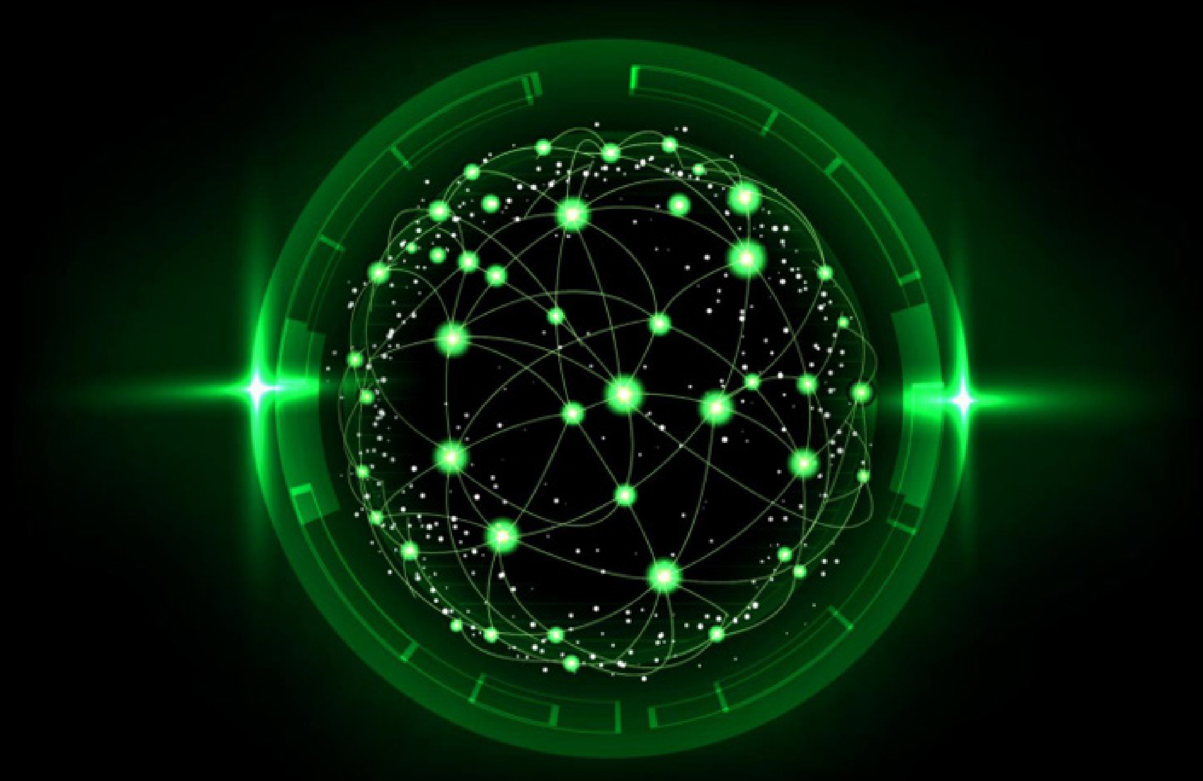
Deloitte.



