Unleashing the power of GenAI through managing cyber risks

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

EY

Shape the future with confidence

GenAI has quickly gained massive popularity and is evolving rapidly

ChatGPT reached 100 million users within two months of launch. Organizations are wrestling with the seemingly urgent need to implement generative AI (GenAI). However, for GenAI to reach its full potential the underlying cyber and related risks need to be recognized and addressed.

Source: ChatGPT reaches 100 million users two months after launch | Chatbots | The Guardian

What are the general risks associated with GenAI?

AI hallucination and unpredictability

GenAI can sometimes produce unpredictable or factually incorrect responses presented in a very convincing manner. To mitigate risks from using incorrect information, companies should provide context and leverage prompt engineering to generate context-aware-answers – and keep humans in the loop for AI oversight.

Discrimination and bias

As GenAI leverages publicly available data for training, responses tend to reflect sentiment rather than objective facts. Bias can be mitigated through techniques such as reinforcement learning with human feedback (RLHF) where the model is taught to be unbiased; yet this method will never be foolproof. Humans must remain in the loop, as part of AI governance.

Copyright infringement

GenAI is trained on publicly available data, much of which is copyright protected. This can lead to lawsuits by intellectual property (IP) holders. Mitigation strategies rely heavily on foundation model providers to obey copyright laws, and for legal frameworks to catch up with GenAI.

Confidentiality and data privacy

When training GenAI models in the public cloud, companies transmit proprietary data which may then be used to further train the model and may be relayed to other users, leading to confidentiality and data privacy concerns. To mitigate this risk, companies should use their own, secure private clouds. Role-based access controls should be used to limit AI network access, and organizations' should understand how and where GenAI data is processed in their private cloud.

Cybersecurity concerns

GenAI models are subject to potential vulnerabilities impairing the model integrity and security which lead to data leakage, unauthorized access and inappropriate output. To mitigate this risk, companies must strengthen cybersecurity protocols, train employees on new security risks and adopt security measures to protect the GenAI models from adversarial attacks.

GenAI, if deployed with inadequate security measures, could lead to significant business implications that extend beyond technical challenges.

What are the general risks associated with GenAI? (cont.)

Reputational damage

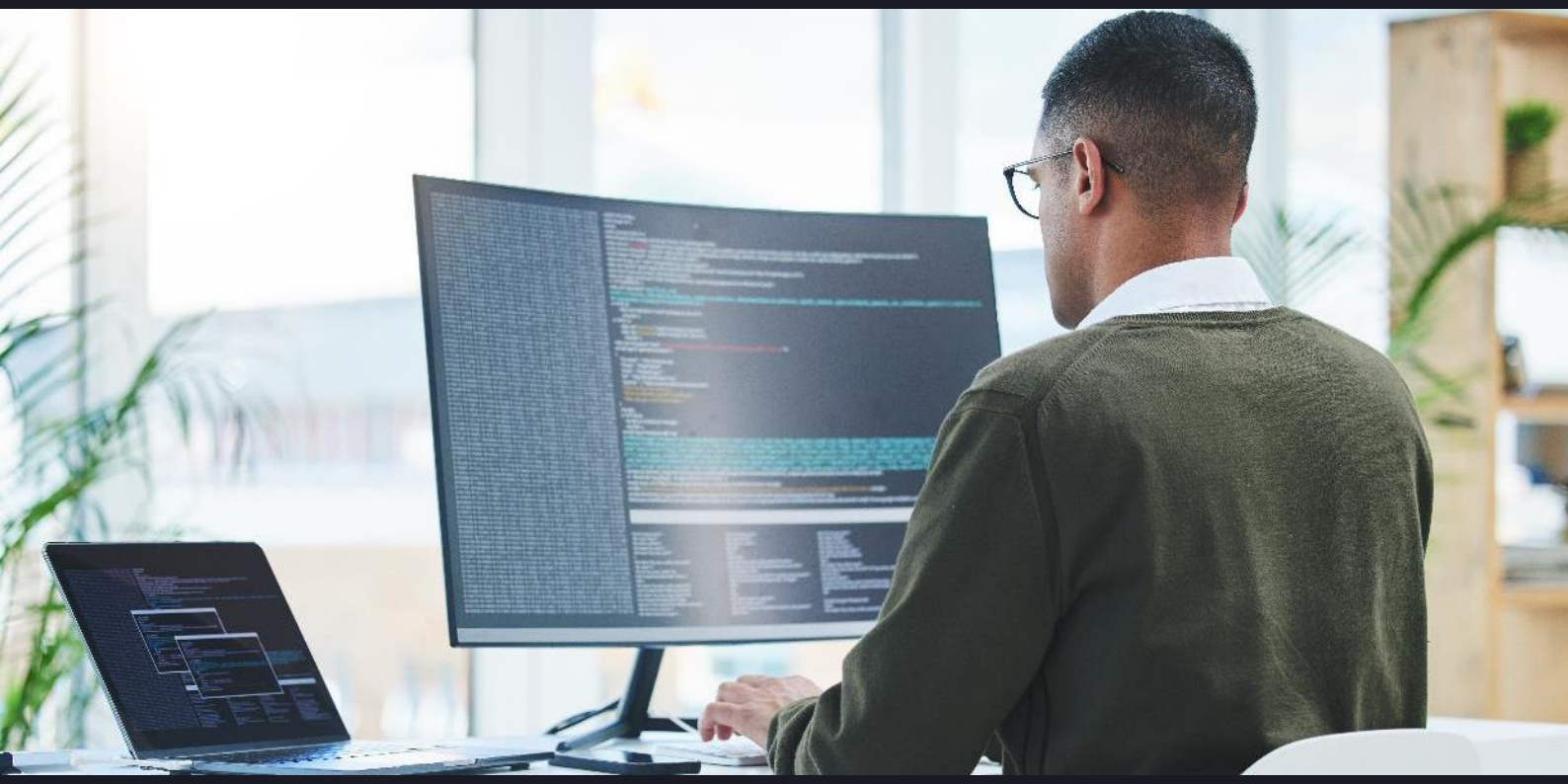
- Customer trust would be significantly affected following security breaches associated with GenAI deployments, especially if customer data is compromised. The impact is not limited to existing customers. It can also deter engagement of potential customers.
- Brand image which has been cultivated over time would be tarnished by security breaches.

