AITI
Authority for Infocommunications Technology Industry of Brunei Darussalam

Guide on

Artificial Intelligence (AI) Governance and Ethics for Brunei Darussalam

Second Edition



Contents

Introduction 2

Definitions & Terminology 5

Purpose & Scope 8

Advancements of AI and Its Risks 12

AI System Lifecycle 15

Guiding Principles on AI Governance and Ethics 18

AI Governance Framework 34

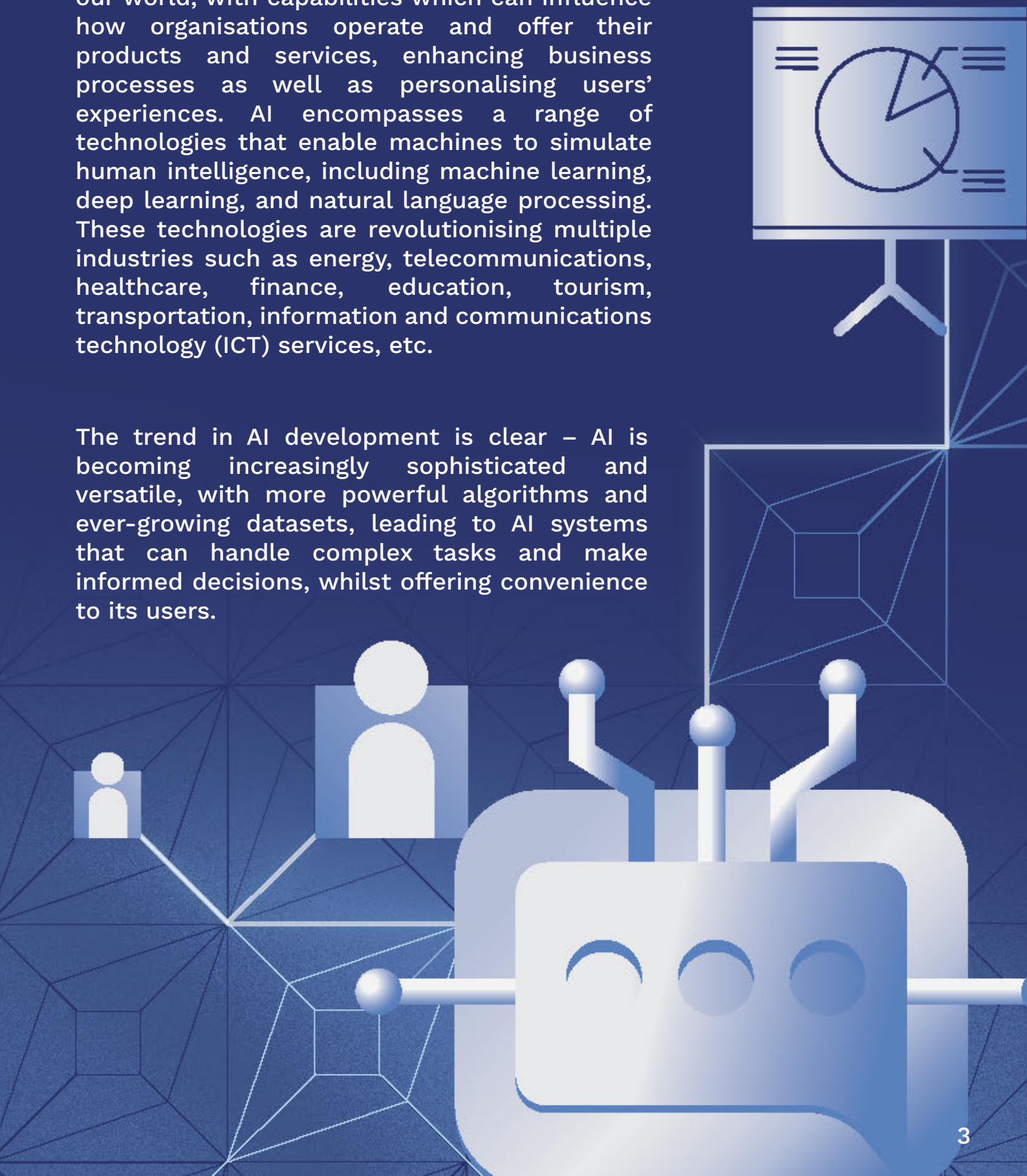
Use Cases on AI Governance and Ethics in Brunei Darussalam 44

Introduction

01

Artificial Intelligence (AI) is rapidly transforming our world, with capabilities which can influence how organisations operate and offer their products and services, enhancing business processes as well as personalising users' experiences. AI encompasses a range of technologies that enable machines to simulate human intelligence, including machine learning, deep learning, and natural language processing. These technologies are revolutionising multiple industries such as energy, telecommunications, healthcare, finance, education, tourism, transportation, information and communications technology (ICT) services, etc.