First risk-tiering based regulation for AI

EU AI ACT

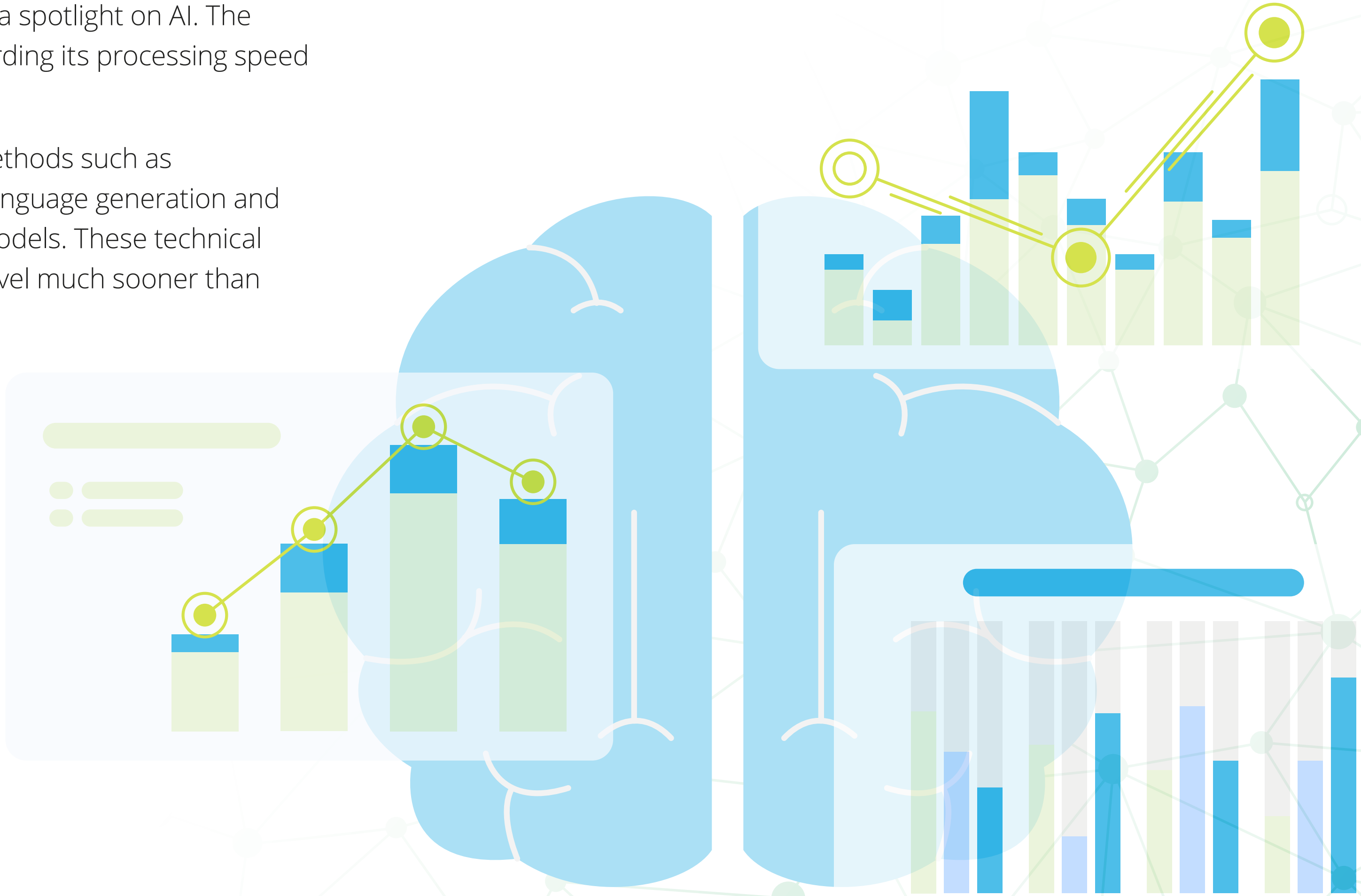
2024

While the idea of AI has been around since the 1950s, recent advances in computer power and the benefits of having a computer that can 'think' have put a spotlight on AI. The exponential improvement in AI has surpassed predictions regarding its processing speed and surprised everyone.

Thanks to new developments like generative AI and learning methods such as reinforcement learning, AI is becoming more skilled in natural language generation and processing, logical reasoning, and creating new patterns and models. These technical capabilities are projected to reach their highest performance level much sooner than initially thought.

