McKinsey
Analytics

Getting to know—and manage—your biggest AI risks

A systematic approach to identifying and prioritizing AI risks can help organizations effectively target mitigation efforts.

by Kevin Buehler, Rachel Dooley, Liz Grennan, and Alex Singla

© Getty Images

May 2021

Many organizations are generating significant value with artificial intelligence (AI) and recognize that this technology will shape the future. At the same time, organizations are discovering that AI could expose them to a fast-changing universe of risks and ethical pitfalls that regulators signal they'll be watching for—and potentially penalizing. Recently, the European Union proposed a set of AI regulations that, if violated, could result in material fines, and the US Federal Trade Commission (FTC) put out notice that it could hold organizations accountable for proliferating bias or inequities through AI.

Just as AI deployment will be core to organizations' future success, leading organizations will be those that actively identify and manage the associated risks. In our latest AI survey, respondents at organizations getting the most value from AI were more likely than others to recognize and mitigate the risks posed by the technology.[1]

The prospect of protecting against a wide and growing range of AI risks might seem overwhelming, but neither avoiding AI nor turning a blind eye to the risks is a viable option in today's competitive and increasingly digitized business environment. So where should companies start?

First, organizations must put business-minded legal and risk-management teams alongside the data-science team at the center of the AI development process. Waiting until after the development of AI models to determine where and how to mitigate risks is too inefficient and time consuming in a world of rapid AI deployments. Instead, risk analysis should be part of the initial AI model design, including the data collection and governance processes. Involving legal, risk, and technology professionals from the start enables them to function as a "tech trust team" that ensures the models conform to social norms and legal requirements while still delivering maximum business value.

Second, because there is no cure-all for the broad spectrum of AI risks, organizations must apply an informed risk-prioritization plan as the initial step in an effective, dynamically updated AI risk-management approach anchored in both legal guidance and technical best practices. As outlined in this article, such a prioritization plan entails creating a catalog of your organization's specific AI risks to define the harm you seek to avoid, and then following a clear methodology to evaluate and prioritize those risks for mitigation.

Identifying AI risks

To create a catalog of specific AI risks, the tech trust team clearly delineates each negative event that could result from a particular AI deployment, so the team can then detail how each of those risks will be mitigated in accordance with the appropriate standards. A helpful way to think through the potential risks is to use a six-by-six framework, mapping risk categories against possible business contexts (exhibit).

We recommend that this process consider at least six overarching types of AI risk:

1. *Privacy.* Data is the lifeblood of any AI model. Privacy laws around the world mandate how companies may (and may not) use data, while consumer expectations set normative standards. Running afoul of these laws and norms can result in significant liability, as well as harm to consumers. Violating consumer trust, even if the data use was technically lawful, can also lead to reputation risk and a decrease in customer loyalty.

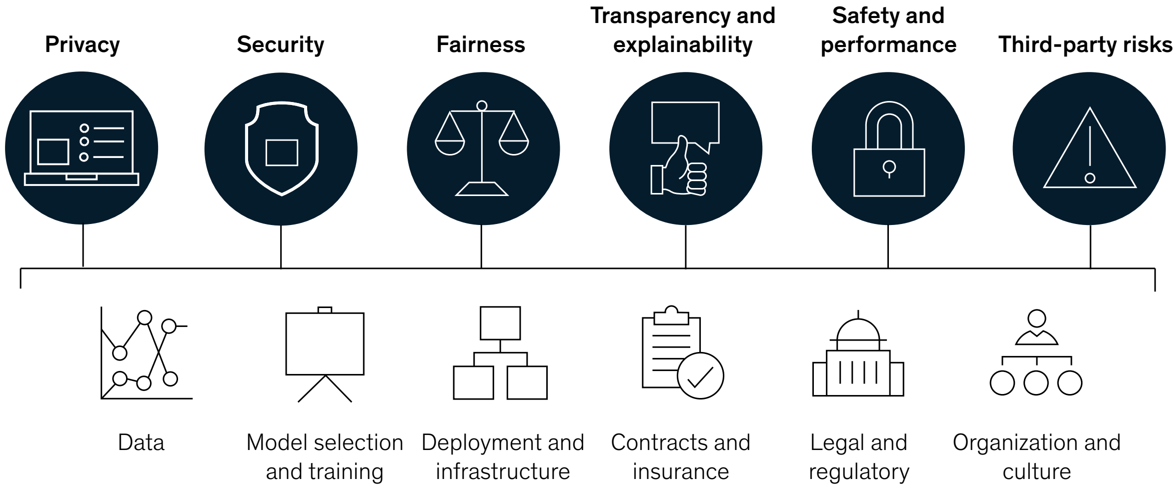
2. *Security.* New AI models have complex, evolving vulnerabilities that create both novel and familiar risks. Vulnerabilities such as model extraction and data poisoning (in which "bad" data are introduced into the training set, affecting the model's output) can pose new challenges to long-standing security approaches. In many cases, existing legal frameworks mandate minimum security standards to meet.