The trend in AI development is clear – AI is becoming increasingly sophisticated and versatile, with more powerful algorithms and ever-growing datasets, leading to AI systems that can handle complex tasks and make informed decisions, whilst offering convenience to its users.



However, it is crucial to acknowledge that alongside the immense potential of AI lies a complex threat landscape, such as the rise of misuse for disinformation, biases in AI algorithms, implications of job displacement and other associated security risks. Proper safeguards are thereby essential to prevent misuse and exploitation of these AI products and services. These considerations are paramount as we move towards the journey of a Smart Nation and the Brunei Vision 2035.

In order to shape a digitally inclusive future for Brunei Darussalam, it is more than about designing, developing, deploying and using emerging technologies, but also leveraging the power of these technologies **responsibly and ethically,** with appropriate cultural and religious considerations as well as alignment with the principles of Melayu Islam Beraja (MIB) and Maqasid of Shariah. Such considerations play an important role in the efforts to build a safe and trusted digital ecosystem in Brunei Darussalam.

Definitions and Terminology

This section will include relevant definitions for the purposes of the Guide on AI Governance and Ethics for Brunei Darussalam ("the Guide").

02

Artificial Intelligence (AI)

Refers to the discipline of making analytical machines intelligent, enabling an organisation to function appropriately and with foresight.

Source: Association of Southeast Asian Nations (ASEAN) Guide on AI Governance and Ethics

AI Systems

Refers to a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to:

(i) perceive real and/or virtual environments;
(ii) abstract these perceptions into models through analysis in an automated manner (e.g. with machine learning), or manually; and
(iii) use model inference to formulate options for outcomes.

AI systems are designed to operate with varying levels of autonomy.

Source: ASEAN Guide on AI Governance and Ethics

Deepfakes

Refers to a Generative AI technique that uses deep learning to generate realistic images, audio or videos.

Source: ASEAN Expanded Guide on AI Governance and Ethics

Deep Learning

Refers to a subset of machine learning. It is very loosely based on the information-processing architecture of the brain – albeit far from matching its ability – enabling systems to find ways of re-representing input data that facilitate making highly accurate predictions.

Source: ASEAN Guide on AI Governance and Ethics

Deployer

Refers to an entity that uses or implements an AI system, which could either be developed by their in-house team or via a third-party developer.

Source: ASEAN Guide on AI Governance and Ethics

Developer

Refers to an entity that designs, codes, or produces an AI system.

Source: ASEAN Guide on AI Governance and Ethics

Foundation Model

Refers to core technologies powering most Generative AI through probabilistic statistical models that are trained on content and designed to generate it.

Source: Expanded Guide on AI Governance and Ethics – Generative AI

Generative AI System

Refers to one or more foundation models working together, combined with the enabling infrastructure, user interface, additional safety measures and application layer that modifies how they are to be used.

Source: ASEAN Guide on AI Governance and Ethics

Human-in-the-loop

Suggests that human oversight is active and involved, with the human retaining full control and the AI only providing recommendations or input. Decisions cannot be exercised without affirmative actions by the human, such as human command to proceed with a given decision.

Source: Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC) Singapore's Model AI Governance Framework, Second Edition

Human-out-of-the-loop

Suggests that there is no human oversight over the execution of decisions. The AI system has full control without the option of human override.

Source: IMDA and PDPC Singapore's Model AI Governance Framework, Second Edition

Human-over-the-loop

Suggests that human oversight is involved to the extent that the human is in a monitoring or supervisory role, with the ability to take over control when the AI model encounters unexpected or undesirable events (such as model failure).

Source: IMDA and PDPC Singapore's Model AI Governance Framework, Second Edition

Machine Learning

Refers to a subfield in AI where algorithms learn by identifying patterns and correlations within data using statistical techniques to enhance performance, all without being explicitly programmed.

Source: ASEAN Guide on AI Governance and Ethics

User

Refers to an entity or person (internal or external) that interacts with an AI system or an AI-enabled service and can be affected by its decisions.

