
How can risk foresight lead to AI insight?
The better the question. The better the answer.
The better the world works.
EY
Building a better working world

“

Success in creating effective AI, could be the biggest event in the history of our civilization. Or the worst.

Unless we learn how to prepare for, and avoid, the potential risks, AI could be the worst event in the history of our civilization. It brings dangers, like powerful autonomous weapons, or new ways for the few to oppress the many. It could bring great disruption to our economy.

I am an optimist and I believe that we can create AI for the good of the world. That it can work in harmony with us. We simply need to be aware of the dangers, identify them, employ the best possible practice and management, and prepare for its consequences well in advance.

Stephen Hawking

1. Infusing Trust by Design into AI

When it can take only one mistake — or a perception of a mistake — for a user to stop trusting AI, how can you earn and sustain user trust?

Every challenge in business is an opportunity for AI. However, organizations are holding back in leveraging these opportunities because of mistrust in AI — and so being cautiously selective in where it is used.

In a recent EY-Microsoft European joint study, it was found that 71% of the respondents considered AI to be an important topic for executive management, and yet only 4% are using AI across multiple processes and to perform advanced tasks. [link]

There is risk inherent in any technology development effort, but even more so in AI which is still early in its innovation cycle. Plus, it involves a wide spectrum of technologies and is being considered for a broad set of use cases. Sustaining trust in AI will not be a 'one-off' exercise. It will be a long-term endeavor which will require state-of-the-art governance and control practices that stay in step with innovations in AI functionality.

Trust is the foundation on which organizations can build stakeholder confidence and active participation with their AI systems. However, in this era of instantly accessible information, mistakes can be costly, and second chances are harder to come by.

Trust by Design helps organizations embed a risk-optimization mindset across the AI lifecycle. It does this by elevating risk from a mere responsive function to a powerful, dynamic and future-facing enabler for building trust.

71% of the companies respond that AI is considered 'an important topic' on the executive management level

4% Only of the companies are actively using AI in 'many processes and to enable advanced tasks'

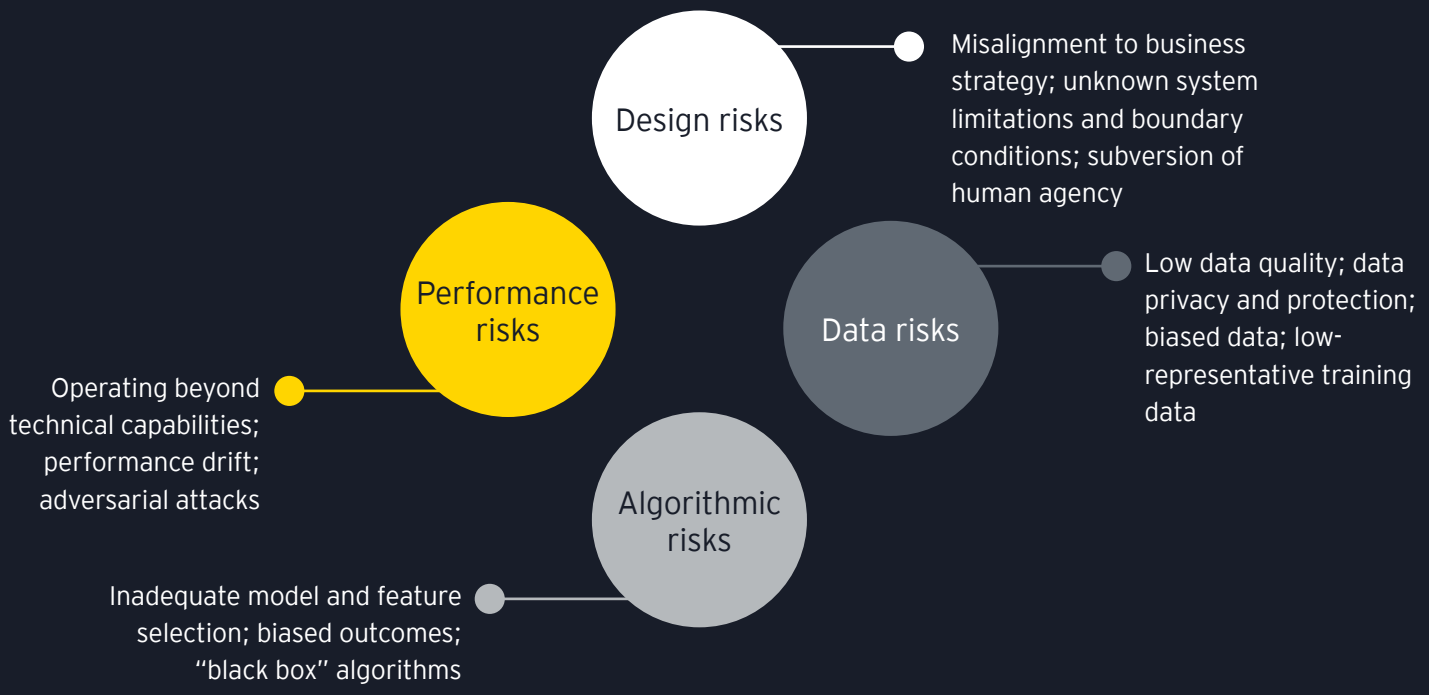
By infusing **Trust by Design** into their AI systems from the outset, organizations can leverage **risk foresight** to accelerate their access to **AI insights.**