3. *Fairness.* It can be easy to inadvertently encode bias in AI models or introduce bias lurking in the data feeding into the model. Bias that potentially

[1] "The state of AI in 2020," November 2020, McKinsey.com.

Exhibit

A systematic approach to identifying AI risks examines each category of risk in each business context.



or actually harms particular classes and groups can expose the company to fairness risks and liabilities.

4. *Transparency and explainability.* A lack of transparency around how a model was developed (such as how data sets feeding into a model were combined) or the inability to explain how a model arrived at a particular result can lead to issues, not the least of which is potentially running afoul of legal mandates. For example, if a consumer initiates an inquiry into how his or her data were used, the organization using the data will need to know into which models the data were fed.

5. *Safety and performance.* AI applications, if not implemented and tested properly, can suffer performance issues that breach contractual guarantees and, in extreme cases, pose threats to personal safety. Suppose a model is used to ensure timely updates of machinery in manufacturing or mining; a failure of this model could constitute negligence under a contract and/or lead to employee harm.

6. *Third-party risks.* The process of building an AI model often involves third parties. For example, organizations may outsource data collection, model selection, or deployment environments. The organization engaging third parties must know and understand the risk-mitigation and governance standards applied by each third party, and it should independently test and audit all high-stakes inputs.

In our experience, most AI risks map to at least one of the overarching risk types just described, and they often span multiple types. For example, a model-extraction attack, in which a model is stolen based on a sample set of outputs, compromises both the privacy and security of the model. Therefore, organizations should ask if each category of risk could result from each AI model or tool the company is considering or already using.

Pinpointing the context in which these risks can occur can help provide guidance as to where mitigation measures should be directed. We identify six such contexts:

1. *Data*. Risks can surface through the way data get captured and collected, the extraction of the data's features (or fields and characteristics), and how data are engineered to train the model.

2. *Model selection and training.* Models are evaluated, selected, and trained based on various criteria, involving choices that present grounds for risk. For example, some models are more transparent than others. While a relatively opaque model might offer better performance, a legal requirement for transparency could necessitate use of a different model.

3. *Deployment and infrastructure.* Models are pushed to production, or deployed, when ready for real-world use. This process and the underlying infrastructure supporting it present risks. For example, the model might fail to perform in the real world as demonstrated by its performance in a lab environment.

4. *Contracts and insurance*. Contractual and insurance guarantees often explicitly address some AI risks. Product and service providers (both B2B and B2C), for example, may include in their service-level agreements parameters around model performance or delivery. Insurance providers might assign liability for incidents such as security or privacy breaches. Such contractual provisions must be surfaced, so teams can ensure they are in compliance.

5. *Legal and regulatory.* Industries, sectors, and regions around the world have varying standards and laws regarding privacy, fairness, and other risks presented in this framework. Therefore, it's important to be aware of applicable laws and regulations based on where, how, and in what sector the model will be deployed.

6. *Organization and culture.* Consider the risk maturity and culture of your organization and how that might affect the way a model or its components are used (for example, the security of the data flowing into it). Broader efforts, such as training programs, resource allocation, and interdisciplinary collaboration among cross-functional teams (such as a tech trust team) play key roles in mitigating risk. To consider the types and likelihood of risks that might arise, it's important to know if these exist, and at what level.

In addition to using this framework to brainstorm risks, the team could consult public databases of previous AI incidents (for example, Partnership on AI's incident database). This review of past risk failures is helpful in AI risk mitigation because AI draws its predictive power from past events. If AI created liability in the past under a similar set of circumstances, there is a higher probability it will do so again.

Further, the risk experts on the tech trust team should examine current use cases within the organization and conduct a "red team" challenge, incentivizing team members to uncover less obvious risks. These could be, for example, risks that flow from second-order model risks, such as risks arising from how the model might be used in practice (as opposed to the first-order risks that arise from how the model was built).

Evaluating and prioritizing the risk catalog

With risks clearly defined and logged, the next step is to assess the catalog for the most significant risks and sequence them for mitigation. With an overwhelming number of potential AI risks, it's impossible to fully mitigate each one. By prioritizing the risks most likely to generate harm, organizations can help prevent AI liabilities from arising—and mitigate them quickly if they do. A strong methodology for such sequencing allows AI practitioners and legal personnel to triage AI use cases so they can focus their often-limited resources on those meriting the most attention.

Existing regulatory frameworks and case law can provide a solid foundation for this prioritization methodology. Liability frameworks across a variety of areas—such as privacy law and anti-discrimination and negligence standards—have been forged over decades to confront many of the practical constraints that data scientists face today. We'll briefly highlight two standards in US laws that can inform successful risk methodologies. The central insight to draw from these legal frameworks is that risks tagged for mitigation should be identified in relation to the likelihood or significance of an incident *and* the costs of a viable alternative approach.

