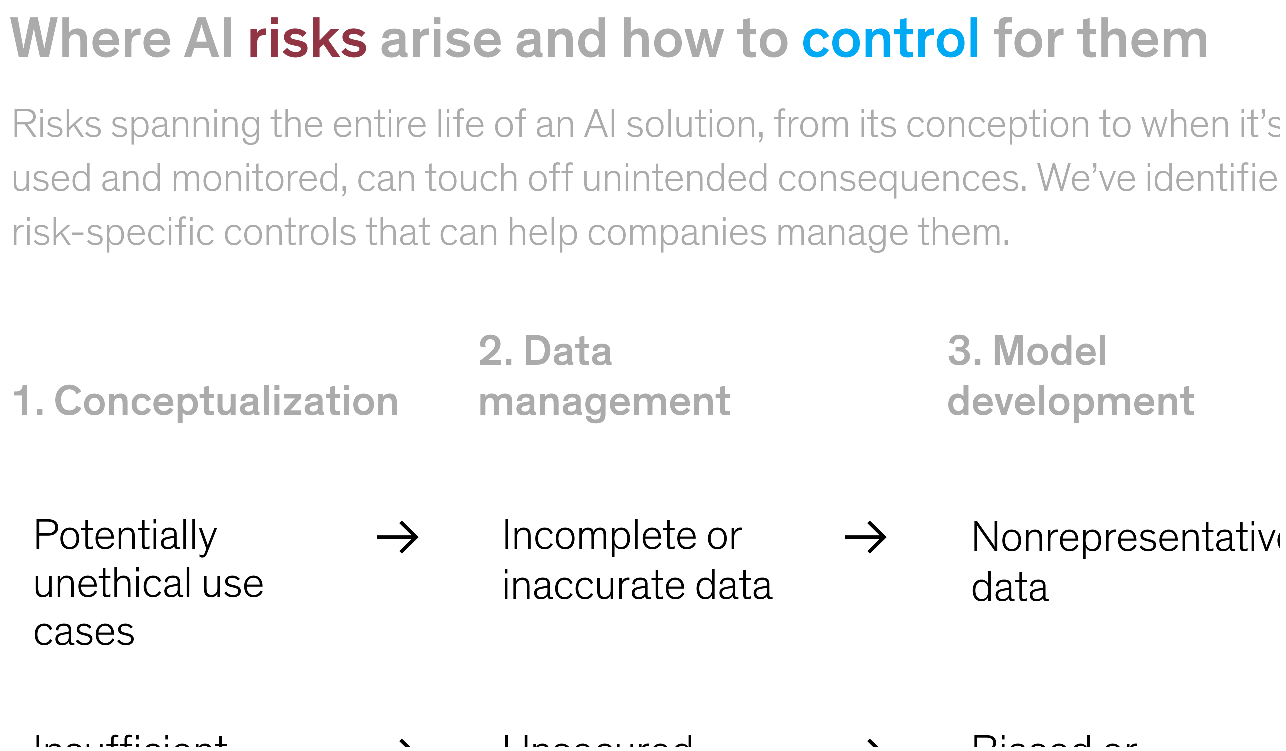
In addition to providing a flavor of the challenges ahead, the examples and categorization above are useful for identifying and prioritizing risks and their root causes. If you understand where risks may be lurking, ill-understood, or simply unidentified, you have a better chance of catching them before they catch up with you.

But you'll need a concentrated, enterprise-wide effort to move from cataloging risks to rooting them out. The experiences of two leading banks help illustrate the clarity, breadth, and nuanced rigor that's needed. The first, a European player, has been working to apply advanced-analytics and AI capabilities to call-center optimization, mortgage decision making, relationship management, and treasury-management initiatives. The second is a global leader, seeking to apply a machine-learning model to its customer-credit decisions.

These banks, like many others in the financial-services sector, have been applying some form of advanced analytics for a number of years, dating back to their early use in credit-card fraud detection and equity trading. They also are subject to a high degree of regulatory oversight and therefore have long been applying and making transparent a wide range of protocols and controls for mitigating the related risks—including cybersecurity risk, where they are frequently on the front lines given the obvious attractiveness of their assets to attackers.

