





1

2

3

4

5

6

About the Deloitte AI Institute

The Deloitte AI Institute helps organizations transform with AI through cutting-edge research and innovation, bringing together the brightest minds in AI to help advance human-machine collaboration in the Age of With™. The Institute was established to advance the conversation and development of AI in order to challenge the status quo. The Deloitte AI Institute collaborates with an ecosystem of industry thought leaders, academic luminaries, startups, research and development groups, entrepreneurs, investors, and innovators. This network, combined with Deloitte's depth of applied AI experience, can help organizations transform with AI. The institute covers a broad spectrum of AI focus areas, with current research on ethics, innovation, global advancements, the future of work, and AI case studies.

Connect

To learn more about the Deloitte AI Institute, please visit www.deloitte.com/us/aiinstitute.

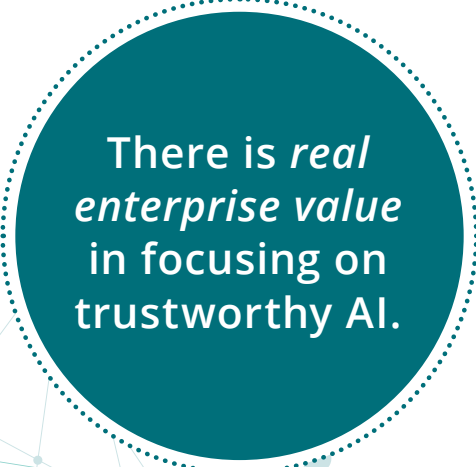


Doing well by doing good with AI

The competitive advantages in trustworthy AI

The full potential of AI hinges on more than function. Organizations need confidence that the tools they deploy behave ethically and are aligned with their values and expectations across a range of dimensions. Yet, like any technology, AI tools have no agency. They are human-designed and driven, and thus the onus is on the enterprise to create AI that is trustworthy and to uphold AI ethics after deployment.

Importantly, this imperative is about more than just mitigating harm. There is real enterprise value in focusing on trustworthy AI. A challenge is to understand the components of ethical AI and bake that insight into enterprise strategy, operational deployment, and stakeholder engagement.



There is *real enterprise value* in focusing on trustworthy AI.



Opportunities inherent in trusted AI systems

There are many examples of cases where AI deployment went wrong: [chatbots that became racist](#), [discriminatory ad algorithms](#), natural language processing with [dubious attention to privacy](#). The consequences from this kind of misbehaving AI (e.g., brand damage, regulatory fines) are easy to understand, which motivates enterprises to take corrective actions.

While a “do no harm” mindset is necessary, if organizations look at AI ethics only through the lens of avoiding consequences, they might miss valuable opportunities that arise because AI is trustworthy. Consider the following benefits:



Compliance

Existing regulatory regimes with an AI nexus (e.g., Europe’s GDPR) and similar forthcoming laws and regulations require attention to AI trustworthiness. Industry groups also develop best practices for AI applications, and individual enterprises have internal requirements for the ethical use of technology. Embedding ethics into the AI life cycle orients the tool toward the ecosystem of regulations and standards, which moves toward efficiency and cost management. Amending or scrapping tools can be expensive.



Reputation

Consumer trust in AI extends to the enterprise using the tool. Trust in AI helps engender consumer confidence and underscores business credibility as a differentiator. What is more, corporate social responsibility (CSR) is a [necessary component of enterprise strategy](#), and ethical AI has a role to play.



Revenue

[Forecasts for AI-derived business value](#) are measured in trillions of dollars, and the bulk of that value is expected to be found in decision support and augmentation. To access this enormous value, enterprises should be able to trust AI outputs and have confidence that the insights they use to make critical business decisions are accurate. In this, trustworthiness is a necessary component of taking part in the AI economy.

