How AI Raises the Stakes for Trust and Risk

AI's potential to solve difficult problems excites those who see it ushering in a new era comparable to the Industrial Revolution of the 18th century. At the same time, the implication that humans are granting inanimate technology a kind of agency and autonomy — and even some degree of personality, as in the case of automated personal assistants who have names and speak to us — creates anxiety for many. While some of those worries belong in the realm of science fiction, using AI does carry real potential risks that organizations must manage.

Precisely because the technology can enable autonomous decision-making and actions by machines, questions about its reliability are more urgent than for other technologies. To establish a baseline for trust in AI, we asked our survey group to rate just how reliable they find the results and recommendations of AI-based systems in two different contexts: in their personal lives, as consumers; and at work, interacting with their organizations' AI-based systems. On a scale of 1 to 10, with 10 being most reliable, respondents' overall rating for personal AI technology was 7, while AI used in their organizations earned a 6.

While those ratings lean positive, they demonstrate that trust in AI is at best guarded. This cautious approach was confirmed when we measured people's degree of concern about eight commonly discussed AI risks and found that most were rated fairly highly — 7 on a 10-point scale (with 10 representing the highest concern). The six risks of high concern were that AI may:

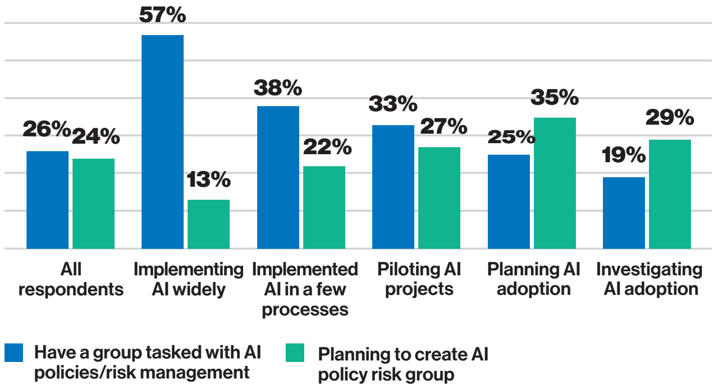
- Deliver inadequate return on investment.
- Produce bad information.
- Be used unethically.
- Support biased, potentially illegal decisions.
- Produce results that humans cannot explain.
- Be too unpredictable to manage adequately.

The two risks eliciting less concern were that AI may disrupt workflows or productivity (rated 3) and that AI may deliver bad customer experiences (rated 4).

New Risks Call For New Risk Management Structures

An important component of building trust in AI is managing the associated risks — particularly through oversight that seeks to understand and verify how models function, mitigate bias, and anticipate unintended consequences. How are respondents doing this? About half are acting to create organizational structures to manage AI risk: 26% have a group that sets policies and manages AI risk, and 24% plan to create one. Organizations that have implemented AI are much more likely to already have a group tasked with setting policies and managing AI risk: 57% of those that have implemented AI broadly in their enterprises do, as do 38% of the point implementers (see Figure 10, "AI Risk

Figure 10: AI Risk Management Structures Are Emerging



As AI becomes more widely implemented, it appears to be driving the creation of groups to develop policies around its use and manage associated risks.

Trust, Risk, and Culture

Our global survey included respondents from all regions, and we found almost no significant variance in response patterns by geography. The one exception was to the questions about AI trust and risk. Europeans surveyed expressed markedly less trust in AI systems used in their personal lives. And North Americans are much more concerned that AI may deliver bad customer experiences, rating that as highly as most of the other risks identified above.

Management Structures Are Emerging"). Large companies of 5,000 or more employees are also more likely to act on AI risk: 34% have set up a group to do this.

According to our survey, the responsibility for managing AI risk is as likely to fall to the CIO, CTO, or CEO as it is to be shared; very few organizations appear to be assigning accountability along traditional lines of risk management, to legal or financial executives. Those implementing AI are more likely to say the responsibility is shared, which is perhaps indicative of their experience working more cross-functionally with this technology. Very few overall have any plans for remediating harm caused by AI applications (see Figure 11, "Who Has a Plan to Remediate Harm Caused by an AI Application?").

An important development among companies that are implementing AI is establishing a management-review mechanism, says Peter Guerra, North America chief data scientist at Accenture.

"The big thing that I'm seeing is a lot of clients, especially those that are highly regulated, are putting together boards that review anything that they want to operationalize from an AI perspective," Guerra says. These boards often have data scientists as well as business leaders, representatives from the legal department, and other relevant experts.