2. Building Trust in AI

AI is introducing new risks and impacts that have historically been the purview of human decision-making, not technology development.

With the risks and impacts of AI spanning across technical, ethical and social domains, a new framework for identifying, measuring and responding to the risks of AI is needed — one that is built on the solid foundation of existing governance and control structures, but also introduces new mechanisms to address the unique risks of AI.

Risks of AI span four key areas:



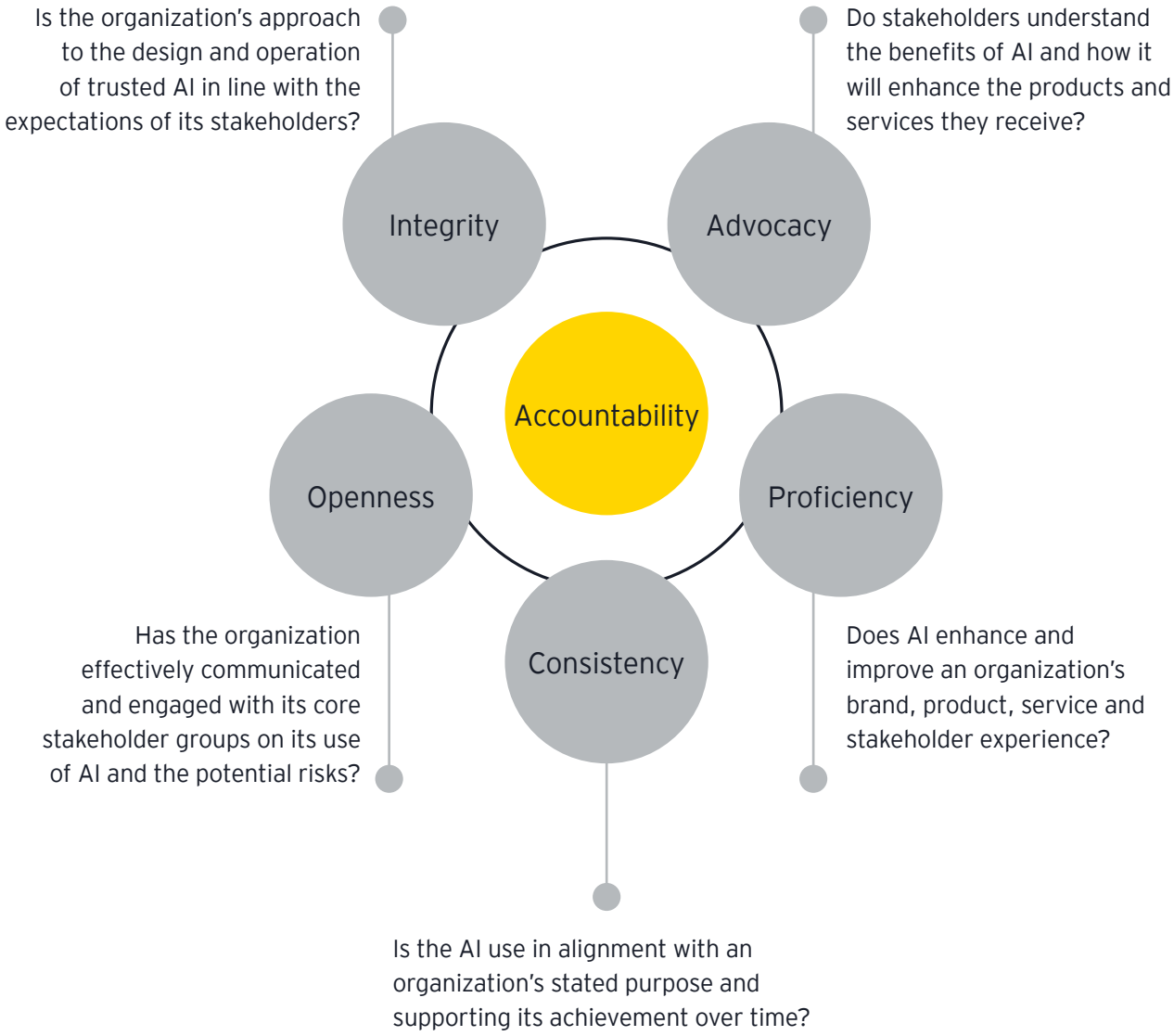
There is a growing consensus that AI should be safe, fair and interpretable and designed to be human-centric, but less agreement on how this can be achieved. Managing the risks of AI is not a passive exercise in which organizations can rely on traditional risk management practices. It's going to require new innovations in managing AI risks, including:

1. Data-based risk analytics
2. Adaptive learning in response to user feedback and model validation
3. Continuous monitoring and supervised response mechanisms
4. Data science infused into risk metrics and dashboards
5. Close collaboration between AI developers and risk professionals

Managing the risks of AI is about more than preventing reputational, legal and regulatory impacts. It's also about being considered trustworthy. With public discourse on AI heavily skewed to its risks, it will take time and active dialogue with stakeholders to build trust in AI systems.

Building trust in AI will take a coordinated approach, leveraging EY's Trust model across the five pillars of trust.

In establishing the five pillars of trust, the overarching element that connects them all is accountability.



Accountability is the foundation on which trust is built and is the inflection point at which an organization translates intentions into behaviors. Regardless of the level of autonomy for an AI system, ultimate responsibility and accountability for an algorithm needs to reside with a person or organization.

By embedding risk management into its design enablers and monitoring mechanisms for AI, organizations can demonstrate their commitment to accountability and for being held to account for AI systems predications, decisions and behaviors.

With understanding still evolving on how AI operates, and when and how risks could develop, many AI systems are considered high risk by default and approached with caution.

To counteract this response, EY has developed the Trusted AI Platform to help organizations quantify the impact and trustworthiness of their AI systems. With greater understanding on the risk profile of each of their AI systems and the drivers of risk, AI designers and operators can respond better to concerns raised by their stakeholders.

> “The rapid pace of disruption is requiring new risk management constructs to balance the opportunities and threats of AI. It’s about flipping our thinking from, “what could go wrong?” to “what has to go right?” Leading AI organizations recognize the role of trust as a competitive differentiator and are building Trust by Design into AI systems from the outset.

Amy Brachio
EY Global Risk and PI Advisory Leader

3. Quantifying the risks of AI

If AI is to reach its full potential, organizations need the ability to predict and measure conditions that amplify risks and undermine trust.

Understanding the drivers of risk in relation to AI requires consideration across a wide spectrum of contributing factors including its technical design, stakeholder impact and control maturity. Each one of these, in their design and operation, can affect the risk level of an AI system.

As an organization works to develop risk mitigation measures, it's important to understand each risk driver's general contribution to the creation or mitigation of risk, and to be able to quantify each driver's relative importance. A third consideration is understanding how the risk drivers interrelate and how this can amplify risk.

Developing an understanding of the risk drivers for an AI system is a complex undertaking. It requires careful consideration of potential stakeholder impacts across the full lifecycle of the AI system. Areas of considerationinclude:

- What goal (e.g., prediction) is the AI system design to achieve?
- What stakeholders are impacted by the AI operation and to what extent?
- What would be the impact of a predictive error by the AI system to stakeholders?
- What data is required for the AI system to make its prediction?
- How trustworthy are the data sources used?
- How complex is the AI system's technical capabilities?
- How stable is the environment the AI system operates within?
- What is the role of human operators in the decision-making process?
- How readily can human monitors understand the decision framework of the AI system?
- How mature are the controls over the AI system?

88%

don't know they are using AI but are using it in their core processes – potentially exposing them to unknown risks.