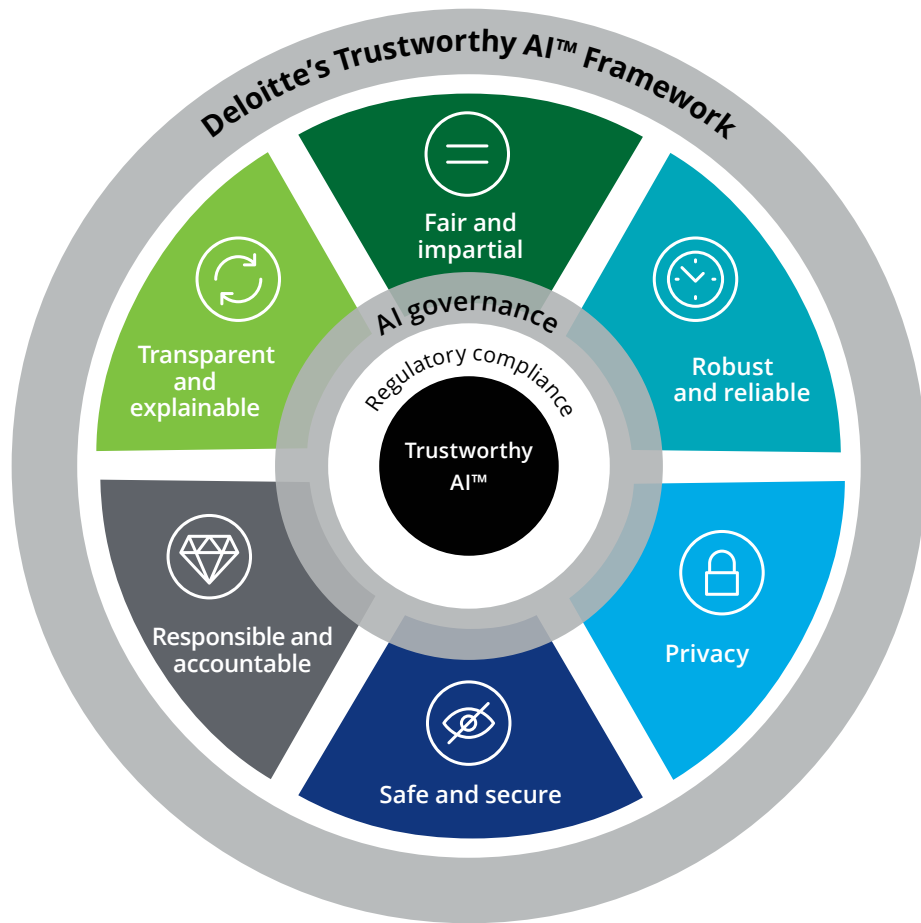


Diversity, equity, and inclusion

There is an enduring public need to further equity in all its forms, and in this, enterprises are both beneficiaries and stakeholders. There is a competitive advantage in using AI tools to reveal nonobvious value in a greater diversity of human capital. Trustworthy AI can help [identify bias that might otherwise go unseen](#), and it can be used to expand recruiting to better applicants, foster meaningful learning, and draw decision-making input from employees whose perspectives matter.

Accessing these and other valuable opportunities with AI requires an enterprise to define just what trustworthy AI means and how it is relevant to the organization.

