Deloitte.

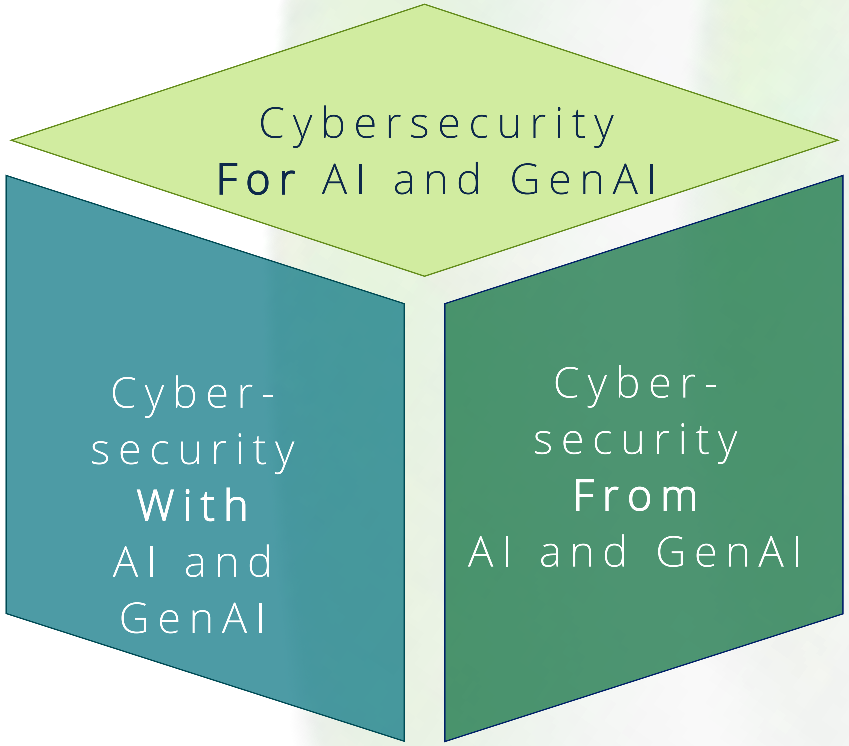
Cybersecurity meets AI and GenAI

October 2024

Table of contents

Dimensions of Cybersecurity within AI and GenAI

Cybersecurity considerations with regards to Artificial Intelligence (AI) and Generative AI (GenAI) have to be viewed from three different angles: Securing AI and GenAI Systems, using AI and GenAI for improving Cybersecurity, and using AI and GenAI for malicious actions.



Cybersecurity **For** AI and GenAI

Cybersecurity **With** AI and GenAI

Cybersecurity **From** AI and GenAI

Cybersecurity *For* AI and GenAI

Protecting AI and GenAI systems from Cybersecurity threats, by providing guidance to secure implemented or planned AI and GenAI use-cases.

Focus of the following

- Cybersecurity for AI and GenAI Framework
- Trusted & Secure AI

Cybersecurity *With* AI and GenAI

Improving Cybersecurity capabilities and boosting Cybersecurity processes by including AI and GenAI.

- Use Case Ideation & Development
- AI and GenAI Training & Labs

Cybersecurity *From* AI and GenAI

Changing Cybersecurity threat landscape due to launch of more sophisticated and new kinds of cyberattacks.

- "Futurecasting" AI and GenAI Tabletop Exercises
- AI Threat Intel & Attack Surface Management