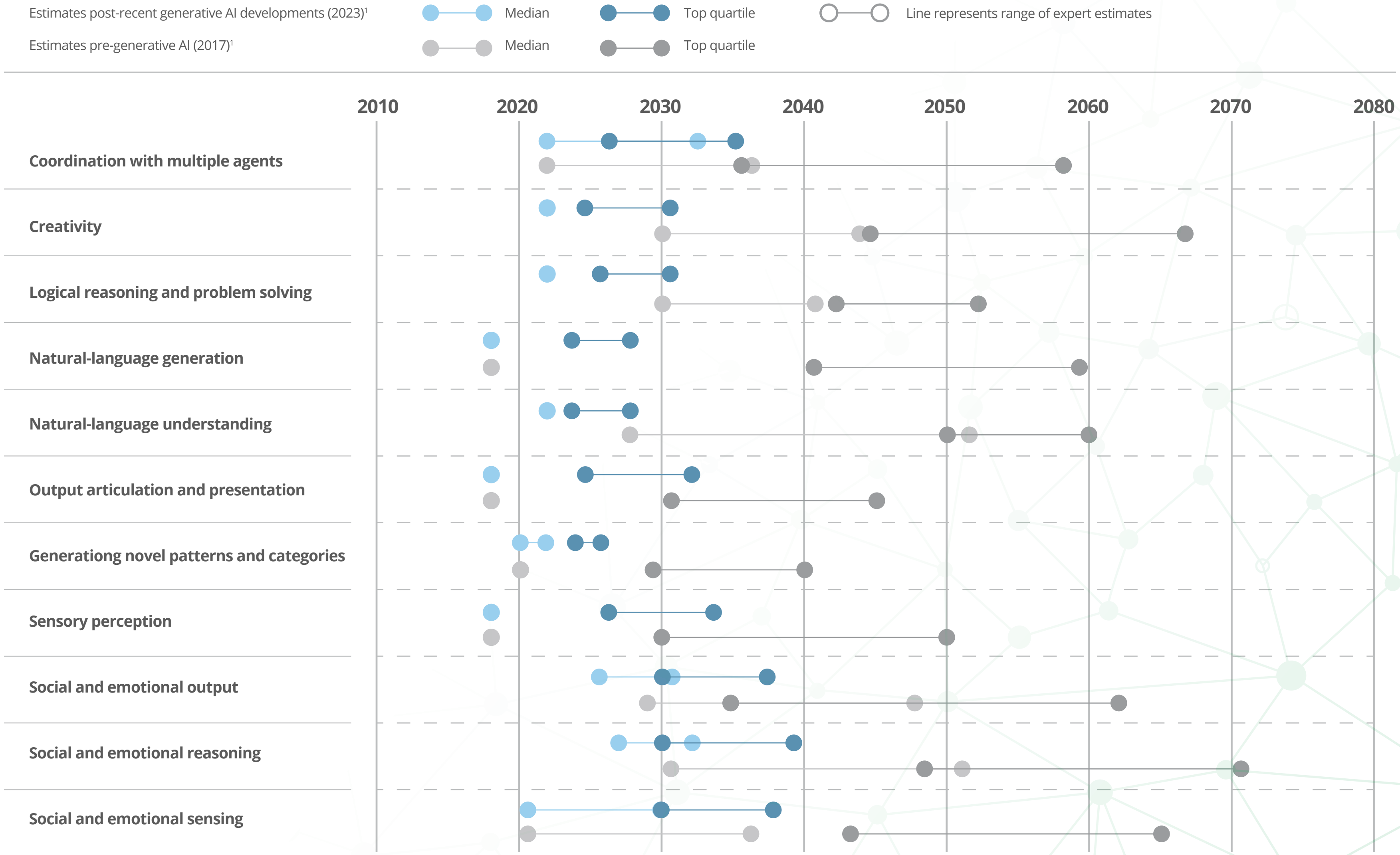
Yet, such a rapid evolution without any regulation has introduced threats to human lives and businesses and introduced technical and social risks. With such an enormous scope, from machine learning algorithms to convolutional neural networks, AI has been a complex topic to cover under an act.

On 9th of December 2023, the EU published its AI legislative act to regulate the paradigm-shifting technology. As it's the first of its kind, the Act puts the EU ahead of any other state or organisation in the race to regulate AI and its potential risks.



Technical capablities, level of human performance achievable by technology

Estimated Projectiles of Technical Capabilities of AI Tools Before and After the Generative AI[1]



[1] Source: McKinsey Global Institute occupation database; McKinsey analysis

The risks and how the act responds to them

The Act focuses on a tiered system to manage risks in high-risk AI systems. It establishes clear rules and mandates a necessary assessment of the impact on fundamental rights for these systems.[2]

1) **Lack of Transparency:** When AI models such as supervised learning models, image processing models and natural language processing models are trained, a dataset is used to train it. Furthermore, many of these models rely on data to process or to feedback (improve its existing model with its output and new data). Such models use complicated mathematical models and computational frameworks in the background that are often unclear or poorly presented to the users.

   With the new Act, new transparency requirements are introduced for high-risk AI systems so that the users can understand the system and use it with ease and clarity. The Act also requires the providers of high-risk AI systems to maintain thorough documentation to demonstrate their compliance with the regulation. This includes records of programming and training methodologies, data sets used, and measures taken for oversight and control.

   The Act empowers users to override the model when necessary to improve outcomes. This clarity enables users to delve into the inner workings of the system, fostering an environment where they can enhance its functionality.

2) **Biased and Discriminating AI:** Some models, such as decision models and affective computing models, risk being biased. This is a significant risk, both in social context and in business. AI is used for numerous decision-making processes affecting our daily and business lives, from qualifying someone for a bank loan to admitting a student to a school, making investment decisions on behalf of the users, to driving a car. Any bias within these processes could pose a severe problem such as discrimination, unfair judgement, loss of assets and intolerance.