Source: ASEAN Guide on AI Governance and Ethics

Purpose & Scope

03

Audience

The Guide is intended for organisations which integrate or adopt AI solutions within existing or upcoming business and IT systems. It is intended **for organisations that design, develop, deploy and/or use AI technologies** in Brunei Darussalam. This includes organisations involved in various aspects of AI, including system designers, developers, researchers, providers, deployers and users of AI technologies as well as those hiring or outsourcing such services to third-party companies which may be located overseas.



Objectives

The overall objective is to build trust among stakeholders in managing the risks related to AI technologies, through promoting governance and ethical practices as outlined in the guiding principles.

Purpose

This document serves as a practical guide for governance and ethics in AI, through a principles-based approach suitable for various organisational contexts. It provides recommendations and guidance on the principles and safeguards which organisations should consider. This includes the following:

Risk Management and Mitigation
To proactively identify, assess and reduce potential harms associated with AI systems such as security threats, biased decision-making, etc.

Public Trust and Confidence
To promote transparency and provide explanations for how the AI systems work.

Ethical and Responsible Use of AI
To embed core ethical principles into the entire AI system lifecycle.

Responsible Innovation
To balance between implementing necessary guardrails whilst fostering rapid technological innovation.

Cross-sectoral Regulatory Compliance
To ensure AI deployment adheres to a rapidly evolving landscape of applicable laws, regulations and standards.

Accountability
To define clear roles and responsibilities for relevant teams involved in the entire AI value chain.

Scope

It is important to note that the Guide is developed to serve as a baseline set of considerations and safeguards in accordance with the following considerations:

Technology-agnostic
The Guide will not focus on specific systems or software and will apply regardless of any technological components within AI systems.

Sector-agnostic
The Guide will serve as a baseline for organisations operating in any sector to adopt. Specific sectors or organisations may choose to include additional considerations and safeguards to meet their needs.[1]

Note

The Guide is a working document developed by Authority Infocommunications for Technology Industry of Brunei Darussalam (AITI), in collaboration with the AI Governance and Ethics Working Group.[2] The Guide may be amended periodically with prior notice, to ensure the document remains relevant to emerging trends, practices and technologies. In essence, the Guide serves as a foundational resource for organisations in Brunei Darussalam to responsibly design, develop, deploy and/or use AI technologies whilst ensuring ethical considerations and risk management factors are adequately addressed.

[1] Sector-specific guides may be developed once the deployment and use of AI technologies are more prevalent and mature for organisations in Brunei Darussalam. Any existing and upcoming guides developed by respective sector agencies would act as a complement to this Guide.

[2] The members of the AI Governance and Ethics Working Group consists of representatives from organisations as indicated under page 53.

Advancement of AI and its Risks

04

Traditionally, AI systems have drawn on effectively using fundamental machine learning and deep learning approaches in order to develop descriptive, predictive or diagnostic outputs.

A major advancement in the use of AI systems can be seen through the use of Generative AI, whereby it is able to draw on large datasets to produce content which appear to be created or written by humans.

Such contents may include text, images, audio, videos, computer code, etc.

These can bring **benefits** to businesses and its users such as the following:

Increases efficiency and productivity

It automates complex tasks and processes data faster than humans, leading to significant productivity gains and reducing operational costs.

Improves decision making

It allows for better analyses of datasets, leading to be more informed, accurate and data-driven decisions.

Improves accessibility

It provides users with ways to customize its solutions across a range of tools and services.

Business continuity

Virtual assistances operating 24/7 ensures continuous services and support.

However such developments in AI can bring about **enhanced risks** raising ethical, legal and societal challenges. Such risks can be seen as follows:

Mistakes and hallucinations

The creation of appealing content which can appear to be highly coherent but includes recommendations which are erroneous or codes which are susceptible to vulnerabilities.

Factually inaccurate responses and disinformation

The generation of false or misleading information which can be amplified, affecting public perception and eroding reliance from more trustworthy sources.

Deepfakes, impersonation, fraudulent and malicious activities

The development of human-like phishing emails and creation of deepfake videos which can impersonate real individuals and deceive others into sharing personal or confidential data.

Infringement of intellectual property rights

The use of copyrighted materials without permission which can lead to legal repercussions if it is used without an appropriate legal basis or if the generated content closely resembles original copyrighted work.

Privacy and confidentiality

The ability to memorise and reproduce training data which can reconstruct sensitive information through malicious querying of the system, compromising privacy of individuals represented by the datasets.