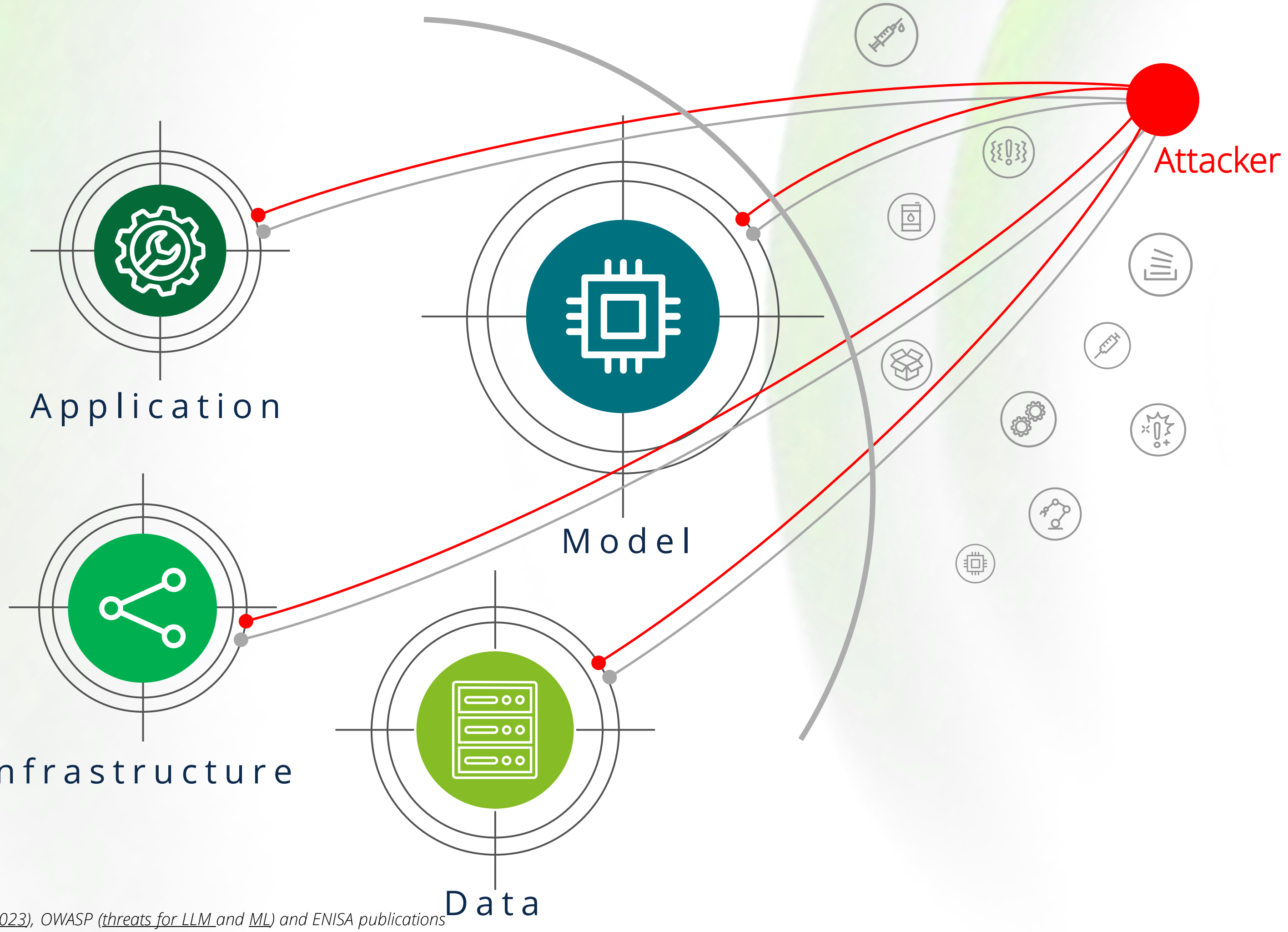
AI and GenAI induced change in Cybersecurity threat landscape*

The rise of AI and GenAI not only comes with new opportunities but also with a change in security-related threats that will continue to evolve, making it imperative to secure AI systems.

Cybersecurity Threat Landscape

The increasing usage and availability of AI and GenAI leads to a change in the Cybersecurity threat landscape. On the one hand it enables attacker to intensify their attack frequency, efficiency and complexity due to the use of AI and GenAI, on the other hand it is leading to completely new threats for AI and GenAI like adversarial attacks.