> **"We need to be able to defend our models and how we made those decisions in front of a regulator, which happens often, actually. Therefore, we can't afford not to do this."**
>
> ERIC MONTEIRO, SUN LIFE

Eric Monteiro, senior vice president and chief client experience officer at Sun Life, says the global financial services and insurance company has established a three-layered approach to govern its use of AI and mitigate the risk of bias in its applications. First, Sun Life provides training to its model developers that includes guidelines and standards to test for bias. Second, a team of experts evaluates every model to examine its behavior after the testing phase, to determine whether the inclusion of additional data introduces bias. Third, Sun Life has another AI model-evaluation team set up specifically for what Monteiro calls "more critical models," such as those that impact the company's financial results, to provide an additional layer of scrutiny.

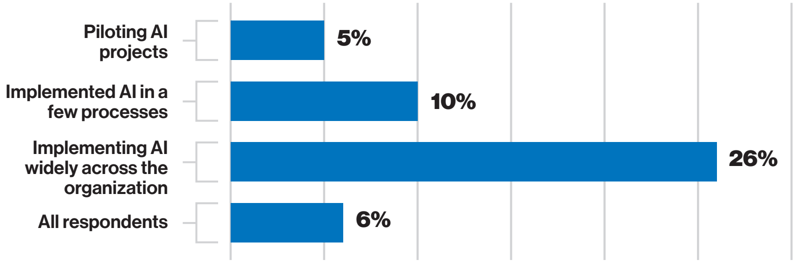
Monteiro says Sun Life takes model validation seriously because it's a regulated financial institution. "We need to be able to defend our models and how we made those decisions in front of a regulator, which happens often, actually. Therefore, we can't afford not to do this," he says.

Linda Zeger, founder and principal consultant at the data analytics and system design consultancy Auroral LLC, says that maintaining documentation about AI projects — including a model's scope, limitations, appropriate and inappropriate uses, and accuracy, as well as notes about users' interactions with it — is an important way to mitigate risks. A company can provide documentation not only internally, but also to partners, customers, and end users to help people understand what an algorithm can and can't do and how they should use it.

"I think one of the greatest risks — and this is why communication is so important — is that the algo-

Figure 11: Who Has a Plan to Remediate Harm Caused by an AI Application?

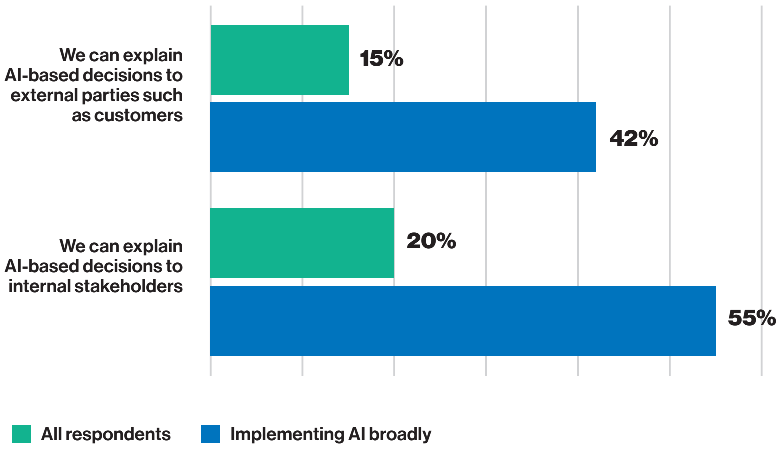
We have a plan for remediating harm that AI applications might cause



AI applications have the potential to cause harm if they malfunction or recommend poor decisions — but only 1 in 4 of the most advanced practitioners has a plan for remediating such harm.

rithm gets used where it's not applicable and where it wasn't designed for," Zeger says. She cites the example of a self-driving car that is tested in some conditions (such as typical weather conditions) but not others (such as unusually heavy smoke). "That is one reason I emphasize that it's really important that the designers of the system document the limitations and scope and accuracy and that the end users receive that documentation — to say, 'This system has only been designed for typical conditions.' It's almost like a black-box warning that's really obvious to the end user," she says.

Figure 12: Ability to Explain AI Results Gains Traction at the Leading Edge



Understanding — and communicating — the logic behind how an AI-based system has produced a result is key to building trust in AI and ensuring that decisions aren't biased.

The Quest for Explainability

An issue that is new to many who consider technology risk management is adequate explainability — the ability to identify the key factors that a model used in producing its results, which may be recommendations in a decision-support system or actions in an automated process. In our survey, the most action on explainability of AI results is being taken by those with the broadest AI implementations. Among this group, 55% make explanations available to internal stakeholders, and 42% make explanations available to external parties such as customers (see Figure 12, "Ability to Explain AI Results Gains Traction at the Leading Edge").

Explainability is important for regulatory compliance in use cases such as hiring or granting credit, where decisions must be shown to be free of illegal bias. It is also enshrined in the European Union's General Data Protection Regulation (GDPR), which gives individuals the right to know how their data has been used to reach a decision. This policy is echoed in emerging data regulations in other jurisdictions.