Propagation of embedded biases

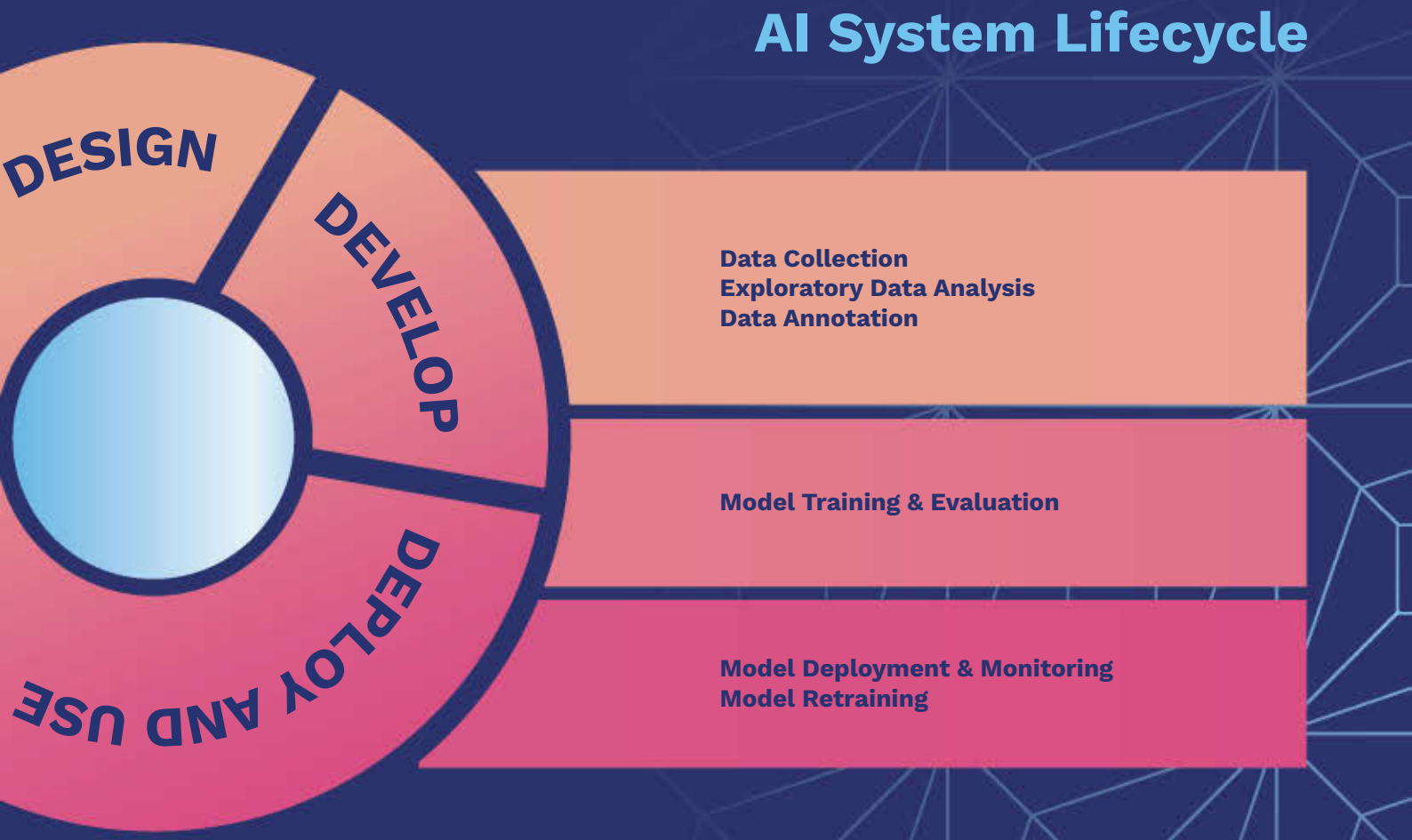
The reinforcement of biases and stereotypes inherited from training data which can lead to biased or toxic outputs that are unfair.

These risks can be mitigated through a shared responsibility amongst system designers, developers, researchers, providers, deployers and users to practice good AI governance as mentioned in Section 7. It is worth noting that mitigating and addressing these risks should not hamper innovation but rather stimulate ethical development and use of AI technologies.