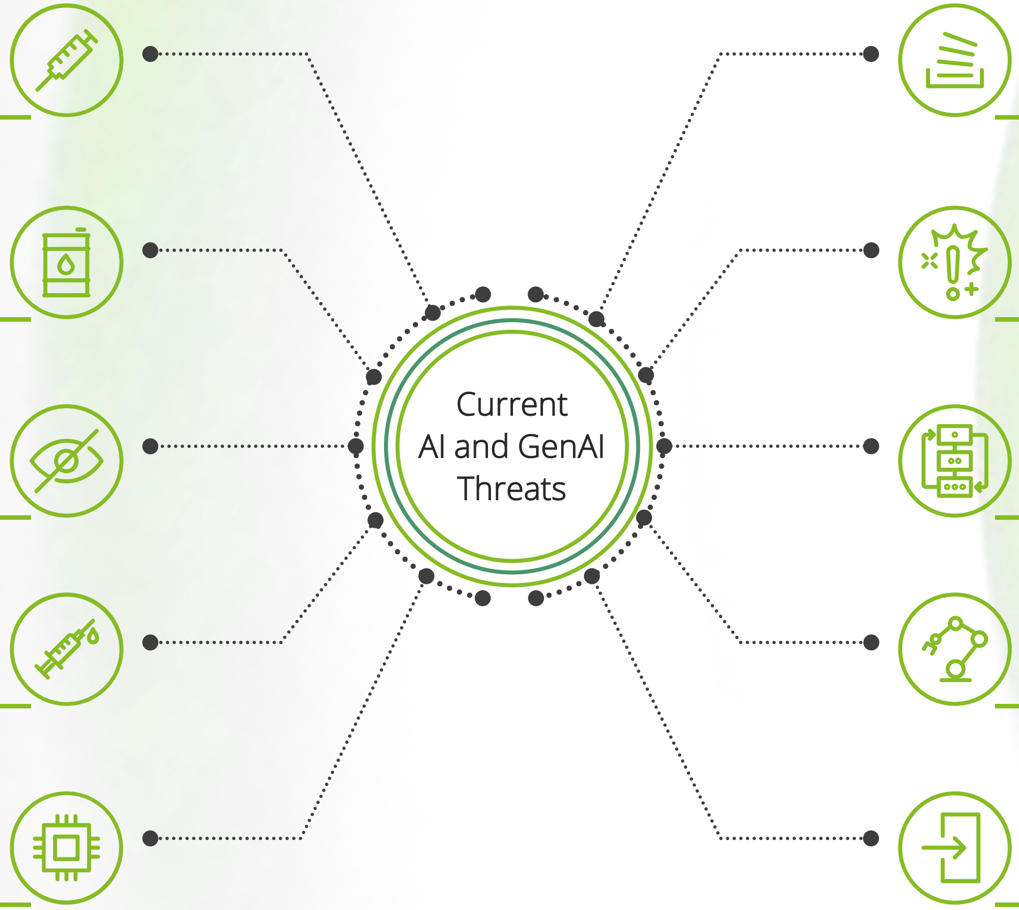
Moreover, the attack surface presented by AI and GenAI solutions is unfamiliar territory for many. It's not only the infrastructure, data and application that require safeguarding, but also the underlying model on which any AI and GenAI System is build. It contains many sensitive information and requires additional protection. Additionally, the increasing amount of data being processed and stored is leading to an increasing focus of data security.



*based on ISO standard (ISO/SAE42001:2023), OWASP (threats for LLM and ML) and ENISA publications

AI and GenAI is expanding the Cybersecurity threat landscape*

Based on the publications of OWASP, ISO and ENISA, Deloitte consolidated the Top 10 threats for AI and GenAI.



Input Injection
Compromising AI applications with malicious inputs that override controls or alter model behavior e.g. Prompt injection for Large Language Models (LLMs).

Training data poisoning
Introducing vulnerabilities or biases into AI models by tampering their training data, compromising security, effectiveness, or ethical behavior.

Information breach
Unauthorized exposure of private data and/or metadata leading to unwarranted data access, privacy (GDPR) violations, and security breaches.

Model poisoning
directly manipulating AI model's parameters to influence its behavior negatively.

Model stealing
Unauthorized access, copying, or exfiltration of an AI model.

Model inversion
Utilizing an AI model's output to reconstruct sensitive data samples used for training, effectively reverse-engineering the model to extract information.

Model denial of service
Triggering resource-intensive operations through inputs that lead to AI model disruptions.

Supply chain vulnerabilities
Incorporating compromised or insecure third-party components like third-party datasets, pre-trained models, and plugins leading to security risks.