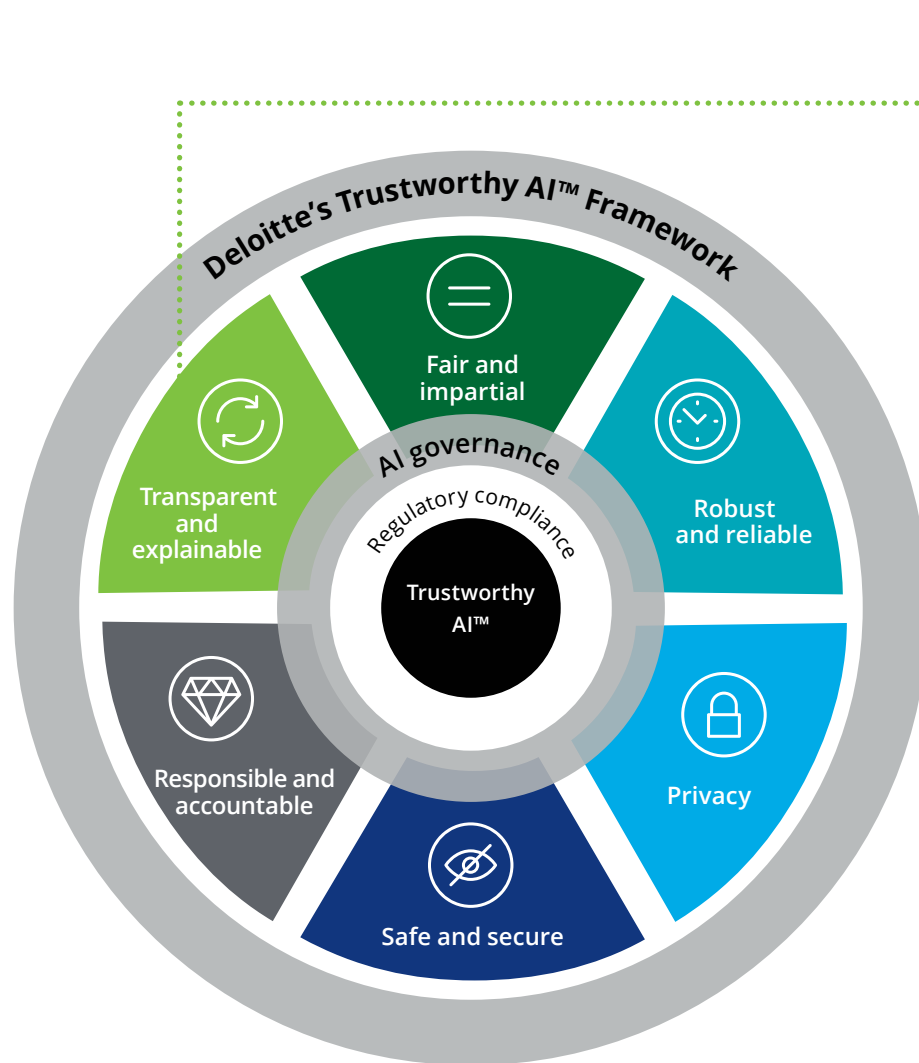
Applying a framework for trustworthy AI



The first step in upholding trustworthy AI is agreeing on the ethical principles that are valuable for the enterprise's strategy and the tools deployed in furtherance of it. What's needed is a framework for consensus-building and decision-making. One example is Deloitte's [Trustworthy AI™ framework](#), which introduces six dimensions through which to evaluate AI behavior: transparency and explainability; fairness and impartiality; robustness and reliability; safety and security; responsibility and accountability; and respect for privacy.

When an organization knows the ethical priorities for assessing its AI tools, it can inform processes, governance, training, and stakeholder buy-in that affects trustworthiness throughout the AI life cycle. Exploring each of the framework's dimensions is a means to more keenly appreciate the inherent value in developing trustworthy AI.