Discrimination and bias

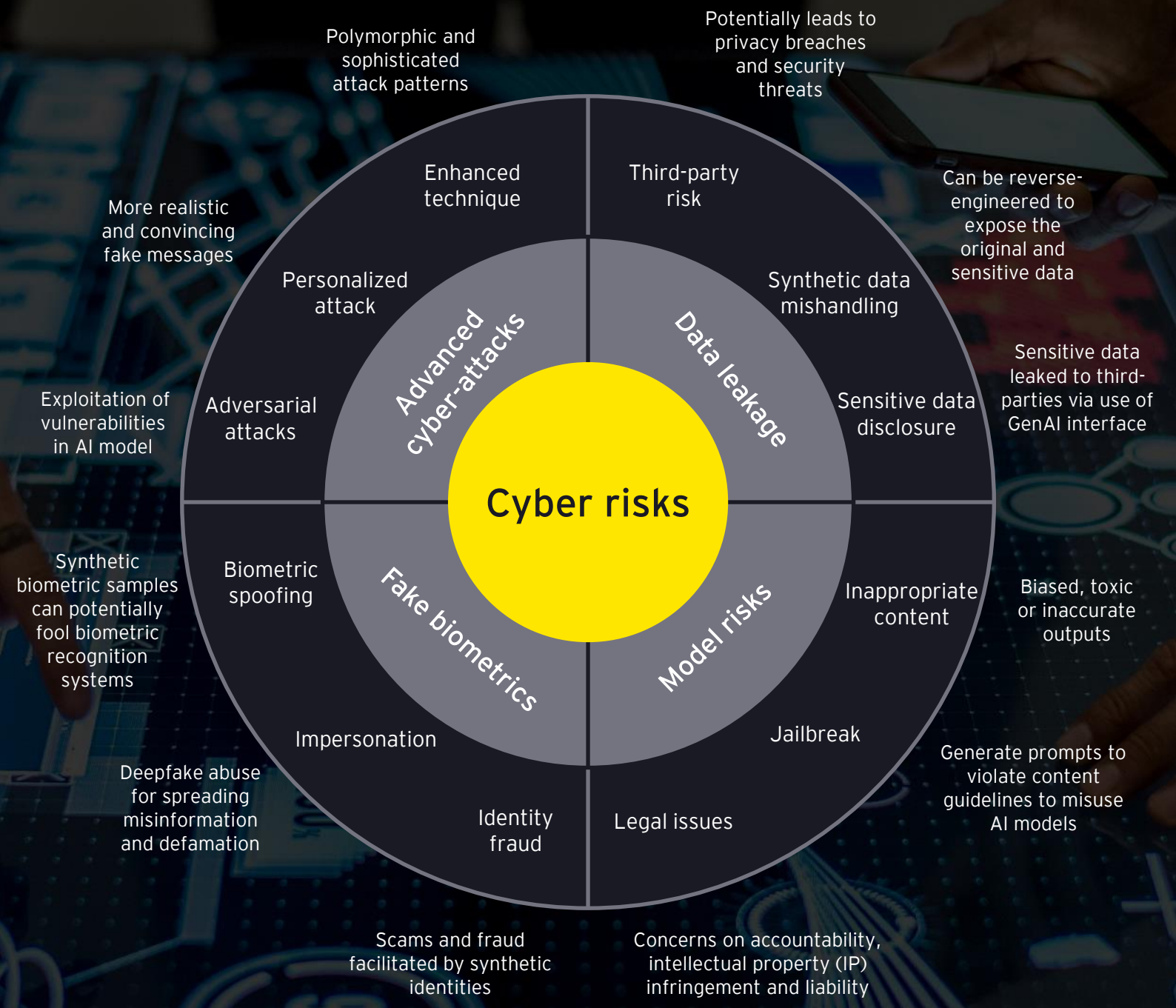
- Legal fines and actions could be substantial in case of security breaches which result in violation of data privacy regulations, particularly due to weak security controls implemented in the GenAI deployment.
- GenAI has been increasingly deployed in customer-facing use cases for service delivery, or business collaboration, as part of contractual commitments. Security breaches resulting in failure in fulfilling contractual obligations could lead to legal challenges and financial penalties.