The first standard derives from a concept familiar to most lawyers in the United States: the "Hand formula," which has been widely influential in shaping negligence standards in the United States.[2] According to the formula, risk is defined as the probability of the harmful event occurring multiplied by the loss the event could generate. Liability ensues any time the burden of preventing an incident is less than the harm the incident could cause.

Take, for example, a hypothetical manufacturer of automated long-haul trucks, which is concerned about the potential for migratory herds to wander onto roads in certain Western states. After performing internal testing, the company estimates an approximate cost of $10 million to gather sufficient data on herds and build an AI model to recognize and avoid the animals. The company also estimates that, at scale, if it did not have a model to help predict where herds might wander onto the roadway, its trucks would be involved in approximately 20 herd-induced accidents per year with an average cost of $100,000 each, increasing to $4 million each in cases with fatalities. Traditional negligence theory, as dictated by the Hand formula, would suggest the company invest in gathering herd-related training data and model creation, because the financial burden of doing so ($10 million) is well under the cost incurred by the herd-induced accidents ($21.5 million if 25 percent result in fatalities). While this example is meant to be simple and straightforward, the same type of risk assessment can help even in cases where harms do not have clear economic measures.

A similar significance standard underlies US anti-discrimination laws, which govern decision making in credit, housing, employment, and other contexts. The standard doesn't mandate perfect fairness or prohibit any harm from occurring; as any data scientist knows, there is no such thing as a perfectly unbiased model. Instead, US laws generally mandate discrimination testing such that no fairer alternative is available, given the intended use case. In plain terms, this means that if data scientists could have trained a model with similar business value and less discrimination but failed to do so, liability could ensue.

A helpful way to start prioritizing your risks is to take a domain-based approach, much as organizations apply AI most successfully. Apply the methodology we describe here to the AI use cases within a core process, journey, or function, and then move onto the next domain. This can save your organization time and money as you reduce risk systematically and efficiently.

Foundations for managing any AI risk

While the list of emerging best practices around AI risk mitigation is long and growing, a couple of practices will prove foundational in enabling the cataloging and weighting of AI risks:

— *Standard practices and model documentation.* Data science, legal, and risk personnel must understand important aspects of AI models and the environment in which they are trained and deployed, so they can surface areas of risk. Simply put, organizations cannot control an AI environment they don't understand. Setting clear standards for model development and documentation can help provide such foundational transparency. Standardized policies for steps in the development life cycle—recording data provenance, managing metadata, mapping data, creating model

[2] Justice Learned Hand (1872–1961) served as a federal district and appellate judge for more than 50 years and had enormous influence on the understanding of the law in the United States, especially of the First Amendment of the US Constitution.

inventories, and more—ensure development teams follow the same sound, approved processes. This in and of itself reduces risk, in addition to providing risk experts in the tech trust team with a map of the AI environment. Consistent model documentation provides another layer of transparency. If individual data scientists document their models differently at varying phases of the AI life cycle, for example, it is harder to do an apples-to-apples comparison of AI models across the organization so that risk standards and reviews can be more easily and consistently applied. An incredibly competitive market for data scientists means it's rare that the same data scientists who trained a model are still around to answer questions when an incident occurs.

— *Independent review.* Existing and proposed AI regulations call for different types of reviews or audits to demonstrate compliance. Some regulators, such as the FTC, explicitly recommend the use of independent standards or expertise to evaluate model fairness, performance, and transparency. As the FTC declared publicly in April 2020, organizations should "consider how [to] hold yourself accountable, and whether it would make sense to use independent standards or independent expertise to step back and take stock of your AI."[3] This year, the FTC added that it could step in to hold companies accountable for AI issues and mandated the deletion of *any* models containing misused customer data, suggesting that serious oversight lies ahead. Systematic internal and external audits can contribute to a solid compliance program for an organization's AI.

Based on recent headlines alone, it's clear that global efforts to manage the risks of AI are just beginning. Therefore, the earlier organizations adopt concrete, dynamic frameworks to manage AI risks, the more successful their long-term AI efforts will be in creating value without material erosion. The approaches we have described enable organizations to begin pinpointing and prioritizing the management of AI risks *right now* as part of a holistic long-term strategy for managing AI risks.

Kevin Buehler is a senior partner in McKinsey's New York office, where **Rachel Dooley** is associate general counsel and **Liz Grennan** is managing counsel; **Alex Singla** is a senior partner in the Chicago office.

The authors wish to thank Andrew Burt, managing partner at bnh.ai and visiting fellow at Yale Law School's Information Society Project, and Patrick Hall, principal scientist at bnh.ai and visiting faculty at George Washington University School of Business, for their contributions to this article.

[3] Federal Trade Commission blog, "Using artificial intelligence and algorithms," blog entry by Andrew Smith, April 8, 2020, ftc.gov.