The framework for trustworthy AI



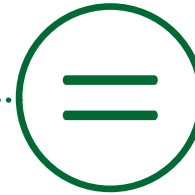
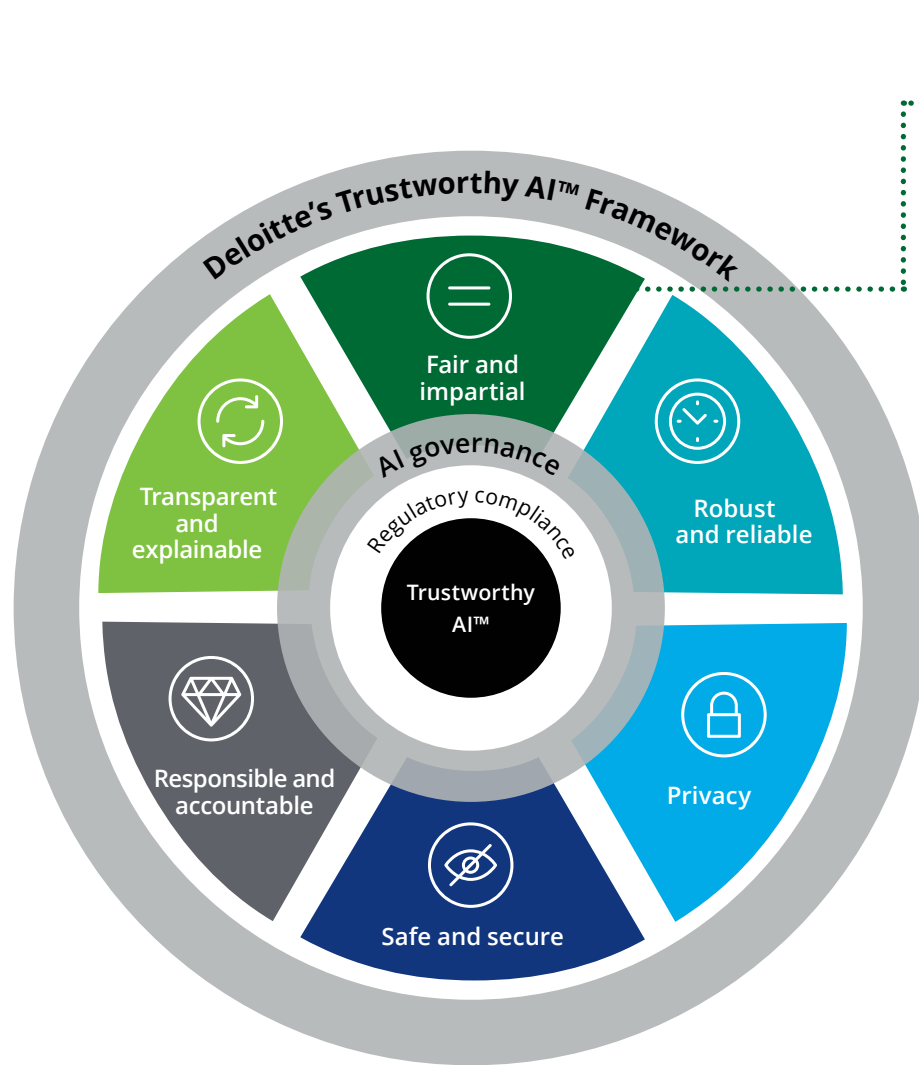
Transparent and explainable

The so-called black box issue with AI frustrates a clear understanding of the precise computational path by which an AI tool arrives at an output. What is needed is a method for interrogating the system, whether that is by building “glass box” solutions or using other tools that allow the algorithms, attributes, and correlations between data to be inspected.

Confidence in AI requires the capacity to understand its flaws, identify biases, perform audits of system accuracy, and, once models are deployed, monitor their drift. This may be accomplished through approaches such as modeling outputs, tracing decisions, or continuous monitoring, each of which has implications for a system's effectiveness.

There are emerging solutions for peering inside the black box. One example is the AI Modeling Insights solution from Chatterbox Labs. The AI-fueled tool is elastic in that it can interrogate both closed and transparent systems, and it presents coherent insights that can be understood by decision-makers who do not have a deep analytic background. In AI, all stakeholders need the capacity to understand a tool's ethical implications from their respective disciplines (e.g., marketing, legal, compliance). For the enterprise, the more decision-makers who can take part in AI application, the more value and opportunity that can be identified in how and where it is used.