   There are different types of AI biases, such as algorithmic prejudice or data bias, negative legacy, temporal bias, over-fitting and implicit bias.[3]

   - Algorithmic or data bias is a type of bias that occurs either because the dataset used to train the model is biased or because the model itself is open to creating a bias.
   - A negative legacy occurs when an AI system uses the existing biases of a human or another model.
   - Temporal bias is a type of bias that happens when a well-working model for a specific period is no longer accurate or valid.
   - Over-fitting is another very common problem of machine learning algorithms which occurs when the model is perfectly aligned with the training dataset but fails to respond correctly to new input data.
   - Finally, implicit bias happens when the AI system unconsciously correlates certain attributes or outputs with specific data types or people groups.

[2] Artificial Intelligence Act: Council and Parliament Strike a Deal On ..., www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/pdf. Accessed 12 Dec. 2023.

[3] "Research Shows AI Is Often Biased. Here's How to Make Algorithms Work for All of Us." World Economic Forum, www.weforum.org/agenda/2021/07/ai-machine-learning-bias-discrimination/. Accessed 12 Dec. 2023.

For this, **the Act dictates that high-risk AI systems must be designed and developed to manage biases effectively, ensuring that they are non-discriminatory and respect fundamental rights.** The AI Act requires human oversight for high-risk systems to minimise risks, ensuring that human discretion is part of the AI system's deployment.

Companies may also need to invest in higher-quality data and advanced bias management tools, potentially increasing operational costs but enhancing AI system fairness and quality.

3) **Privacy and Security:** AI poses a significant risk to privacy. Models require data to train, process, test, validate and feedback. **Some of this data is inevitably private data. This can lead to abuse of personal data, distrust in customers and users for the model or its providers and breach of confidential information.** Either by security breaches or by reidentification (also known as the Mosaic Effect, which occurs when an individual data or dataset stays anonymous on its own becomes identified by being combined with other anonymous or identified data[4]), this risks data privacy and data minimisation.

The Act addresses this risk by stating that training, validation and testing data sets shall be subject to appropriate data governance and management practices. The statement emphasizes that when processing special categories of personal data, it should be done with suitable safeguards to protect the fundamental rights and freedoms of individuals. These safeguards include implementing technical limitations on re-use and utilising state-of-the-art security measures. Privacy-preserving techniques, such as pseudonymisation or encryption, are recommended, especially when anonymisation might significantly impact the intended purpose.

4) **Misinformation and Manipulation:** The emergence of Generative AI has brought forth new avenues for disseminating misinformation and manipulating public opinions. Deepfakes, models to manipulate consumer and social behaviour, and emotion detections are among many AI tools used for this intent. According to a Stanford Study, this is one of AI's biggest threats.[5]

The Act prohibits multiple systems and apps such as decision models, affective computing, models that deploy subliminal techniques, computer vision models used for social scoring, exploiting vulnerabilities, categorising sensitive data, emotion detection and behavioural manipulation.

5) **Techno-Solutionism:** Considering the AI as an omnipotent solution to all problems is another issue and another source of risk. AI is a highly sophisticated, state-of-the-art tool. However, **the tendency to apply AI solutions in every scenario and against every problem causes new problems. Problems vary from discrimination to loss of lives by self-driving car malfunctions.** Moreover, this creates an ambiguity upon the responsible authority of the error. Who do we blame in the case of self-driving car malfunction? The service provider? The user? The hardware provider? Natural causes? Or does anyone else happen to be at the scene of the incident?

[4] "Mosaic Effect Definition." Law Insider, www.lawinsider.com/dictionary/mosaic-effect. Accessed 12 Dec. 2023.

[5] "SQ10. What Are the Most Pressing Dangers of Ai?" *One Hundred Year Study on Artificial Intelligence* (AI100), ai100.stanford.edu/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence-ai100-2021-1-0. Accessed 12 Dec. 2023.

The Act states the necessity of integrating human oversight into high-risk AI systems will require system design and deployment changes, along with potential staff training

The documentation and record-keeping requirements will impose a significant administrative burden, potentially affecting the time to market for new AI products.