Narrative Science

In developing the **Trusted AI Platform**, EY distilled these considerations into three important components to managing the risks of an AI system:

1 Technical risk – evaluates the underlying technologies, technical operating environment and level of autonomy

2 Stakeholder impact – considers the goals and objectives of the AI agent and the financial, emotional and physical impact to external and internal users, as well as reputational, regulatory and legal risk

3 Control effectiveness – considers the existence and operating effectiveness of controls to mitigate the risks of AI

EY developed the Trusted AI Platform to provide an integrated approach to evaluate, quantify and monitor the impact and trustworthiness of AI.

The Trusted AI Platform uses interactive, web-based schematic and assessment tools to build the risk profile of an AI system, and then an advanced analytical model to convert the user responses to a composite score comprising technical risk, stakeholder impact and control effectiveness of an AI system. The technical risk score is subject to a complex multiplier based on the impact on stakeholders, taking into account unintended consequences such as social and ethical implications. An evaluation of governance and control maturity acts as a mitigating factor to reduce residual risk of an AI system.

> “Building and maintaining trust in AI will require investments into new, innovative risk management techniques, including integrated human-automated monitoring and rapid response mechanisms. It also requires going beyond an understanding of the potential risks and impacts and developing a deeper measurement system for risk drivers.

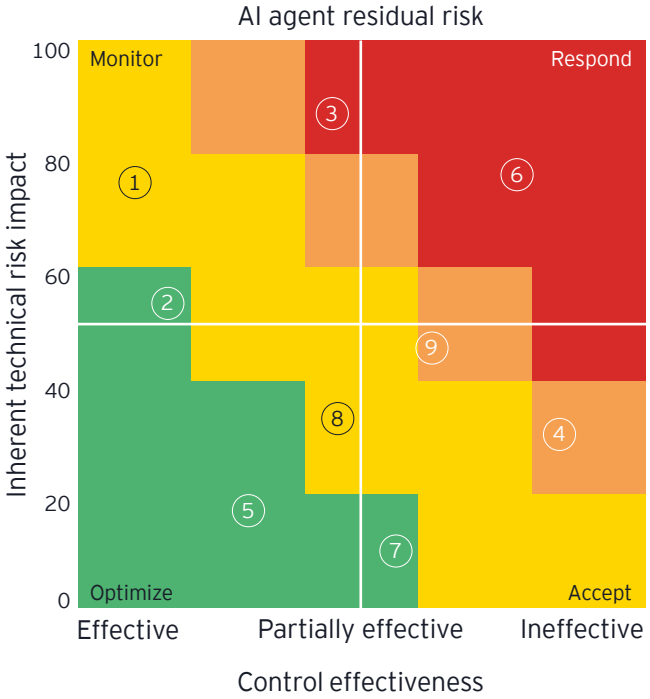
Cathy Cobey
EY Global Trusted AI Advisory Leader

The combination of these three scores enables the EY Trusted AI Platform to calculate the residual risk of an AI system's design, using the following formula:

$$\text{AI Residual risk} = \left[\text{Technical risk} \times \text{Stakeholder impact}\right] - \text{Control effectiveness}$$

The advantage of developing a quantitative score of an AI system's residual risk is that the residual risk can be mapped across its AI portfolio and against an organization's risk tolerance level.

Another key benefit of this type of model is the ability to drill down into the drivers of risk for each of the three components. In developing risk mitigation strategies, it is crucial to understand the source of the risk.



For example, is risk stemming from a decision to rely solely on automated decision-making with no human oversight, or is it the use of sensitive data attributes with a lack of effective controls over bias?

A third key benefit is the ability to perform dynamic risk management by forecasting the impact on risk when an AI design changes – such as an AI agent's functional capabilities or level of autonomy. This allows for a better understanding of the risk profile of an AI system and helps foster fact-based evaluations of systems against organizational risk metrics.

The EY Trusted AI Platform can be leveraged by organizations to develop this risk quantification during a robust desk-top design and challenge function at the beginning of their AI project. Embedding trust requirements in the design of AI systems from the outset will result in more efficient AI training and higher user trust and adoption.

4. Responding to the risks of AI

Responding to the risks of AI will require the use of new, innovative control practices that can keep pace with AI's fast-paced adaptive learning techniques.