The framework for trustworthy AI

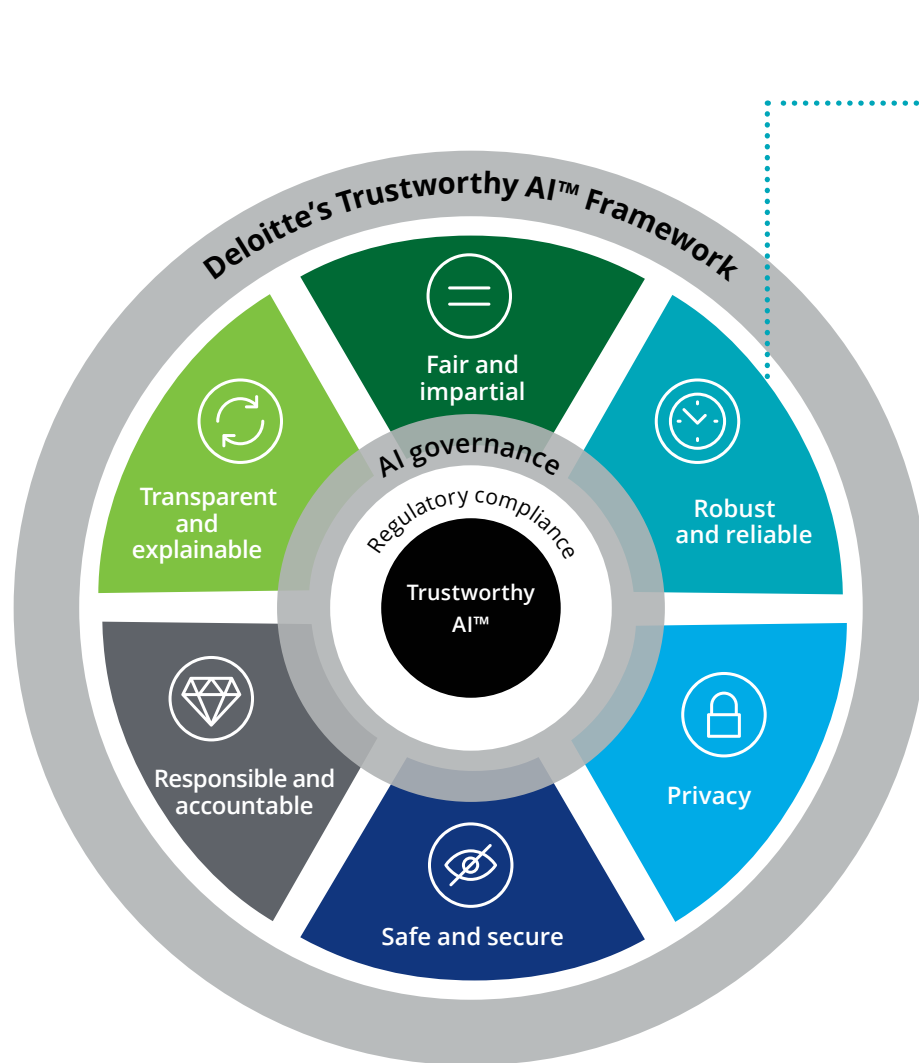


Fair and impartial

The concept of fairness can be somewhat nebulous, and organizations should establish a clear definition in the context of its application. Generally, however, trusted AI should be equitable in its operation. Impartiality rests in many ways on the quality of the data on which a model is trained. Enterprises are charged with ensuring data sets do not contain hidden biases, that real-world domain data is assessed after deployment, and that there are internal and external checks embedded in processes to continuously monitor and validate AI fairness.

As an example in practice, consider the Agmis EasyFlow AI vision system, which can be used to gauge whether personal protective equipment is worn correctly in places where it is required. While vision systems can be susceptible to bias, there is a policy component to EasyFlow deployment. The system informs a manager or other authority of a potential problem, and it is the human who follows up to investigate. This approach leverages the potential in analyzing real-world data to enhance human decision-making, and more than that, the outcome promotes health and safety in the workplace.