Excessive agency abuse
Causing AI applications to gain excessive authority or using such authority to perform unintended actions beyond desired functionality.

Adversarial examples
Utilizing adversarial learning to create malicious inputs, which deceive AI models during the inference phase, e.g. by causing a misclassification.

**based on ISO standard (ISO/SAE42001:2023), OWASP (threats for LLM and ML) and ENISA publications*

Cybersecurity for AI and GenAI Framework*

Overarching & AI and GenAI domain specific security capabilities ensure the secure development, implementation, and usage of AI and GenAI solutions.

AI and GenAI Domains

The domains constitute the core structure of AI and GenAI systems and are used to cluster security capabilities

The Data Domain includes all data handled by the model during training, testing, validation, and for inference after deployment.

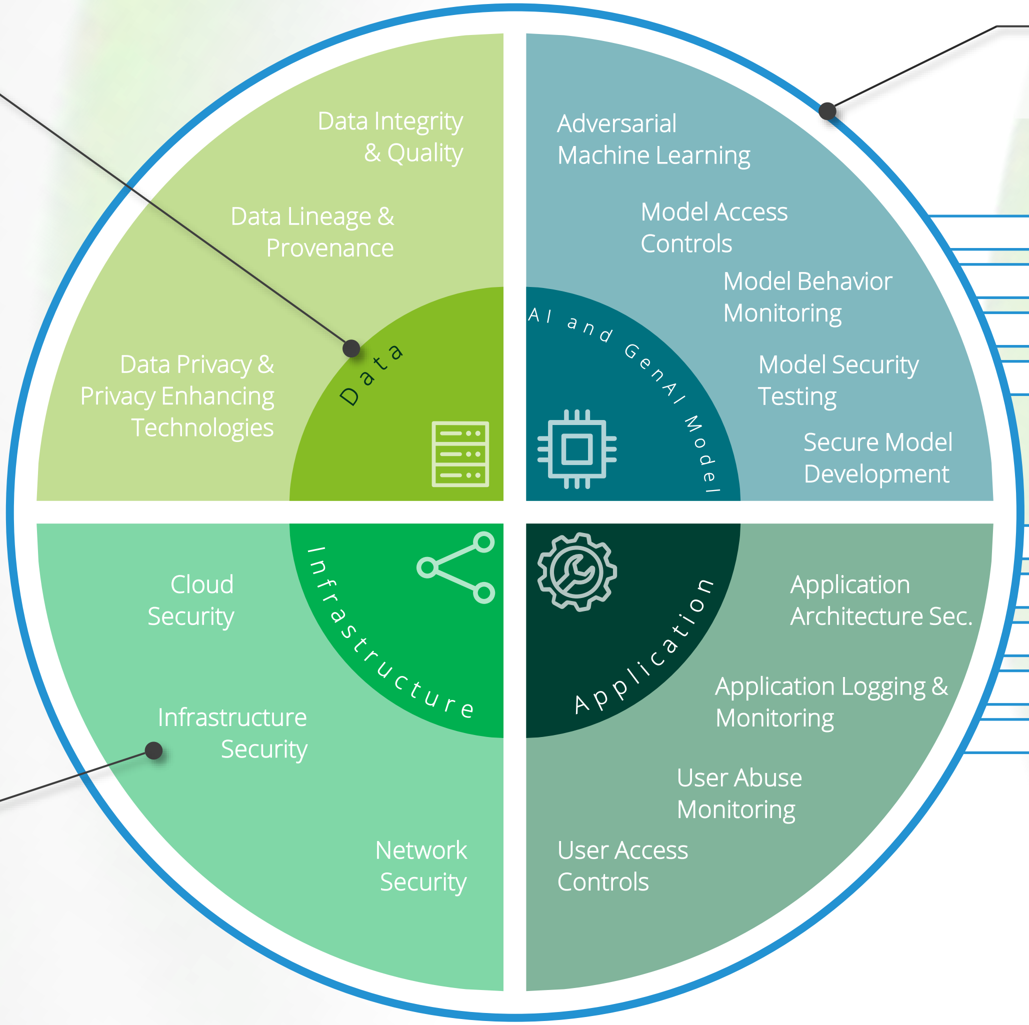
The AI and GenAI Model Domain involves the model architecture, training, testing and validation processes, in addition to the model's unique parameters.