Loss of competitive advantage

- Organizations adopt GenAI to drive innovation, personalization and efficiency to gain competitive advantage and differentiate themselves. Security breaches in GenAI could immediately undermine this competitive edge.
- Failure to secure proprietary information – especially commercial secrets – in GenAI deployments could lead commercial rivals gaining access to this information with a resulting loss of competitive advantage.



What are the cyber-specific risks in GenAI deployments?



Mitigating the key risks ...

Adopt Security by Design and Privacy by Design principles in the end-to-end process of GenAI model deployment, to ensure that cyber risks are identified and mitigated before deployment – noting existing security controls can often be leveraged and extended.

Key cybersecurity domains	Implementation measures
Environmental security ▪ The infrastructure, including on-premise and cloud environments, supporting the GenAI systems and data should be secured to mitigate the risks of unauthorized access, known vulnerabilities, insecure and unnecessary services, etc.	▪ **Physical security:** Safeguarding physical premises hosting the computational resources through access control, surveillance system, environmental controls, etc ▪ **Network security:** Securing the access to the relevant networks, including cloud environment, through firewalls, intrusion detection systems and network monitoring systems ▪ **Runtime protection:** Protecting the GenAI model and data during its operation through strong authentication and access control mechanisms ▪ **Trusted execution environments:** Shielding the GenAI model operations from potential threats and unauthorized access through secure enclave technology to enable the execution of GenAI models in a secure and isolated run-time memory space ▪ **Vulnerabilities management:** Patching of software supporting the GenAI models operations to ensure all known vulnerabilities are patched on a timely basis ▪ **Auditing and monitoring:** Detecting anomalies, potential vulnerabilities and unauthorized access attempts through continuous monitoring of the processing environment and regular audits
Data security ▪ Data should be protected throughout the data life cycle from collection, processing, storage, transmission and disposal - via governance frameworks - to ensure compliance with all relevant regulations.	▪ **Data privacy:** Implementing privacy controls in the data collection process to ensure consent from users is obtained for use in training AI models; only necessary personal data is collected with reference to the purpose of the GenAI model; and data de-identification measures are implemented to mitigate the risk of data breach ▪ **Data protection:** Securing the data through encryption, role-based access controls and multi-factor authentication to protect the confidentiality, integrity and availability of data ▪ **Data retention and secured data disposal:** Implementing secure data disposal process to remove data with reference to relevant regulations which mandate data retention period ▪ **Third-party security:** As third parties are likely to be involved in GenAI deployment, it is important to govern the access of third parties' access to GenAI assets through restricting access rights, incorporating contractual security requirements and ongoing monitoring

Mitigating the key risks ...