The framework for trustworthy AI



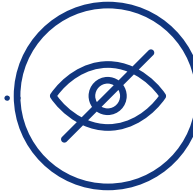
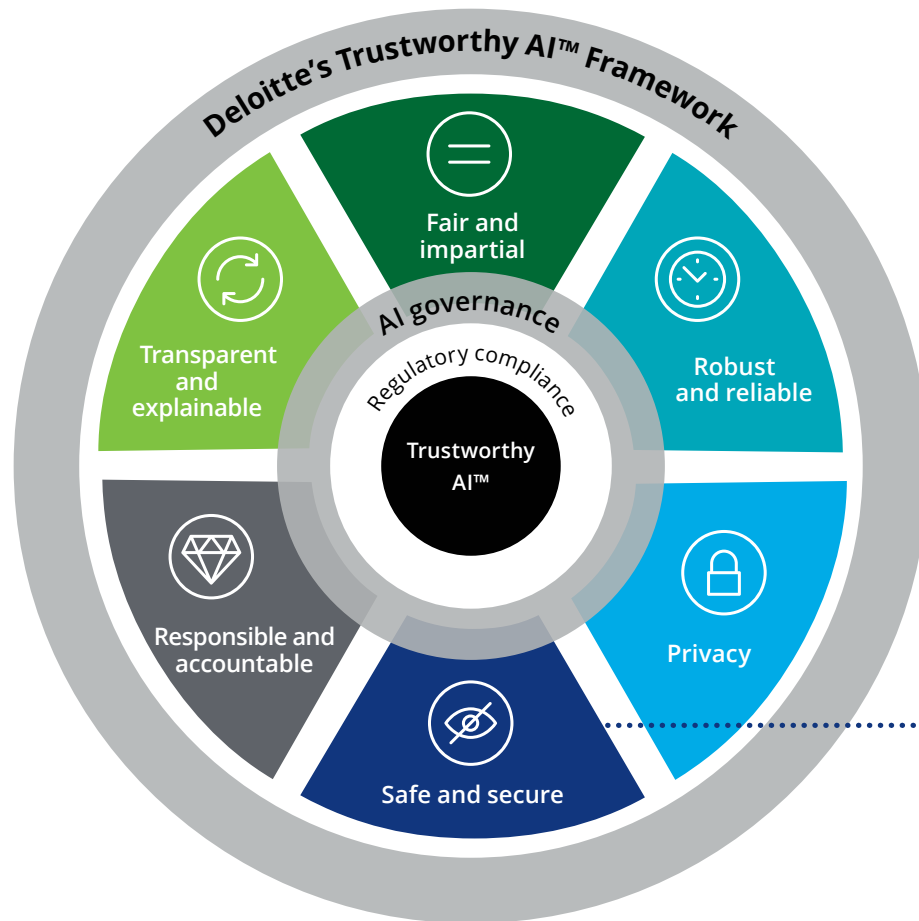
Robust and reliable

Consistency is key. Trustworthy AI systems need to behave reliably and as expected, even when encountering unexpected data. They should also be robust enough to remain reliable at scale. What is more, there is a security component; the system should be robust in a way that shields it from malicious efforts to mislead the model into delivering inaccurate outputs and behaviors.

Model drift is a challenge, and thus a valuable component of robust AI is that it can become more effective and accurate as it is used. The startup Senergy developed a tool to collect and process data from smart meters to identify instances of energy theft. Algorithms identify anomalies in consumption patterns, and importantly, the neural network models continuously receive data from a variety of sources (e.g., historical outages, GIS data, and vegetation conditions) so that the model trends toward greater accuracy with new information from changing weather patterns. The more attention that is directed at AI reliability, the more value an organization can extract from its use.

Kairos, a facial recognition company, partnered with Untangle to audit its deep learning models to detect bias and preserve algorithmic accountability. One auditing method is the "facial relevance map," which shows why the model is making decisions. This is an example how companies are taking accountability to the next level by making their solution explainable.

