How do you teach AI the value of trust?

The better the question. The better the answer.
The better the world works.

Building a better working world

"AI is different from traditional IT systems and brings with it a new set of opportunities and risks. To build trust in AI organizations will need to go beyond monitoring the AI system's reliability and performance and consider whether it has the appropriate level of explainability and accountability. Consideration should also be given to broader ethical and societal impacts."

Keith Strier
EY Global & EY Americas Advisory AI Leader

Trusted artificial intelligence (AI) explained

AI is not a single technology, but a diverse set of methods and tools continuously evolving in tandem with advances in data science, chip design, cloud services and end-user adoption. The most common examples of AI methods and tools include natural language processing, machine learning, deep learning, computer vision, conversational intelligence and neural networks.

One fundamental difference between AI and non-AI systems is that a traditional program is coded to execute commands, while an AI is coded to learn. In this way, an AI system has the unique ability to improve performance over time (whether through supervised or unsupervised learning), which is a human-like capability.

Although AI is frequently the headline, the real narrative is much broader. EY recommends a systems-view that goes beyond AI, and emphasizes how robotic, intelligent and autonomous systems are the new tools of digital transformation. As a practical matter, enterprises that are further along in their digital journey will be able to more quickly adopt and realize benefits from AI. This would be characterized by enterprise-scale mobile and cloud infrastructure, agile IT processes, comprehensive data integration and governance, and most critically, a culture that encourages experimentation and rewards constructive failures.

AI is being applied toward an ever-wider set of business scenarios, automating activities and tasks traditionally performed by humans. Consequently, it is increasingly important for designers, architects and developers of such systems to be fully aware of downstream and adjacent implications, including social, regulatory and reputational issues. Apart from these, they should also be aware of best practices emerging in ethically aligned design, and governance of intelligent and autonomous systems.

It's well established that robotic, intelligent and autonomous systems can malfunction, be deliberately corrupted, and acquire (and codify) human biases in ways that may or may not be immediately obvious. The first step in minimizing these risks is to promote awareness of them, and then proactively design trust into every facet of the system from day one. This trust should extend to the strategic purpose of the system, the integrity of data collection and management, the governance of model training and the rigor of techniques used to continuously monitor system and algorithmic performance.

AI technologies differ significantly on the opportunities and risks they create, and therefore it's important that organizations consider what type of AI is appropriate for their particular use case. Before starting an AI project organizations should ensure that the following four conditions have been considered and met to the degree required for their specific use case:

Ethics — The AI system needs to comply with ethical and social norms, including corporate values. This includes the human behavior in designing, developing and operating AI, as well as the behavior of AI as a virtual agent. This condition, more than any other, introduces considerations that have historically not been mainstream for traditional technology including moral behavior, respect, fairness, bias and transparency.

Social responsibility — The potential societal impact of the AI system should be carefully considered, including its impact on the financial, physical and mental well-being of humans and our natural environment. For example, potential impacts might include workforce disruption, skills retraining, discrimination and environmental effects.

Accountability and explainability — The AI system should have a clear line of accountability to an individual. Also, the AI operator should be able to explain the AI system's decision framework and how it works. This is about demonstrating a clear grasp of how AI uses and interprets data, how it makes decisions, how it evolves as it learns and the consistency of its decisions across sub-groups.

Reliability — The AI system should be reliable and perform as intended. This involves testing the functionality and decision-framework of the AI system to detect unintended outcomes, system degradation or operational shifts — not just during the initial training or modelling but also throughout its ongoing operation.