In addition to GDPR, Melvin Greer, chief data scientist for the Americas at Intel, points to Canada's federal privacy laws and moves in 26 U.S. states to develop laws like GDPR. "I think it's safe to say that we are moving into a regulatory and legal environment that is going to put more emphasis on data privacy and protection," Greer adds.

Heavily regulated industries such as financial services and health care are well ahead in understanding and managing explainability.

Monteiro says that Sun Life invests in its process for examining and validating models, and it experiments with new tools to increase its ability to describe the work of algorithms.

"If somebody came in for an insurance application and didn't get approved, we need to be able to tell the regulators why," he says. "Regulators want to make sure that it's not creating systemic bias or societal problems."

Monteiro says that emerging software tools are getting better at describing correlations that occur within the algorithms. "You can say with this model, 'The key variables that mattered for this decision were X, Y, Z. And the model took these into account and got to these outcomes,'" he says.

Explainability also plays a role in building trust in AI internally, by providing visibility into how a model works and why it is making a particular recommendation to business stakeholders. Because credit agency Equifax serves financial services companies with its AI-driven scoring tools, explainability is particularly important. Thus, the company has invested heavily in innovating effective AI models that are also explainable.

Vickey Chang, vice president for data and analytics at Equifax's U.S. Information Services unit, says the company previously used a neural network model that she regarded as a black box because the underlying logic could not be seen. Equifax subsequently developed an in-house machine learning model that provides the reasons behind a recommendation, such as why a given credit application was approved or denied. She says this newer model, which uses technology that Equifax has patented, outperforms its neural network predecessor.

At Telenor Group, explainability has become a steeper challenge as the Norwegian telco's AI group has advanced from machine learning to deep learning and its larger, more complex models. Astrid Undheim, vice president of analytics and AI at Telenor, says deep learning models are black boxes that she describes as being "too big to be able to explain properly." That means the company takes a cautious approach.

"You need to test the system much more rigorously to be able to trust it," she says. "I think another thing is that these models are not mature enough to be used in critical domains because of this black-box problem. So in my view, we need to start using them for problems with low risk and then, in that same process, learn, and also put focus on advancements in deep learning that involves, for instance, explainability and interpretability of the model. These are unsolved problems at the moment."

AI Puts Ethics Conversations on the Table

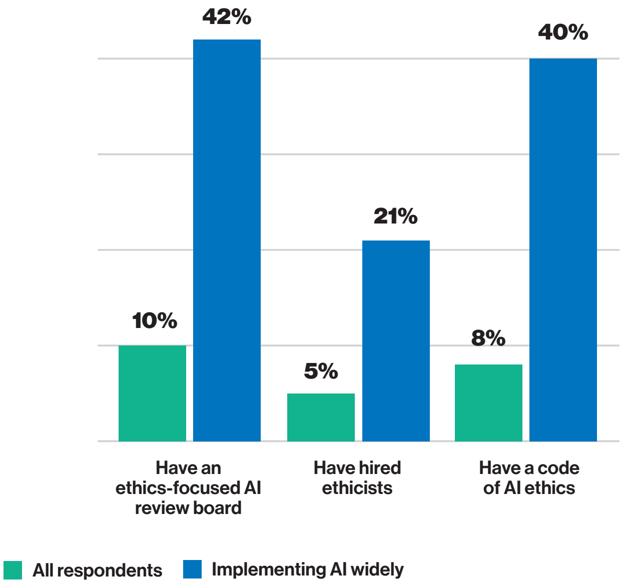
Making the workings of AI models more transparent is also aligned with efforts to ensure that this powerful technology is used ethically and that developers consider how a model developed for one purpose might potentially be misused for another. Advanced practitioners such as Capt. Michael J. Kanaan, cochair for AI at U.S. Air Force headquarters, believes that "if you are in the business of developing AI solutions, building them, or providing them to someone, you have a moral obligation to talk about these topics and the dual-use purposes."

While it's possible that ethics concerns are being addressed via broader AI review groups that the majority of respondents are committed to, just 10% of respondents to our survey have set up an ethics-focused AI review board, and 14% are planning to establish one. Those with broad implementations of AI are ahead on this issue: 42% report that they have an AI ethics oversight group, and 40% have adopted an AI code of ethics (see Figure 13, "Focus on AI Ethics"). Hiring for the emerging job of AI ethicist lags, with about 1 in 5 of the most advanced implementers having done so, compared with 1 in 20 of respondents overall.

DBS Bank in Singapore is one enterprise that has established principles and a process for evaluating proposed data use cases to ensure responsible data use. The review process is a natural outgrowth of the bank's experience with developing its data governance practices. DBS also decided to add AI capabilities to its board and increase its focus on using data responsibly, launching a Responsible Data Use Committee that brings together a diverse array of viewpoints to vet use cases.