6) **Technical Risks:** A sophisticated and rather complicated tool such as AI comes with numerous technical challenges. Those challenges include processing power and hardware provision, storage, optimisation, and choosing the right algorithm, right model, and right sample set. Any failure to meet these challenges can lead to other risks and problems. For instance, choosing a wrong or non-uniformly chosen dataset can lead to overfitting or bias. Or not having enough processing power or insufficient hardware can lead to a failure in model training. So, a reasonable risk assessment, as well as optimisation and design protocol, is essential.[6]

The Act also addresses technical risks. It demands that users and providers consider eliminating or reducing the risks as much as possible by appropriately designing and developing the AI system. If the risk cannot be eliminated, appropriate mitigation and control measures should be put in place.

Provide all information to AI-system users, particularly the risks using the AI as intended and the risks that may occur when the AI is misused and provide training to the users where appropriate. When eliminating or reducing the risks resulting from using the AI system the following should be considered: technical knowledge, experience, education, training, environment in which the system is intended to be used.

As a penalty, the Act specifies that failure to comply may result in significant fines. These fines vary based on the nature of the infringement and the size of the company, ranging from €35 million or 7% of global turnover to €7.5 million or 1.5% of turnover.

[6] Marr, Bernard. "The 15 Biggest Risks of Artificial Intelligence." Forbes, Forbes Magazine, 5 Oct. 2023, www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/.

Key considerations for organisations under the AI Act

Organisations need to consider comprehensive strategies and protocols when contemplating the integration of AI at any operational stage. Such measures not only help organisations to comply to the AI Act, but also allows a smoother and more optimised AI experience.

AI Strategy: Formulating and executing a centralised AI strategy and management framework in response to the evolving regulatory landscape ensures preparedness for upcoming requirements. Investigating such a strategy assists companies in establishing and operating other AI protocols as well.

In addition to these, analysing the completeness and accuracy of your AI inventory and addressing any challenges in the process are also crucial steps to follow.

Such a comprehensive strategy and set of actions help mitigate the technical and operational risks as well as ensure the compliance to the EU AI Act and any other regulations.

Guidance and Training: Organisations should also focus on guidance and training for their personnel, thus creating awareness on the acceptable usage of AI-based solutions, including generative AI within the organisation.

Informing the approved use cases of each potential AI-based solution and updating relevant information security policies are also a part of above-mentioned training and guidance protocol.

This set of actions addresses the risks of manipulation, misinformation, operation, techno-solutionism, and transparency.

AI Risk Framework & Governance: Organisations must formulate controls specific to AI risks, based on regulations and industry standards. Beyond design and implementation, the framework should include monitoring the effectiveness of these controls and implementing governance mechanisms to identify and manage misinformation.

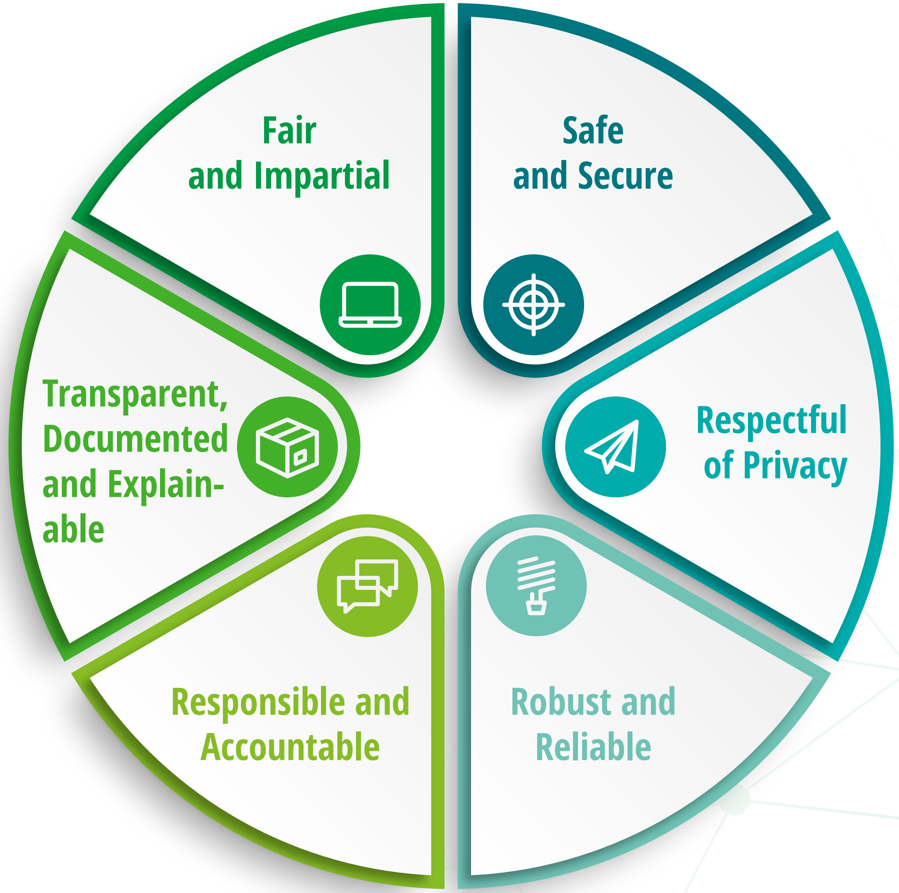
Monitoring: Implementing a comprehensive AI monitoring system, utilising established technologies like Secure Web Gateways and Endpoint DLP Solutions, is crucial. This system should monitor the use of generative AI within the organisation, ensuring the maintenance of all the measures and actions mentioned above.