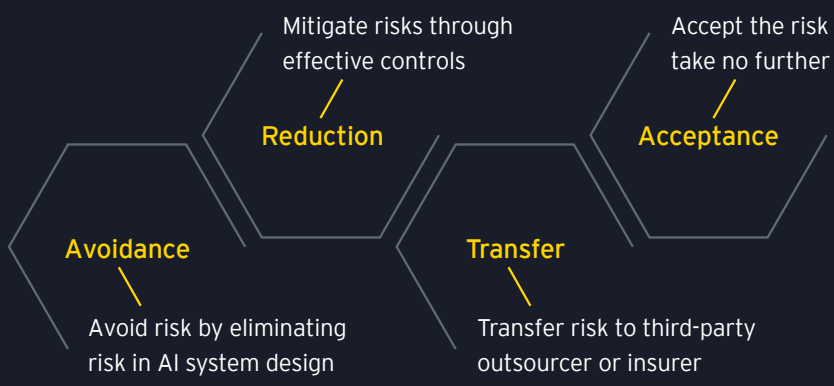
Developing an appropriate AI risk mitigation plan can be difficult as there is no consensus on what good looks like, and stakeholder expectations are still evolving. Traditional risk management methods and tools need to be supplemented with new methods, and yet investments are lagging behind.

In developing a risk mitigation strategy it's important for an organization to use an integrated approach which considers the objectives of the AI system, the potential impacts to stakeholders (both positive and negative), the technical feasibility and maturity of control mechanisms and the risk tolerance of the AI operator.

Consideration should also be given to compliance with existing laws and regulations and voluntary adherence to ethical AI guidelines.

In deciding what to do about each risk there are four strategies which an organization can employ. These strategies can be applied individually or in combination.

Strategies to address the risks of AI



Each of these strategies has its pros and cons and has a role to play in developing a trusted AI system.

The EY Trusted AI Platform includes a mechanism for organizations to select their chosen risk mitigation strategy for each of the identified risks, with the option to select from a pre-determined list of suggested control practices if a risk reduction strategy is selected.

Avoid

In some cases, an organization will want to avoid a risk. This strategy may be chosen when the impacts are serious, or the probability of a negative outcome is so high that you want to eliminate the risk. This involves avoiding the activity altogether such as choosing to not use AI in a particular use case (e.g., diagnosing illness or facial recognition citizen screening). The disadvantage of this strategy is that it may result in the loss of benefits in using AI.

Reduce

A common first approach to risk management is to reduce the risk. This could involve a strategy either to reduce the impact if a negative event occurs or reduce its likelihood. This strategy allows organizations to reap the benefits of AI but reduce its risks to an acceptable level through effective control practices. A possible downside of this approach is an over-confidence in the adequacy of existing control practices or an under-investment in new controls required for AI. This could result in an expectation that the risk is adequately mitigated when it isn't.

Transfer

Another approach may be to transfer the risk to an insurer or third party. Insurance or use of third-party outsourcers may address the remediation and impact costs of AI but do very little in mitigating against reputational damage and perceived accountability. When involving a third party in its AI lifecycle, organizations are still responsible for the risks of their AI systems even if a portion of the AI lifecycle is transferred and must include AI trustworthy and ethical criteria into their third-party risk management framework.

Accept

For minor or low impact risks the acceptance of a risk may be cost-effective. By accepting a risk, you get full access to the AI benefits without a cost outlay to develop and implement control practices. However, there needs to be a serious consideration of the likelihood and cost implications of a failure, as the cost of remediation and reputation could far outweigh the cost of putting a control in place.

5. Monitoring the risks of AI

To monitor AI effectively, continuous monitoring mechanisms are needed to confirm that an AI system is operating as intended.

With AI, which can continue to learn and adapt its decision framework after it's put into production, it's important that strong monitoring mechanisms are in place to establish trust. Organizations need to be able to continually evaluate whether an AI system is operating within acceptable performance levels and identify when a new risk is forming.

Human oversight (e.g., human-in-the-loop) is an important risk mitigation strategy for AI systems, but organizations need to recognize the limitations of human capabilities and leverage automated mechanisms as well.