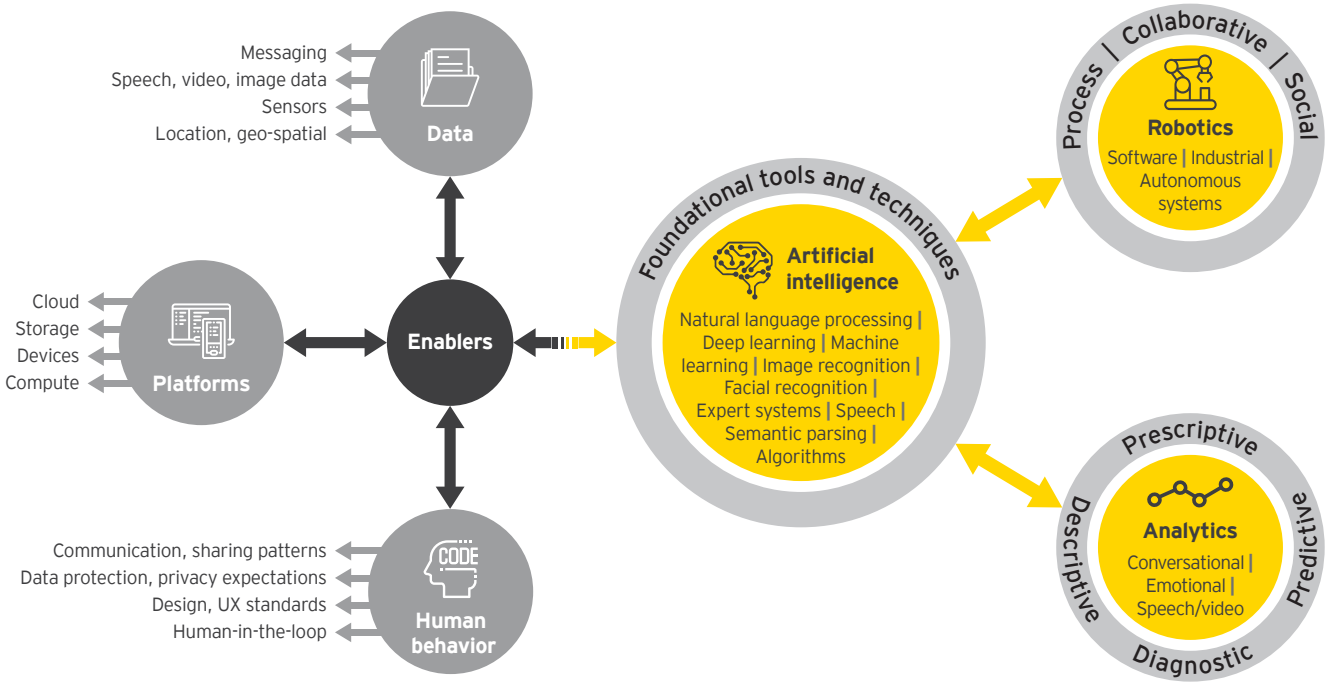
Building trust in AI

EY developed a trusted AI framework to help enterprises understand the slate of new and expanded risks that may undermine trust not only in these systems, but also in products, brands and reputations.

The implications of a failed AI cascade beyond operational challenges. It may also lead to litigation, negative media attention, customer churn, reduced profitability and regulatory scrutiny. Of course, this is more than an academic concern, given recent media and regulatory attention around the potential misuse of personal data to power algorithms that influence, if not shape behavior at the societal level.

Core to EY's framework is a unique emphasis on the systems in which AI is embedded. This systems-oriented view holds that the risks of AI go beyond the underlying mathematics. To achieve and sustain trust in AI, an enterprise must understand, govern, fine-tune and protect all of the components embedded within and around the AI system. These components include data sources, sensors, firmware, software, hardware, user interfaces, networks as well as human operators, and users.

AI system view



Illustratively, consider the complexity of components within an autonomous vehicle that must work together to deliver its intended value. A network of sensors feed data to an onboard AI system that in turn controls multiple mechanical systems. Each of these components plays a critical role in the successful operation of the whole system, and can also represent a single point of failure in the reliability and performance of that system. Therefore, trusting an autonomous vehicle to fulfill its purpose requires that we collectively trust every component of that system in its individual design and performance. Put differently, trust is achieved, sustained or lost at the system level.

Case study: Autonomous driving bus

An autonomous bus nears a stop sign and must decide how quickly to stop. Its intelligent navigation executes a complex set of near-instantaneous decisions and communications between the bus's physical sensors, software and braking mechanism. Multiple components enable the vehicle to perform human-like cognitive functions such as recognizing and identifying the stop sign, interpreting location through GPS positioning, evaluating surrounding objects, and controlling and calibrating the speed of a braking mechanism – all based on its perception of speed, road conditions and distance to the stop sign to balance safety with comfort for passengers.

Each action introduces the potential for failure:

- See and recognize stop sign
- Scan surrounding environment
- Determine distance to STOP sign with GPS positioning
- Assess current speed
- Calculate desired braking pressure
- Apply brakes
- Adjust, as conditions change

"With the increasing impact AI is having on business operations, boards need to understand how AI technologies will impact their organization's business strategy, culture, operating model and sector. They need to consider how their dashboards are changing and how they can evaluate the sufficiency of management's governance over AI, including ethical, societal and functional impacts. They need to take a proactive role in understanding how AI is being used across their business operations and its impact on their risk management and finance functions."

Jeanne Boillet
EY Global Assurance Innovation Leader

Creating trust in AI will require both technical and cultural solutions. To be accepted by users, AI must be understandable, meaning its decision framework can be explained and validated. It must also perform as expected and be incorruptible and secure.

EY's trusted AI framework emphasizes five attributes necessary to sustain trust:

Performance: The AI's outcomes are aligned with stakeholder expectations and perform at a desired level of precision and consistency.

Bias: Inherent biases arising from the development team composition, data and training methods are identified, and addressed through the AI design. The AI system is designed with consideration for the need of all impacted stakeholders and to promote a positive societal impact.

Transparency: When interacting with an AI algorithm, an end user is given appropriate notification and an opportunity to select their level of interaction. User consent is obtained, as required for data captured and used.

Resiliency: The data used by the AI system components and the algorithm itself is secured from unauthorized access, corruption and adversarial attack.