Deloitte's six dimensions for trustworthy AI

Fair and Impartial
AI systems should make decisions that follow a consistent process and apply rules fairly, as well as incorporate internal and external checks to remove biases that might lead to discriminatory or differential outcomes, to help ensure results that are not merely technically correct but considerate of the social good.

Transparent, Documented and Explainable
AI systems may not operate as "black boxes"; all parties engaging with an AI should be informed that they are doing so and be able to inquire as to how and why the system is making decisions.

Responsible and Accountable
The increasing complexity and autonomy of AI systems may obscure the ultimate responsibility and accountability of companies and human beings behind the decisions and actions of these systems; policies should be in place to clearly assign liability and help ensure that parties impacted by AI can seek appropriate recourse.



Safe and Secure
Just as we currently depend on the consistent performance of human professionals to help ensure that our daily activities are safe and healthy, we should be able to depend on equivalent or even greater reliability as we enable more of our systems with AI.

Respectful of Privacy
As AI systems often rely on gathering large amounts of data to effectively accomplish their tasks, we should ensure that all data is gathered appropriately and with full awareness and consent, and then securely deleted or otherwise protected from further, unanticipated use.

Robust and Reliable
As AI systems take greater control over more critical processes, the danger of cyberattacks and other harms expands significantly. Appropriate security measures should be put in place to help ensure the integrity and safety of the data and algorithms that drive AI.

How can we help

Our Trustworthy AI framework helps us to address our AI related services with diligence and, helps our clients to understand, respond to and mitigate AI risks in a responsible, sustainable manner.

Addressing each of the risks and considerations discussed earlier, Deloitte provides professional, industry-level services designed to moderate and resolve potential challenges and problems.

Donal Murray
Partner, Risk Advisory
donmurray@deloitte.ie
+353 1 417 8587

Colm McDonnell
Partner, Risk Advisory
cmcdonnell@deloitte.ie
+353 1 417 2348

Hilary Lemass
Director, Risk Advisory
hlemass@deloitte.ie
+353 1 574 9938

Onatkut Varis
Director, Risk Advisory
ovaris@deloitte.ie
+353 1 417 3295

Deloitte.

Important notice

At Deloitte, we make an impact that matters for our clients, our people, our profession, and in the wider society by delivering the solutions and insights they need to address their most complex business challenges. As the largest global professional services and consulting network, with over 312,000 professionals in more than 150 countries, we bring worldclass capabilities and high-quality services to our clients. In Ireland, Deloitte has over 3,000 people providing audit, tax, consulting, and corporate finance services to public and private clients spanning multiple industries. Our people have the leadership capabilities, experience and insight to collaborate with clients so they can move forward with confidence.

This publication has been written in general terms and we recommend that you obtain professional advice before acting or refraining from action on any of the contents of this publication. Deloitte Ireland LLP accepts no liability for any loss occasioned to any person acting or refraining from action as a result of any material in this publication.
Deloitte Ireland LLP is a limited liability partnership registered in Northern Ireland with registered number NC1499 and its registered office at 19 Bedford Street, Belfast BT2 7EJ, Northern Ireland.

Deloitte Ireland LLP is the Ireland affiliate of Deloitte NSE LLP, a member firm of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee (“DTTL”). DTTL and each of its member firms are legally separate and independent entities. DTTL and Deloitte NSE LLP do not provide services to clients. Please see www.deloitte.com/about to learn more about our global network of member firms.

Designed by CoRe Creative Services. RITM1655967