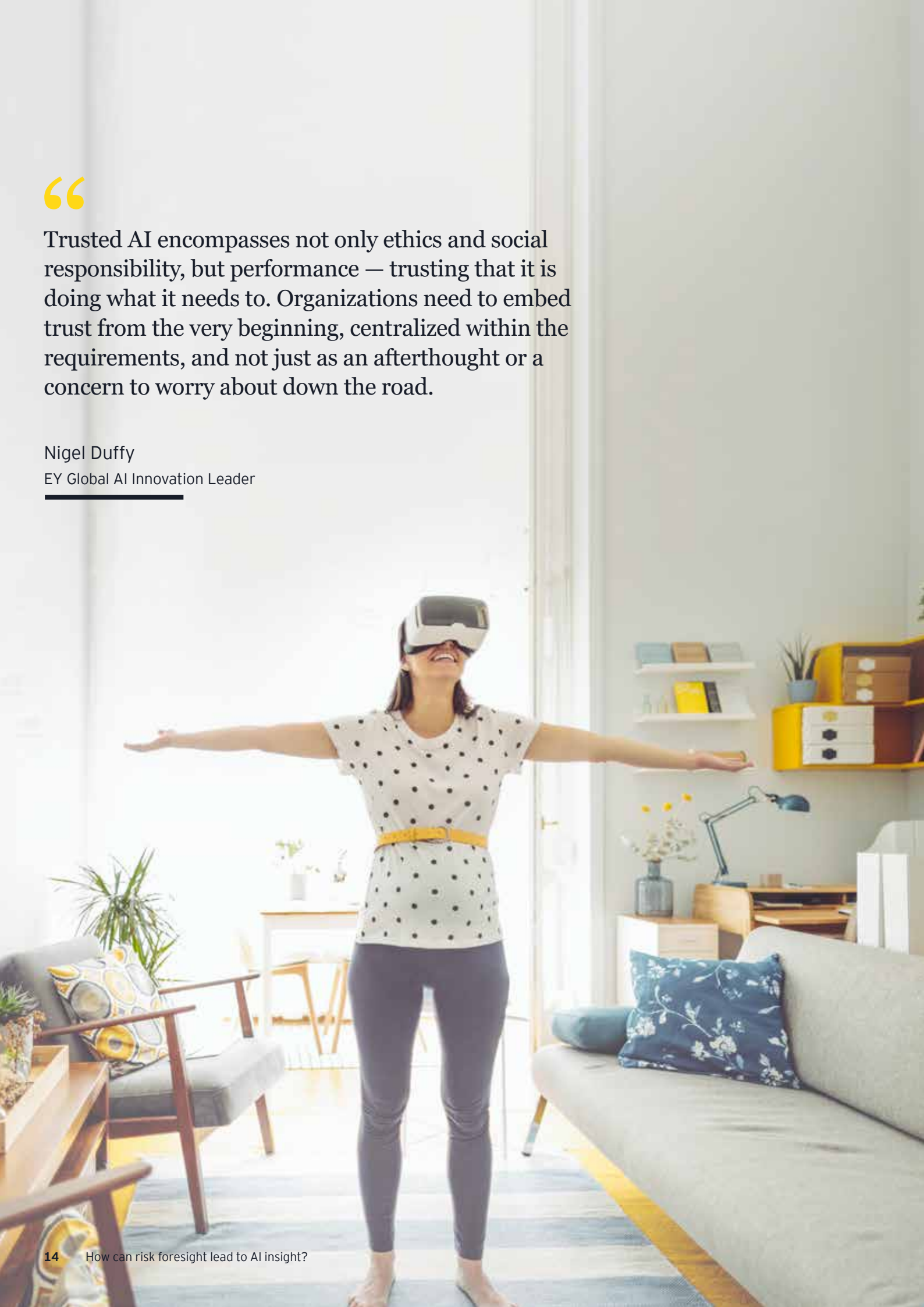
Human operators are best suited to respond to incidents of short duration that require high-level cognitive analysis of a small quantity of disparate information. Humans are not good at maintaining focus over a long time, particularly in low incident situations and have limitations on the quantity of information that can be processed at any one time.

Automated systems, particularly those that are AI-enabled, can overcome these shortcomings. These systems bring their own challenges as they are limited by the information at their disposal and their training.

Leveraging human insight in responding to alert conditions provided through real-time monitoring by automated systems can provide a powerful monitoring mechanism for AI.

An effective monitoring and alert mechanism will also require the identification of acceptable behavior for the AI agent (e.g., what does good look like?) and the codification of these behavior conditions into metric-driven analytics. For example, defining the tolerance for differences in the level of precision and confidence levels for a prediction of credit worthiness across all sub-classes of ethnicity, gender and age.