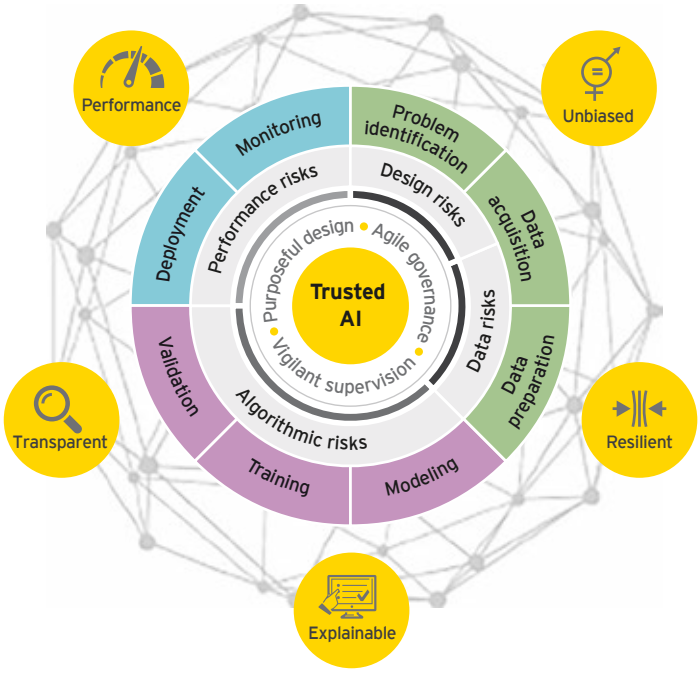
Explainability: The AI's training methods and decisions criteria can be understood, is documented, and is readily available for human operator challenge and validation.

In practical terms, AI is not implemented, but applied, and when it is applied in the following continuous three-step innovation process with these attributes in mind, the outcome is trusted AI:

Purposeful design: Design and build systems that purposefully integrate the right balance of robotic, intelligent and autonomous capabilities to advance well-defined business goals, mindful of context, constraints, readiness and risks.

Agile governance: Track emergent issues across social, regulatory, reputational and ethical domains to inform processes that govern the integrity of a system, its uses, architecture and embedded components, data sourcing and management, model training, and monitoring.

Vigilant supervision: Continuously fine-tune, curate and monitor systems to ensure reliability in performance, identify and remediate bias, and promote transparency and inclusiveness.



"Teaching AI is analogous to parenting a child – you need to teach AI not only how to do a task but also all the social norms and values that determine acceptable behaviour. Training AI in an immersive fashion requires that developers build in ethical and risk considerations at the outset of AI design and development."

Cathy Cobey
EY Global Trusted AI Advisory Leader

Emerging governance practices in establishing trusted AI

Although there is a growing consensus on the need for AI to be ethical and trustworthy, the development of AI functionality is outpacing developers' ability to ensure that it is transparent, unbiased, secure, accurate and auditable. There is a need for organization's to develop an AI governance model that embeds ethical design principles into AI projects and overlays existing technology governance structures.

Leading practices on establishing a trusted AI ecosystem include:

AI ethics board 	A multi-disciplinary advisory board providing independent advice and guidance on ethical considerations in AI development. Advisors should be drawn from ethics, law, philosophy, technology, privacy, regulations and science. The advisory board should report to and/or be governed by the Board of Directors.
AI design standards 	AI design policies and standards for the development of AI, including an AI ethical code of conduct and AI design principles. The AI design standards should define and govern the AI governance and accountability mechanisms to safeguard users, follow social norms and comply with laws and regulations.
AI inventory and impact assessment 	An inventory of all algorithms, including key details of the AI, that is generated using software discovery tools. Each algorithm in the inventory should be subject to an impact assessment to assess the risks involved in its development and use.
Validation tools 	Validation tools and techniques to ensure that the algorithms are performing as intended and are producing accurate, fair and unbiased outcomes. These tools can also be used to monitor changes to the algorithm's decision framework.
Awareness training 	Educating executives and AI developers on the potential legal and ethical considerations for the development of AI, and their responsibility to safeguard an impacted users' rights, freedoms and interests.
Independent audits 	Undergoing independent AI ethical and design audits by a third-party against your AI and technology policies and standards, and international standards to enhance users' trust in your AI system. An independent audit would evaluate the sufficiency and effectiveness of the governance model and controls across the AI lifecycle from problem identification to model training and operation.

Contacts

Rob Walker
EY Risk Leader

+ 44 7917 000 052
rwalker@uk.ey.com

Richard Brown
EY Technology Risk Leader

+44 20 7951 2000
rbrown@uk.ey.com

Piers Clinton-Tarestad
EY Technology Risk, Digital Trust Leader

+44 121 230 1337
pclintontarestad@uk.ey.com

Kevin Duthie
EY Technology Risk, Innovation and Digital Leader

+44 1224 653 120
kduthie@uk.ey.com

Sofia Ihsan
EY Technology Risk, Emerging Technology Assurance Leader

+44 161 333 2762
sofia.ihsan@uk.ey.com

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. For more information about our organization, please visit ey.com.

EYG no. 03880-183Gbl

EY-000075634.indd (UK) 09/18.
Artwork by Creative Services Group London.

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