The Application Domain is the external layer of the AI and GenAI system that hosts the model and sits on the infrastructure. It serves as the user interface.

The Infrastructure Domain encompasses the underlying hardware and networking components that are used for developing and hosting the AI and GenAI system.

AI and GenAI Security Capabilities

A Security Capability is a category for grouping of controls that are designed to help address specific Cybersecurity threats in each domain.



Data

Data Integrity & Quality

Data Lineage & Provenance

Data Privacy & Privacy Enhancing Technologies

AI and GenAI Model

Adversarial Machine Learning

Model Access Controls

Model Behavior Monitoring

Model Security Testing

Secure Model Development

Infrastructure

Cloud Security

Infrastructure Security

Network Security

Application

Application Architecture Sec.

Application Logging & Monitoring

User Abuse Monitoring

User Access Controls

Overarching Security Capabilities

Capabilities to maintain the security of AI and GenAI solutions across each domains.

Lifecycle Security

- Secure development process
- Secure supply chain
- Asset management
- Data loss prevention (DLP)

Governance, Risk & Compliance

- Third-party risk management
- AI and GenAI security risk management
- AI and GenAI specific policies, standards & architecture
- Business continuity management
- Regulatory compliance

*based on ISO standard (ISO/SAE42001:2023), OWASP (threats for LLM and ML) and ENISA publications

AI and GenAI Cybersecurity roadmap – discover your next steps

The AI and GenAI Cybersecurity Roadmap is designed to help organizations on the journey toward secure implementation, deployment, and usage of AI and GenAI applications.

01 Hold a AI and GenAI Cybersecurity lab

Understand the basics: delve into foundational concepts of Cybersecurity for AI and GenAI including threat landscape, encryption, network security, and access controls with AI labs and future casting table-top exercises (TTX)

Familiarize yourself with AI and GenAI: gain a basic understanding of AI and GenAI principles and its implementation, algorithms, and its applications in Cybersecurity.

02 Assess your AI risk level (AIRL)

Measure your maturity/risk level: to gauge your organization's readiness and maturity, we have Framework devised a broad assessment and security. Deloitte's solution can help you define the AIRL-of each component you are hoping to secure scoring it on a 1 to 5 scale. Your components' AIRL will inform the specific controls families to be considered and prioritized.

03 Identify tailored Security controls

Identify the tailored set of Cybersecurity controls. Deloitte's Cybersecurity for AI and GenAI Framework has over 500 controls from legislation, industry standards, and existing frameworks mapped to four domains to meet the needs of your AIRL.

04 Implement tailored Security controls

Begin with individual quick wins: Deloitte's approach begins with implementing individual quick wins derived from our maturity assessment to focus on near-term security enhancements. These quick wins are actionable steps that are designed to help you yield significant improvements in your security posture.

05 Engage in continuous monitoring

Actively monitor your AI and GenAI systems. Though implementing tailored security controls is crucial to harden your AI and GenAI systems against threats, the ever-evolving threat landscape and new legislation/regulations are pushing organizations to continuously monitor their environments to stay abreast of attacks. To bolster your confidence in the security of your AI systems, Deloitte's attack surface monitoring (ASM) and threat intel, and AI red teaming services are next steps to help you bolster the Cybersecurity for and from AI systems and threat actors.

Cybersecurity for AI and GenAI Framework legend

The Deloitte Cybersecurity for AI and GenAI Framework synthesizes best practices from ISO standards, MITRE ATT&CK, OWASP and ENISA to secure your AI and GenAI system to combat known and novel threats.

External drivers

Threats

As a special type of software, there are new and traditional Cybersecurity threats to AI and GenAI systems.

(Gen)AI is expanding the cybersecurity threat landscape

Based on the publications of OWASP, ISO and ENISA, Deloitte consolidated the Top 10 threats for (Gen)AI

Regulations

We have conducted geography-specific research to comply with regulations and follow standards.