AI System Lifecycle

05

Understanding AI's potential and threats requires examining its development process, known as the AI system lifecycle. It is important to incorporate or align this lifecycle into organisation's system development and project management framework. This lifecycle refers to various stages that an AI system goes through and can be broadly divided into these key stages:



Design

Data Collection, Analysis and Annotation

This initial stage involves defining the system's purpose, its intended users, the amount of data, its use cases, secure data management practices[3] and the ethical considerations. It encompasses the collection and processing of data to ensure data quality, data integrity, relevance, and absence of bias. At this stage, organisations should consider the sourcing of reliable data as well as engaging relevant and diverse stakeholders to ensure the system meets its main purpose.

[3] This is in reference to any applicable standards, guidelines, policies and/or any relevant laws within Brunei Darussalam.

Development

Model Training and Evaluation

Upon model creation, rigorous testing is crucial. This stage ensures the AI system performs as intended, delivers accurate results and avoids unintended consequences. It also involves assessing the system's fairness[4], robustness, and explainability which can provide humans with the understanding on how the AI system arrives at its decisions. To ensure consistency, organisations may want to continue to incorporate continuous feedback loop from the relevant and diverse stakeholders to refine and improve the AI model throughout its development.

Deployment and Use

Model Deployment and Monitoring

Once validated, the AI system is deployed in the real world. Continuous monitoring is essential to track the system's performance, identify any biases or errors that may emerge over time and ensure it adapts to changing circumstances.



At this stage, it may be essential to evaluate the safeguards which may include levels of human interventions and checks on the data inputs, algorithm and model outputs.

Model Retraining

Data gathered during monitoring is used to refine the AI system. This may involve retraining the model with new data or modifying the algorithms to address identified issues. This iterative process of feedback and improvement is crucial for maintaining the system's effectiveness and mitigating potential risks. Such periodical review is important to ensure that the system continues to be automated and adapts itself in accordance to ethical frameworks, laws and regulations.

[4] For example, minimising bias in algorithm selection and training the AI model to use only high-quality or representative data.

Guiding Principles on AI Governance and Ethics

06

7 Guiding Principles

The seven (7) guiding principles[5] below help to gain users' trust in AI systems and tools through guidance on how it should be designed, developed, deployed and used by organisations. The principles below take into consideration its impact to technology development, harnessing the foundation of a safe and trusted digital economy for Brunei Darussalam.

- Transparency and Explainabilty
- Data Protection and Data Governance
- Security and Safety
- Robustness and Reliability
- Fairness and Equity
- Human Centricity
- Accountability and Integrity

[5] The principles cite elements of the ASEAN Guide on AI Governance and Ethics.

Transparency and Explainability

Transparency involves disclosing to users that an AI system is being used and contributed in the decision-making, the types of data it uses, and its purpose. This ensures that users understand that such system is utilised, therefore enabling users to make an informed choice of whether to use the AI-enabled system.

Explainability refers to the ability of an AI system to convey the reasoning behind its decision and conclusion in a format that is simple and understandable. This allows users to understand the factors contributing to the AI system's recommendation.

It is worth noting that there may be situations where organisations may not be able to fully practice the above principles and certain exceptions may apply with justifications. For example, disclosing confidential or proprietary information at risk such as information containing organisation's Intellectual Property (IP).



AI systems should enable the explicit labelling of AI-generated content to assist users in differentiating between automated or human-created information, thereby eliminating any ambiguity in making an informed choice.

Developers

Organisation A is a company that provides real-time predictive insights and offers its cost-predictor solutions to clients in the healthcare sector.

It practices the Transparency principle by disclosing to its clients the exact parameters used in developing their AI model and providing detailed explanation on all algorithms that had any foreseeable impact on its decision. Further, it practices the Explainability principle by training its AI model to explain each prediction result to its users without the need to go through Organisation A's support staff.

Deployers

User A is the Finance Officer of a healthcare company who uses an AI system as a tool to assist in predicting the cost of medical treatments for its customers.

She practices the Transparency principle by disclosing to her supervisor that she has used the AI system as a tool to recommend the potential cost of the selected treatment.

Further, she practices the Explainability principle by describing to her supervisor the datasets in which the AI system is trained on and the reasonable steps she has taken to validate and confirm the actual cost.

Data Protection and Data Governance

AI systems should have proper mechanisms in place in order to maintain and protect the quality and integrity of data throughout their entire lifecycle.