Key cybersecurity domains

Model security

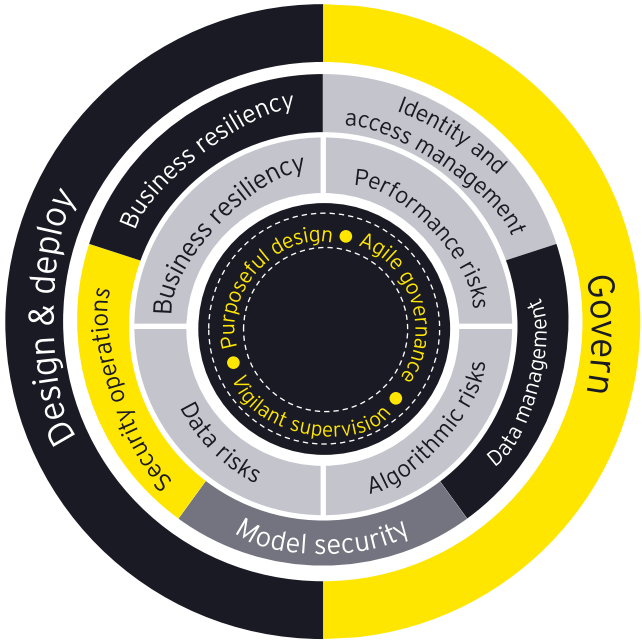
- GenAI models should be secured to mitigate the risk of producing biased, toxic or inappropriate output and guard against adversarial attacks through robust model testing and validation, legal review and compliance assessment, and ongoing monitoring and audit.

Implementation measures

- **Adversarial training:** Incorporating adversarial samples during the training phase to enable models to recognize and correctly classify potential deceptive inputs in order to enhance model resilience
- **Generalization:** Implementing techniques such as regularization and cross validation to enhance the generalization capabilities of the model to prevent attackers from identifying the individual data points in the training data
- **Anomaly detection:** Monitoring the predictions generated by the model for anomalies or unexpected patterns and follow up on the inconsistencies identified
- **Regular retraining:** Retraining the model on a clean and validated dataset periodically to ensure that the effect of malicious patterns introduced by attackers during model training phase, if any, are nullified
- **Gradient masking:** Obscuring or masking models' gradients to limit the information available to attackers to craft adversarial samples to deceive the model
- **Input validation:** Implementing input validation controls to validate for potential adversarial modification and establishing mechanisms to flag or reject suspicious input to prevent the model from being deceived
- **Limiting interactions:** Restricting interactions with the model to trusted environments or application programing interfaces (APIs) and monitoring the interactions with external parties to prevent the model from being accessed by unauthorized parties
- **Ongoing assessment:** Conducting regular assessments to evaluate the ongoing reliability of the adopted technology solutions and assess the capability in addressing emerging fraud schemes or attacks relevant to the GenAI models
- **Adversarial testing:** Conducting penetration and red team testing on the model to evaluate model resilience and identify vulnerabilities with timely remediations

EY Generative AI risk and governance framework

A robust risk management framework and governance processes that establish organizational standards for ethical and responsible use of GenAI based on NIST AI RMF, MITRE ATLAS, OWASP Top 10 for LLM, ENISA, HITRUST, ISO/IEC 23894, ISO 42001, etc.



What next?

While leading organizations are rightly enthusiastic in adopting GenAI, they are still at the early stages in terms of embedding cybersecurity into their deployments.

The most successful organizations will be those which can recognize and articulate the value of cybersecurity in the GenAI era, giving their stakeholders the confidence to move at speed on their adoption journeys.

Establish GenAI principles and guardrails to support experimentation

As businesses rapidly experiment and adopt GenAI, it is essential for organization to move quickly to protect and accelerate the rate of innovation.