Nonetheless, these banks' stories only illustrate a subset of the risk-specific controls organizations should be considering. Exhibit 2 presents a more complete list of potential controls, spanning the entire analytics process, from planning to development to subsequent use and monitoring. Our hope is that taken together, the tool and examples will help leaders who must confront a wide range of issues—from avoiding bias in recommendation engines to eliminating personal-identity risk to better tailoring the responses of customer-service bots to the needs of specific customers, and many more beyond.

Exhibit 2



Where AI risks arise and how to control for them

Risks spanning the entire life of an AI solution, from its conception to when it’s used and monitored, can touch off unintended consequences. We’ve identifie risk-specific controls that can help companies manage them.

1. Conceptualization

2. Data management

3. Model development

Potentially unethical use cases → Incomplete or inaccurate data → Nonrepresentativ data

Insufficient → Unsecured → Biased or

Clarity: Use a structured identification approach to pinpoint the most critical risks

The European bank’s COO started by assembling leaders from business, IT, security, and risk management to evaluate and prioritize its greatest risks. Inputs to this exercise included a clear-eyed look at the company’s existing risks and how they might be exacerbated by AI-driven analytics efforts under consideration, and at new risks that AI enablers, or the AI itself, could create. Some were obvious, but others less so. One that unexpectedly neared the top of the list was the delivery of poor or biased product recommendations to consumers. Such flawed recommendations could result in a significant amount of harm and damage, including consumer losses, backlash, and regulatory fines.

What the bank's leaders achieved through this structured risk-identification process was clarity about the most worrisome scenarios, which allowed them to prioritize the risks encompassed, to recognize controls that were missing, and to marshal time and resources accordingly. Those scenarios and prioritized risks will naturally vary by industry and company. A food manufacturer might prioritize contaminated-product scenarios. A software developer might be particularly concerned about disclosure of software code. A healthcare organization might focus on issues such as patient misdiagnosis or inadvertently causing harm to patients. Getting a diverse cross-section of managers focused on pinpointing and tiering problematic scenarios is a good way both to stimulate creative energy and to reduce the risk that narrow specialists or blinkered thinking will miss major vulnerabilities. Organizations need not start from scratch with this effort: over the past few years, risk identification has become a well-developed art, and it can be directly deployed in the context of AI.

Breadth: Institute robust enterprise-wide controls

Sharpening your thinking about show-stopping risks is only a start. Also crucial is the application of company-wide controls to guide the development and use of AI systems, ensure proper oversight, and put into place strong policies, procedures, worker training, and contingency plans. Without broad-based efforts, the odds rise that risk factors such as the ones described previously will fall through the cracks.

Concerned with the potential risk from poor or biased product recommendations, the European bank began adopting a robust set of business principles aimed at detailing how and where machines could be used to make decisions affecting a customer's financial health. Managers identified situations where a human being (for example, a relationship manager or loan officer) needed to be "in the loop" before a recommendation would be delivered to the customer. These workers would provide a safety net for identifying if a customer had special circumstances, such as the death of a family member or financial difficulties, that might make a recommendation ill-timed or inappropriate.

The bank's oversight committee also conducted a gap analysis, identifying areas in the bank's existing risk-management framework that needed to be deepened, redefined, or extended. Thorough and consistent governance at the bank now ensures proper definition of policies and procedures, specific controls for AI models, core principles (supported by tools) to guide model development, segregation of duties, and adequate oversight. For example, model-development tools ensure that data scientists consistently log model code, training data, and parameters chosen throughout the development life cycle. Also adopted were standard libraries for explainability, model-performance reporting, and monitoring of data and models in production. This governance framework is proving invaluable both for in-house AI-development efforts and for evaluating and monitoring third-party AI tools such as an SaaS fraud model the bank had adopted.

In addition, bank policies now require all stakeholders, including the sponsoring business executives, to conduct scenario planning and create a fallback plan in case AI model performance drifts, data inputs shift unexpectedly, or sudden changes, such as a natural disaster, occur in the external environment. These fallback plans are included in the bank's regular risk-review process, giving the board's risk committee visibility into the steps being taken to mitigate analytics-driven and AI-related risks.