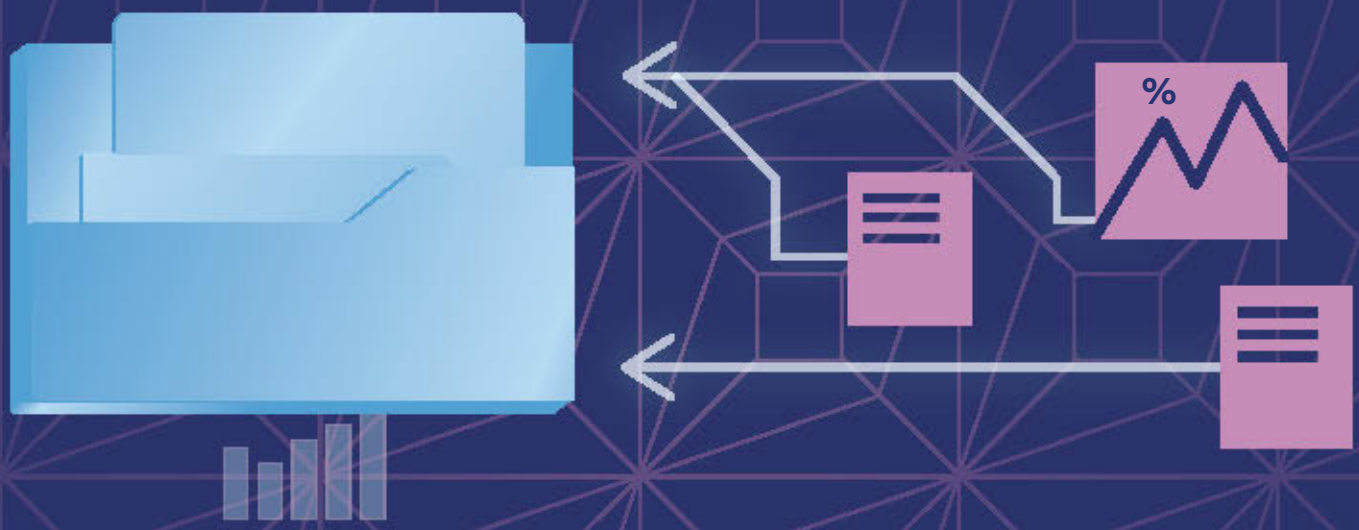
Data protection by design and privacy by design principles should be considered throughout the entire lifecycle of the AI system. The way data is collected, used, disclosed, processed, stored, generated, and deleted throughout the AI system lifecycle must comply with applicable laws and regulations, including Brunei Darussalam's Personal Data Protection Order, 2025.



Organisations should be transparent about their data collection practices, including the types of data collected, where it is sourced, how it is used, who has access to it and whether such data will be disclosed and/or anonymised.

Organisations should ensure that necessary consent is obtained from individuals before collecting, using, or disclosing personal data for AI development and deployment, or otherwise have appropriate legal basis to collect, use or disclose personal data without consent. In addition, organisations should consider placing greater sensitivity in the design of AI systems, in particular where it collects, uses or discloses personal data of minors and vulnerable groups.

Data governance frameworks should also be considered and adhered to by designers, developers, deployers and users of AI systems and tools.[6] Organisations should ensure a well-documented and effective audit trail. These frameworks should also be periodically reviewed and updated in accordance with relevant law and regulations.



[6] Organisations may refer to voluntary guides such as "ASEAN Data Management Framework (DMF)" and "ASEAN Model Contractual Clauses (MCC)".

Developers

Organisation B is a company that provides data analytics solutions to clients in the telecommunications sector.

It practices the Data Protection principles by ensuring that it anonymises its client data at source before using it for development. Furthermore, it built an AI model with the ability to automatically detect any personal data, where its algorithm was trained to minimise the use of this data in developing the AI model.

It practices the Data Governance principles by documenting the lineage of data and logging data consistently across multiple components in a secure and centralised log storage. In addition, it had also ensured proper documentation of data sources by obtaining reliable data directly from clients instead of using unreliable third-party datasets.

Deployers

User B is the Data Protection Officer of a telecommunication company that uses an AI tool to conduct data analytics for business improvement purposes.

She practices the Data Protection principles by ensuring that the training of the AI system will only be based on the necessary data only. She consults relevant teams to weigh the pros and cons of training the AI system with either anonymised or necessary personal data.

She practices the Data Governance principles by clearly documenting the reasons and justifying the selection of the proposed datasets. Where personal data is involved, she ensures the necessary data protection and security measures are in place before its use.



Security and Safety

Safety involves placing emphasis on a risk-prevention approach by identifying and mitigating potential risks in order to ensure the safety of developers, deployers and users of AI systems through proper safeguards.