Help the business get use cases to market faster

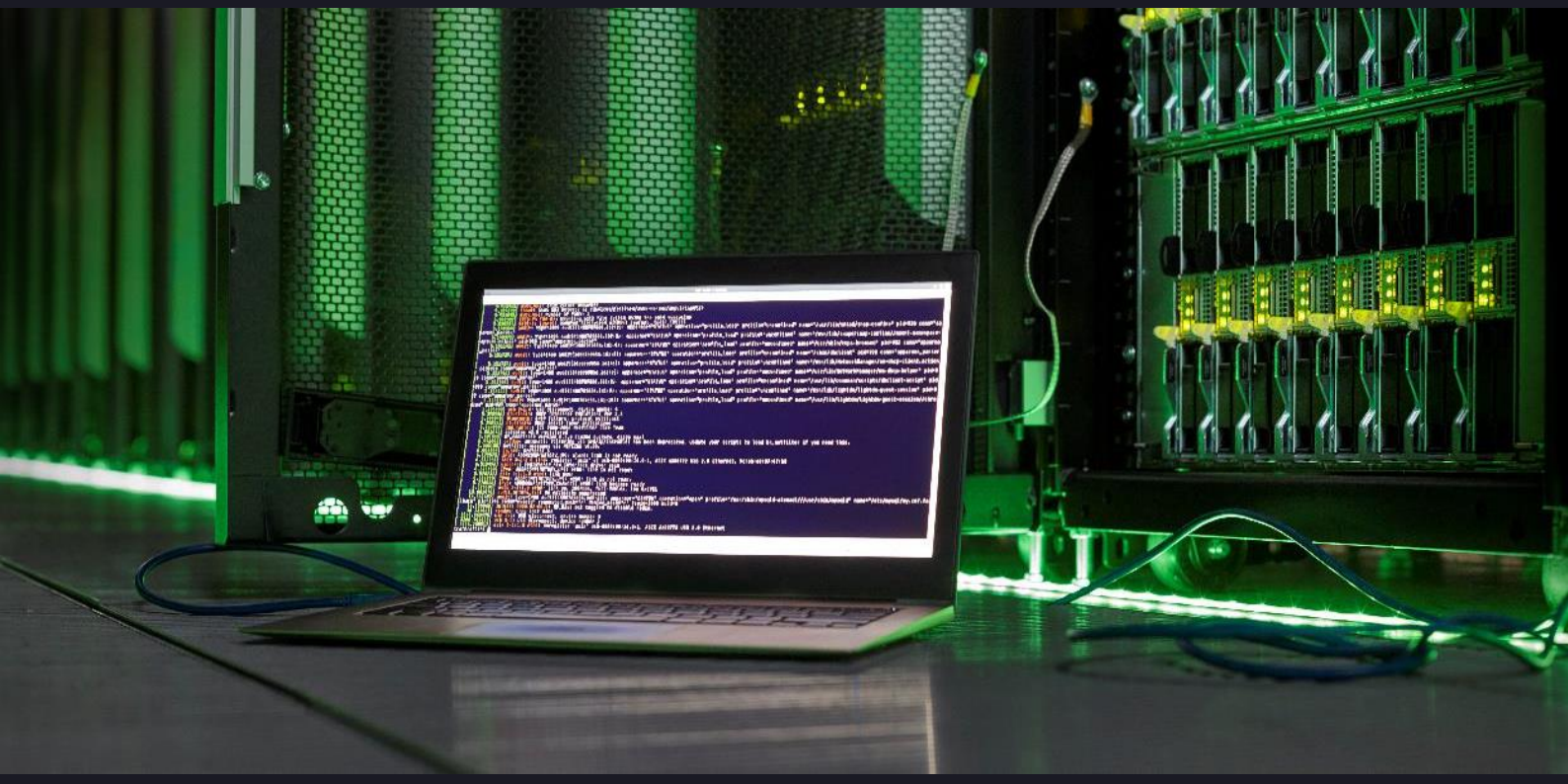
Develop a pre-configured and pre-sanctioned set of architectures, integration patterns and technology stack components to support business use cases. Make secure by design the fastest route to market in your organization.

Target cybersecurity enablement

Cyber executives can be valuable collaborators by developing a practical GenAI security and risk framework that aids in getting to "yes" for the business, while remaining within risk tolerances.

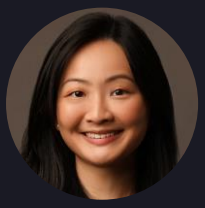
Gain visibility of the AI attack surface and third-party ecosystem

Organizations with existing strong cybersecurity foundations should have strategies in place to manage all cyber risks across the attack surface and their third-party ecosystem. Expanding this to cover new GenAI attack surfaces and third parties will allow organizations to adopt GenAI with greater confidence.



Authors

Jeremy Pizzala
EY Asia-Pacific Cybersecurity Consulting Leader
jeremy.pizzala@hk.ey.com



Yi Fang Chua
EY Oceania Consulting Cybersecurity Partner, Financial Services
yi.fang.chua@au.ey.com

Alan Lee
Partner, Financial Services, Consulting
Ernst & Young Advisory Services Limited
alan.lee@hk.ey.com

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

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. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

© 2024 EYGM Limited.
All Rights Reserved.

EYG no. 008560-24Gbl
ED None

This material has been prepared for general informational purposes only and is not intended to be relied upon as accounting, tax, legal or other professional advice. Please refer to your advisors for specific advice.

ey.com