The framework for trustworthy AI



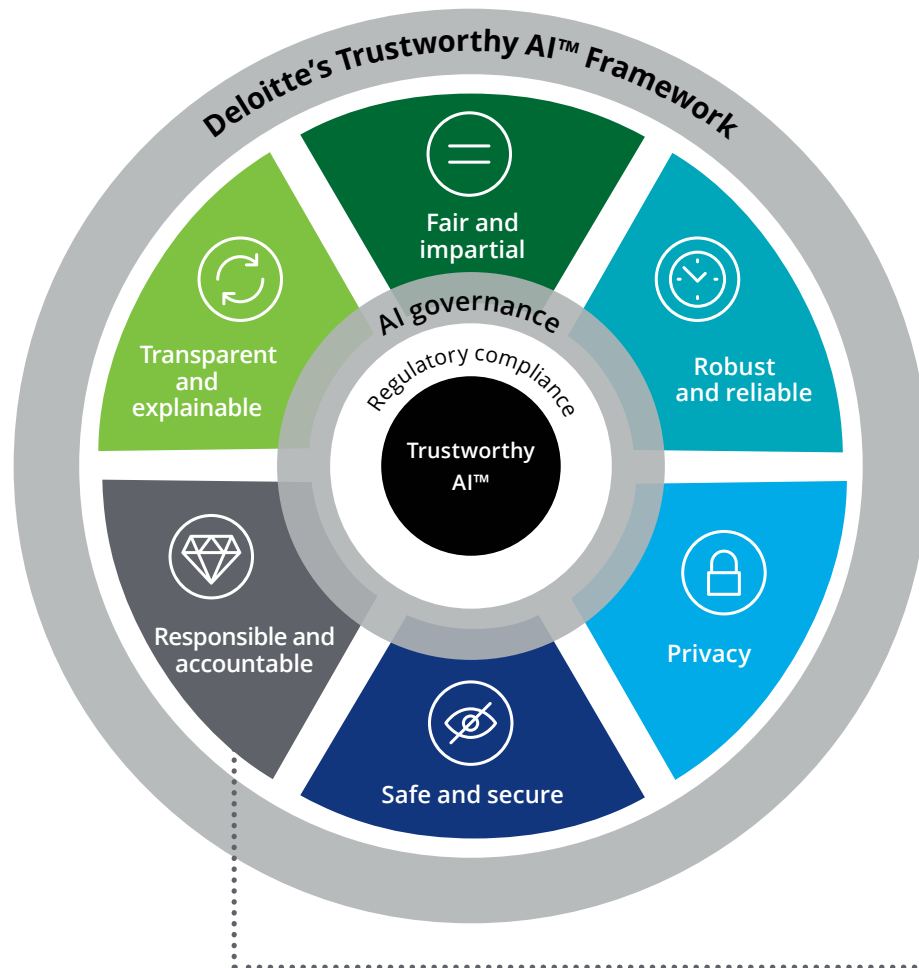
Safe and secure

Risk management is a component of building, deploying, and managing AI. What are the external, internal, and even physical risks surrounding a system's behavior? Organizations should not only build toward systems that are resilient against known (and unknown) risks, but also communicate them to system users and stakeholders. In practice, this may mean adversarial model training, employing layered processes with redundancy built into the infrastructure, and rigorous governance and reporting that permeates the AI life cycle.

Purposeful security considerations in AI can build trust between the organization and its stakeholders. Consider that in technology ecosystems, if one system is compromised, all may become vulnerable. The company Olive offers a workforce model that uses machine learning to automate administrative tasks in health care organizations, and as it relates to security, the model does not require an addition to the existing technology stack. Instead, it employs proven and secure tools the organization already uses.

More fundamental is the security of the data that fuels the AI model. Facial recognition systems raise valid privacy concerns. As an example of intentional data safeguards, Kairos reports using high-level encryption and tokenization to protect its most sensitive data. For organizations that use these kinds of AI tools, the ability to point to rigorous security is a competitive necessity.