This can be done by conducting regular impact or risk assessments to identify suitable precautionary measures that enable human intervention when required. An example scenario is where autonomous vehicles may cause injury to pedestrians.

Safety of the public and the users of AI systems should be of utmost priority in the decision-making process of AI systems and risks should be assessed and mitigated to the best extent possible.

Before deploying AI systems, organisations should conduct risk assessments and relevant testing or certification and implement the appropriate level of human intervention to prevent harm when unsafe decisions take place. This may be verified further by engaging an internal or external agency to evaluate and carry out risk assessments from a third-party perspective. The risks, limitations, and safeguards of the use of AI should be made known to the user.



Security refers to ensuring the cybersecurity of AI systems[7], which includes mechanisms against malicious attacks specific to AI such as data poisoning, model inversion, the tampering of datasets, as well as other attacks designed to reverse engineer personal data used to train the AI.

Organisations should implement technical security measures like robust authentication mechanisms and encryption. Organisations should also involve its cybersecurity leader throughout the AI lifecycle to maintain a strong overall security posture.



Developers

Organisation C is a technology start-up company which provides a digital care programme to tackle obesity-related chronic diseases.

It practices the Safety principles by tackling potential risks at an early stage and conducting rigorous tests on its AI models by using a build-and-deploy process that included automated code testing such as unit testing and static analysis. This allowed the company to highlight possible vulnerabilities in the code before running its AI model.

It practices the Security principles by implementing measures to strengthen their AI models by conducting annual penetration tests with an independent third-party security firm and performing post-mortem analyses to identify root causes and implementing future controls.

[7] Organisations may refer to international best practices such as "Guidance for Secure AI System Development" developed by National Cyber Security Centre, United Kingdom.



Organisations should also implement safeguards to protect AI systems against cyberattacks, data security attacks and other digital security risks. These may include ensuring strong authentication mechanisms, data security safeguards, regular software updates to AI systems, incident response plans, proper access management for critical or sensitive systems and security testing such as penetration tests or vulnerability scanning.

Developers

Organisation D manufactures AI-enabled autonomous vehicles and practices the Safety principle by ensuring the designers, developers and deployers of AI systems provide for human intervention mechanisms for the human driver to easily resume manual driving, where necessary. Such precautionary safeguards allow for the autonomous vehicle system to safely disengage itself in the event that it were to make unsafe decisions.

It practices the Security principles by encrypting all data transmissions to prevent unauthorised access and ensuring real-life Geographic Information System (GIS) monitoring for timely actions to avoid any mishap.

Deployers

User C is the Health, Safety and Environment officer of an oil and gas company who uses an AI-powered CCTV for health and safety of its employees.

He practices the Safety principles by utilising the AI system as a tool to analyse the footage and the validity of its recommendations. Upon validation, he confirms the significant recommendations which could lessen the potential risks of the workplace hazards.

He practices the Security principles by ensuring that the use of such CCTV is strictly for the use of improvement of worker safety only and ensures such data is encrypted and given access to authorised personnel only.

Robustness and Reliability

AI systems should be sufficiently robust to cope with errors during execution and unexpected or erroneous input, or cope with stressful environmental conditions. It should also perform consistently. AI systems should, where possible, work reliably and have consistent results for a range of inputs and situations. AI systems may have to operate in real-world, dynamic conditions where input signals and conditions change quickly.

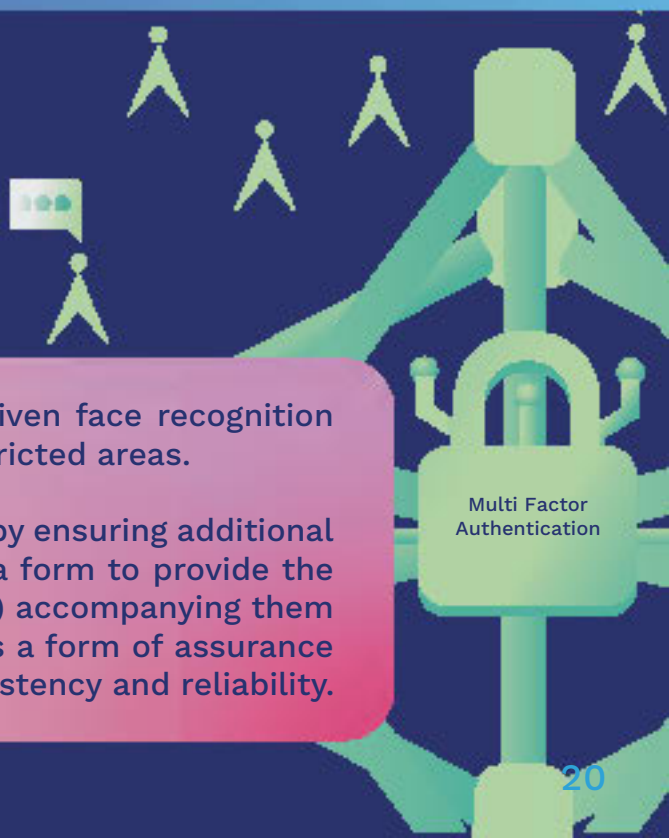
To prevent harm, AI systems need to be resilient to unexpected data inputs, not exhibit dangerous behaviour and continue to perform according to the intended purpose. Notably, AI systems are not infallible and organisations should ensure proper access control and protection of critical or sensitive systems and take actions to prevent or mitigate negative outcomes that occur due to unreliable performances.