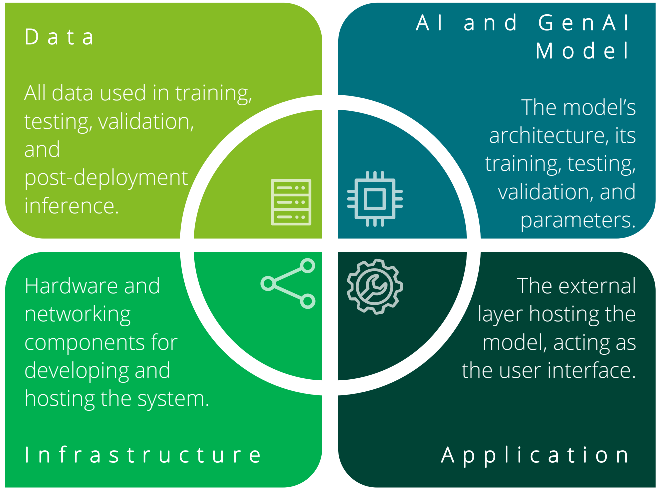
This is a major focal point of this Framework and is being continuously updated to keep up with this evolving space.

Cybersecurity for AI and GenAI Framework structure

System domains

AI and GenAI threats identified can be exploited in different stages of the AI and GenAI lifecycle.

Building appropriate mitigation mechanisms against the new and traditional vulnerabilities, the AI and GenAI system is organized in four different domains.



Data

All data used in training, testing, validation, and post-deployment inference.

AI and GenAI Model

The model's architecture, its training, testing, validation, and parameters.

Hardware and networking components for developing and hosting the system.

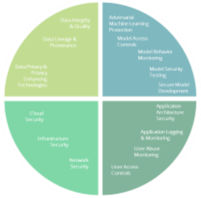
Infrastructure

The external layer hosting the model, acting as the user interface.

Application

Domain-specific capabilities

For each domain, there are different capabilities which group security controls to address all the potential attack vectors and threats for AI and GenAI.



Data Integrity & Quality
Data Lineage & Provenance
Data Privacy & Privacy Enhancing Technologies
Adversarial Machine Learning Protection
Model Access Controls
Model Behavior Monitoring
Model Security Testing
Secure Model Development
Cloud Security
Infrastructure Security
Network Security
Application Architecture Security
Application Logging & Monitoring
User Abuse Monitoring
User Access Controls

Overarching capabilities

As AI and GenAI is a software at its core, overarching capabilities ensures Governance, Risk, and Compliance for the AI and GenAI Lifecycle across domains.

Lifecycle Security	Governance, Risk & Compliance
Secure development process	Third-party risk management
Secure supply chain	AI and GenAI security risk management
Asset management	AI and GenAI specific policies, standards & architecture
Data loss prevention (DLP)	Business continuity management
	Regulatory compliance

Controls

For each of the capabilities, security controls are summarized to mitigate all potential threats.

Extract of controls

- Robust access control
- Regular validation of SBOM
- Watermarking
- Input injection testing
- Data Minimization
- Data Separation

Stay tuned for our next publication!

Deloitte.

Thank you.

Volker Burgers
Partner
vburgers@deloitte.de

David Caswell
Managing DIrector
dcaswell@deloitte.com

Jordan McKenzie
Manager
jormckenzie@deloitte.de

Lucie Wollenhaupt
Manager
lwollenhaupt@deloitte.de

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited (DTTL), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.

Deloitte provides industry-leading audit and assurance, tax and legal, consulting, financial advisory, and risk advisory services to nearly 90% of the Fortune Global 500® and thousands of private companies. Our people deliver measurable and lasting results that help reinforce public trust in capital markets, enable clients to transform and thrive, and lead the way toward a stronger economy, a more equitable society, and a sustainable world. Building on its 175-plus year history, Deloitte spans more than 150 countries and territories. Learn how Deloitte's approximately 457,000 people worldwide make an impact that matters at www.deloitte.com.

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited (DTTL), its global network of member firms or their related entities (collectively, the "Deloitte organization") is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related identities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. DTTL and each of its member firms, and their related entities, are legally separate and independent entities.