Jeffery Lee, executive director of legal and compliance at DBS, says the bank's approach to ensuring that data use cases are appropriate builds on foundational data-governance principles: The baseline for any use case

Figure 13: Focus on AI Ethics



As with AI risk management overall, those with the most advanced AI practices have a much greater likelihood of having taken specific actions to apply ethical considerations to their use of the technology.

is that the relevant data is of good quality and will be used in a way that conforms to regulatory and security requirements. Then, proponents of a particular use case must be able to articulate how the proposed use of the data is purposeful, unsurprising (meaning that customers and employees would not be shocked to learn how data was being used), respectful to customers and employees, and explainable; this is where the Responsible Data Use Committee may be convened to consider the use case. Finally, there is a model governance process that looks at the development, testing, validation, documentation, deployment, communication, and ongoing review of a particular model.

This review process safeguards the bank's work in an emerging field by asking challenging questions, says Lam Chee Kin, DBS managing director and head of the bank's Legal, Compliance, and Secretariat group. "Why do we feel comfortable that this data use case is responsible? Can we hold our heads up high to society so that we're actually using data in a responsible way? This requires very, very broad and diverse societal perspectives on whether you should allow something or not," he says. "We've chosen to staff that committee along the lines of a broad cross section of perspectives. For example, our sustainability team is there, along with human resources, legal, and compliance. We have a country perspective as well, because what may be right for Singapore may not be right for India or China."

What's Working: Telenor Handles 'Black Boxes' With Care

As Telenor Group's AI practice has advanced to using deep learning with larger, more complex models that are difficult to explain, it has taken a cautious approach regarding which domains this technology is used in. Until model explainability and interpretability improve, Telenor believes deep learning is best applied to low-risk problems and advocates much more rigorous testing to build trust.

DBS's third stage of review, model governance, is a work in progress, says Lam. "How do you govern an AI [system], and do you in some situations have to disregard the AI? How do you determine whether an AI is fit for purpose? In what situations do you or do you not involve humans? All these questions are really important," he says.

"The reason for putting a ton of time and effort into this is that if we get it right, it is potentially a competitive advantage," Lam says. "If we get this correct, if you can develop a framework that can be applied broadly — when do you release an AI, how do you turn off the AI when it's going wrong, when do you involve humans, how do you deal with this properly? — if we get that part correct, that's a real differentiator."

Working to Mitigate Bias

Managing the risk of bias in AI applications is fundamental to sound data science, because inadequate data sets can skew a model that predicts a machine failure just as easily as they can distort a model with an impact on humans. However, the question of managing bias becomes more urgent with models that make recommendations affecting people, where the potential consequences of doing the wrong thing are greater. Identifying and mitigating bias risk requires organizations to scrutinize data sets for adequate diversity and to bring diverse points of view to the table as models are developed.

"No one at any of the companies that have had issues on this ethics piece got up in the morning and said, 'I'm going to make a discriminatory AI,'" says Kanaan of the U.S. Air Force. But such cases have "illuminated inherent biases that we have to address. Now we get to have that conversation."

Kanaan says efforts to mitigate bias start with understanding the purpose of any project and then evaluating whether there is appropriate data to build an unbiased model. He and other experts interviewed cite recruiting as a use case where it may be particularly difficult to assemble a data set that does not perpetuate bias. Training data based on past successful hires will almost inevitably reproduce biased decision-making, no matter how often a model is tweaked.

Intel’s Greer says he sees a significant increase in practitioners’ conversations around ethics, with bias being top of mind. “I think people are determined to try and get it right, which is really, really a positive,” he says.

Mitigating bias is a constant challenge, Greer says. “It does require some diligence and some forethought in order to ensure that we aren’t selecting data sets that simply reinforce our predefined conclusions.” Likewise, he wants to ensure that data analysts aren’t being drawn from a homogeneous population and that the tools they’re using aren’t inherently biased.

The work of reducing the risk of bias extends to the practice of one data science team adopting another’s work. “The common process is to adopt, then adapt, other people’s algorithms or training to models. However, in doing so, it can lead to this hereditary insertion of the same biases that were used to train the original models into this new analysis,” Greer says.

Greer also advocates for diverse participation in efforts seeking to establish principles for the ethical use of data (see “Frameworks for AI and Data Ethics”). He is involved with Data for Democracy, a group run by volunteers that is actively building a diverse community of people that includes professional data scientists, citizen data scientists, ethicists, and sociologists. The ideal, he says, is to assemble “a cross-functional diverse team of people who are able to look at the use of data and project ways to ensure that the guidelines and policies that are created benefit all communities.”

And ultimately, assembling similarly diverse data science teams, and ensuring that ethical guidelines are followed, will be essential to any organization if AI is to gain the trust needed to deliver on its promise. ●

MIT SMR Connections develops content in collaboration with our sponsors. It operates independently of the *MIT Sloan Management Review* editorial group.