Worker training and awareness are also prominent in the bank's risk-mitigation efforts. All affected employees receive comprehensive communications about where AI is being used; what steps the bank is taking to ensure fair and accurate decisions and to protect customer data; and how the bank's governance framework, automated technology, and development tools work together. Additionally, business sponsors, risk teams, and analytics staff receive targeted training on their role in identifying and minimizing risks. For instance, business sponsors are learning to request explanations on model behavior, which they are using to provide feedback on business assumptions behind the model. Meanwhile, the risk team has trained up on how to better identify and mitigate legal and regulatory-compliance issues, such as potential discrimination against protected groups or compliance with GDPR.

Monitoring AI-driven analytics is an ongoing effort, rather than a one-and-done activity. As such, the bank's oversight groups, including the board's risk committees, regularly review the program to stay on top of new risks that might have emerged as a result of regulatory changes, industry shifts, legal interpretations (such as emerging GDPR case law), evolving consumer expectations, and rapidly changing technology.

Nuance: Reinforce specific controls depending on the nature of the risk

Important as enterprise-wide controls are, they are rarely sufficient to counteract every possible risk. Another level of rigor and nuance is often needed, and the requisite controls will depend on factors such as the complexity of the algorithms, their data requirements, the nature of human-to-machine (or machine-to-machine) interaction, the potential for exploitation by bad actors, and the extent to which AI is embedded into a business process. Conceptual controls, starting with a use-case charter, sometimes are necessary. So are specific data and analytics controls, including transparency requirements, as well as controls for feedback and monitoring, such as performance analysis to detect degradation or bias.

Our second example sheds valuable light on the application of nuanced controls. This institution wanted visibility into how, exactly, a machine-learning model was making decisions for a particular customer-facing process. After carefully considering transparency requirements, the institution decided to mitigate risk by limiting the types of machine-learning algorithms it used. Disallowing certain model forms that were overly complex and opaque enabled the institution to strike a balance with which it was comfortable. Some predictive power was lost, which had economic costs. But the transparency of the models that *were* used gave staff higher confidence in the decisions they made. The simpler models also made it easier to check both the data and the models themselves for biases that might emerge from user behavior or changes in data variables or their rankings.

As this example suggests, organizations will need a mix of risk-specific controls, and they are best served to implement them by creating protocols that ensure they are in place, and followed, throughout the AI-development process. The institutions in our examples implemented those protocols, as well as enterprise-wide controls, at least in part, through their existing risk infrastructure. Companies that lack a centralized risk organization can still put these AI risk-management techniques to work using robust risk-governance processes.

There is much still to be learned about the potential risks that organizations, individuals, and society face when it comes to AI; about the appropriate balance between innovation and risk; and about putting in place controls for managing the unimaginable. So far, public opinion and regulatory reaction has been relatively tempered.

But this is likely to change if more organizations stumble. As the costs of risks associated with AI rise, the ability both to assess those risks and to engage workers at all levels in defining and implementing controls will become a new source of competitive advantage. On the horizon for many organizations is a reconceptualization of "customer experience" to encompass the promise as well as the pitfalls of AI-driven outcomes. Another imperative is to engage in a serious debate about the ethics of applying AI and where to draw lines that limit its use. Collective action, which could involve industry-level debate about self-policing and engagement with regulators, is poised to grow in importance as well. Organizations that nurture those capabilities will be better positioned to serve their customers and society effectively; to avoid ethical, business, reputational, and regulatory predicaments; and to avert a potential existential crisis that could bring the organization to its knees.

About the author(s)

Benjamin Cheatham is a senior partner in McKinsey's Philadelphia office and leads QuantumBlack, a McKinsey company, in North America; **Kia Javanmardian** is a senior partner in the Washington, DC, office; and **Hamid Samandari** is a senior partner in the New York office.

The authors wish to thank Roger Burkhardt, Liz Grennan, Nicolaus Henke, Pankaj Kumar, Marie-Claude Nadeau, Derek Waldron, and Olivia White for their contributions to this article.