The framework for trustworthy AI

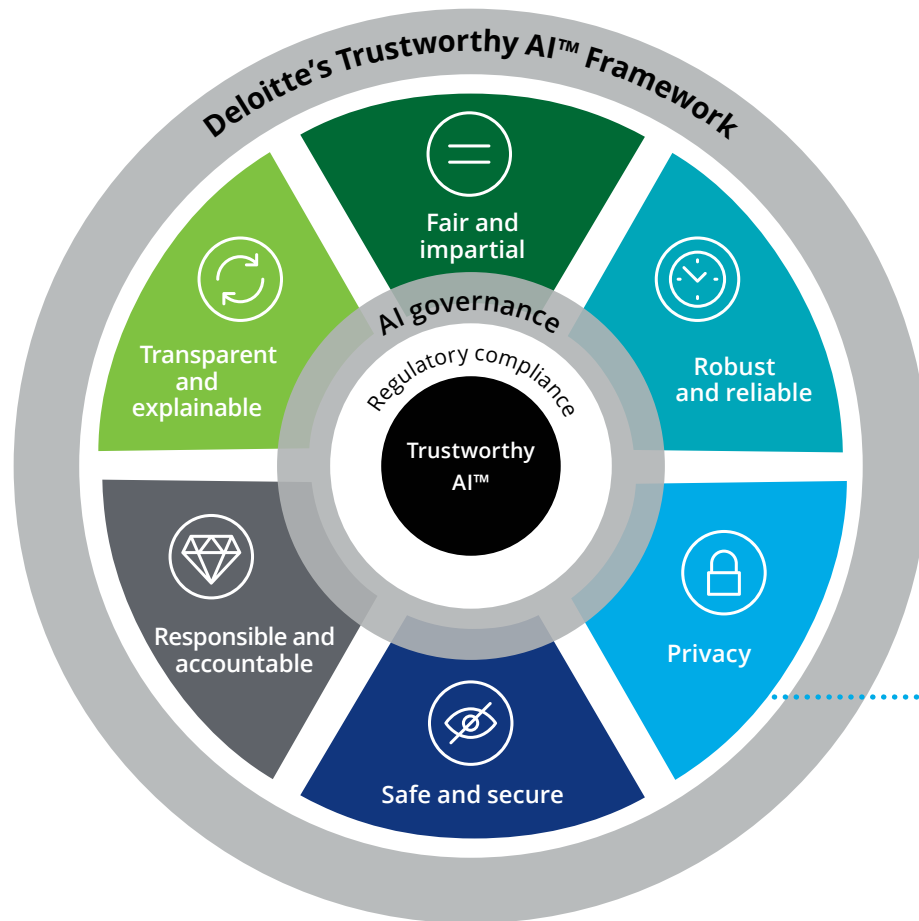


Responsible and accountable

Behind every AI tool is a chain of humans who made decisions on how the system is constructed and used in the real world. Trusted AI is that which is based on human accountability and responsibility for the tool's behavior. This is one of the largest challenges in AI, and addressing it requires a clear chain of responsibility for correcting problems that manifest in operation.

Part of addressing this is training throughout an organization. Stakeholders should be incentivized to think about AI ethics as a component of the workflow and report concerns or insights to the right decision-makers in the enterprise. Toward this end, one approach may be to integrate AI ethics training as an extension of the integrity training many organizations already conduct.

The framework for trustworthy AI



Respectful of privacy

Privacy concerns abound in the technology space, as so much of the data used to fuel innovative technologies is generated by users. For trustworthy AI, organizations should ensure that training data is guarded, inputs are restricted to the model owner, outputs are delivered only to the user, and the model itself is shielded from exposure. In practice, users need clear opportunities to understand what data is being used and in what way, the choice to opt in or out of data-sharing, and a channel for communicating feedback and concerns.

One component of AI that respects privacy is empowering the end user to make informed decisions about whether they want to use the tool. An example that may set a positive trend is Apple's "nutrition label." With it, developers are required to list what data an app will collect, and this is presented in the App Store, allowing users to opt in (or out) before downloading the software. Building these kinds of capabilities into AI models can answer consumer privacy interests, and it also anticipates and aligns with data privacy regulations emerging around the world.

A method for transforming the organization

The world is at the beginning of the Age of With, in which human capacity for insight, decision-making, efficiency, and innovation is dramatically expanded by cognitive systems. We are also entering an age of trust. The tools that unleash world-changing capabilities should be trusted to act in line with human expectations for ethics and appropriateness. The full potential of AI may hinge on that confidence.

Yet, trustworthy AI may be best viewed as a competitive opportunity, rather than as a burden. When an enterprise takes a holistic view of how AI systems operate with regard to the dimensions of trustworthiness, it can change the business. Leadership moves toward accountability, such as with a chief AI ethics officer or governance board. Ethics training and stakeholder buy-in proliferate throughout the enterprise, shifting the business culture toward an ethical mindset. Processes are expanded with focused efforts to continually assess whether AI is remaining within the boundaries of what “trustworthy” means for the organization.

In this, ensuring trustworthy AI is the vehicle for transforming the DNA of the enterprise, putting it on a forward-leaning footing that is ready to thrive and compete in the Age of With. Rather than playing defense against emerging regulatory regimes and unethical AI behavior that engenders customer mistrust, the enterprise sees growing value in confidently leveraging AI capabilities. In that, the business can be uplifted overall, known not just as a user of trustworthy tools, but also as a trustworthy organization.

When an enterprise takes a holistic view of how AI systems operate with regard to the dimensions of trustworthiness, it can change the business.

Author

Arnab Bera

Manager

Data Science & Analytics

Deloitte Consulting LLP



This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional adviser.

Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States, and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.