Lastly, an effective monitoring mechanism for AI will need to be continually evaluated and updated for changes in operating conditions and technological advances. The best approach is to adopt an iterative feedback process in which the results of user feedback, operator issues, AI system validation, real-time monitoring and broader environmental scans affect AI retraining and redesign, and investments in control practices.



“

Trusted AI encompasses not only ethics and social responsibility, but performance — trusting that it is doing what it needs to. Organizations need to embed trust from the very beginning, centralized within the requirements, and not just as an afterthought or a concern to worry about down the road.

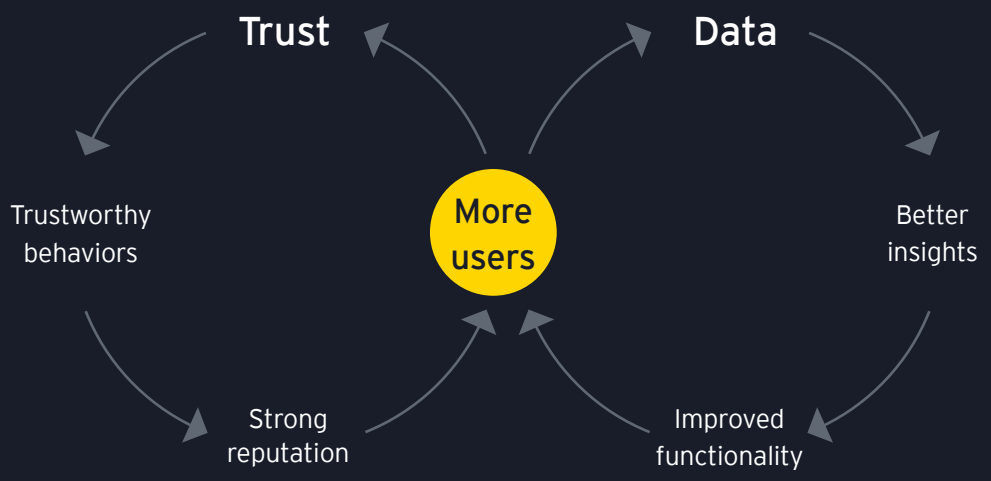
Nigel Duffy

EY Global AI Innovation Leader

6. Leveraging trust in AI as a competitive advantage

Organizations that will thrive in an AI world will be those that leverage trust in AI to attract more users and accelerate their adoption of AI.

AI has already begun to disrupt the way that we work and live. Organizations that will thrive in an AI world will be those that can optimize both data and trust feedback loops to attract more users and accelerate their adoption of AI.



By acting in good faith, developing a robust trustworthy behaviors AI risk management system and involving users in their AI journey, organizations will go a long way in establishing user trust as a competitive differentiator and accelerator of AI adoption.

EY's Trusted AI Platform can assist an organization in this journey by providing insights on the sources and drivers of risk and guiding an AI design team in developing targeted risk mitigation strategies.

The EY Trusted AI Platform provides an organization with the ability to:

- Identify the drivers of AI risk
- Conduct a qualitative and quantitative assessment of AI risks
- Forecast the risk impact when an AI component changes
- Conduct a fact-based evaluation to design risk mitigation strategies
- Plot the risk profile across their AI portfolio
- Develop a robust AI risk management system

The EY Trusted AI Platform helps companies leverage risk foresight to accelerate their access to AI insights. This will enable them to build trust and derive sustained value from AI.

Contacts

Cathy Cobey
EY Global Trusted AI Advisory Leader

Amy Brachio
EY Global Risk and PI Advisory Leader

Nigel Duffy
EY Global AI Innovation Leader

EY | Assurance | Tax | Transactions | Advisory

About EY
EY is a global leader in assurance, tax, transaction and advisory services. The insights and quality services we deliver help build trust and confidence in the capital markets and in economies the world over. We develop outstanding leaders who team to deliver on our promises to all of our stakeholders. In so doing, we play a critical role in building a better working world for our people, for our clients and for our communities.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. For more information about our organization, please visit ey.com.

EYG no. 004392-19Gbl

BMC Agency
GA 1012754

ED None

In line with EY's commitment to minimize its impact on the environment, this document has been printed on paper with a high recycled content.

ey.com