Measures such as proper documentation of data sources, tracking of data processing steps, and data lineage can help with troubleshooting AI systems.

Developers

Organisation E is a company that leverages deep learning techniques and combines biometric, geo-location and behavioural analytics with multi-factor to authenticate user identities.

It practices Robustness and Reliability principles by involving engineering, product, sales and research teams in the design and development stages. It conducts rigorous testing before the deployment of AI systems to ensure consistent results across a range of situations and environments. In addition, it ensured proper access controls on protection via a multi-level assurance framework through relevant department's Head of Unit.



Deployers

User D is a Security Officer who utilises an AI-driven face recognition system to record visitors temporary access to restricted areas.

He practices Robustness and Reliability principles by ensuring additional measures are taken by requiring visitors to fill in a form to provide the purposes of the visit, the name of the employee(s) accompanying them and the time in and out of the areas. This provides a form of assurance to record the visits and to monitor AI system consistency and reliability.

Fairness and Equity

Organisations should have safeguards in place to ensure that algorithmic decisions do not further exacerbate or amplify existing discriminatory or unjust impacts across different demographics. The design, development, and deployment of AI systems should not result in unfair biasness or discrimination. An example of such safeguards would include human interventions and checks on the algorithms and its outputs.

Organisations should conduct regular testing of such systems to confirm if there is bias and where bias is confirmed, make the necessary adjustments to rectify imbalances to ensure equity. With the rapid developments in the AI space, AI systems are increasingly used to aid decision-making.

Developers

Organisation F is an Institute of Higher Learning which offers diploma courses to students. Every year, the institution conducts an early admissions exercise which is an aptitude-based admission process that allows students to apply for admissions prior to receiving their final grades.

The institution practices Fairness and Equity principles by adopting a human-over-the-loop approach to invite the unsuccessful applicants for a face-to-face interview to determine the AI system's output and validity. It also ensures that a second review would be conducted by relevant staff (such as admissions committee or academic members of staff) to ensure reliability on the shortlisting of candidates.



If not properly managed, decisions made using an AI system's outputs could perpetuate existing bias against specific demographics. To mitigate discrimination, it is important that the design, development, and deployment of AI systems align with fairness and equity principles. In addition, the datasets used to train the AI systems should be diverse and representative. Appropriate safeguards should be taken to mitigate potential biases during data collection and pre-processing, training, and inference.

AI

Deployers

User E is a Professor of an Institute of Higher Learning who prohibits the use of generative AI as a primary source of information in written assignments. She uses an AI system to identify any work whereby students may not uphold such academic integrity.

The Professor practices Fairness and Equity principles by adopting a human-in-the-loop approach to treat the AI system's recommendations as a suspected incident before validating the output. Upon validation, she invites the suspected students with a face-to-face interview to respond to questions about the situation before determining whether there will be any formal records of misconduct.

Excellence School
Ami n



Human Centricity

AI systems should respect human-centred values and pursue benefits for society, including human well-being. It is key to ensure that people benefit from AI design, development, and deployment while being protected from potential harms and not take advantage of individuals, in particular minors and vulnerable individuals.[8] It is imperative for AI systems to be accessible, especially in instances where AI systems are used to make decisions about humans or aid them.

Human-centricity should be incorporated throughout the AI system lifecycle, starting from the design to development and deployment. Actions must be taken to understand the way users interact with the AI system, how it is perceived, and if there are any negative outcomes arising from its outputs.

Developers

Organisation G is an organisation in the education sector which has developed an AI-enabled Adaptive Learning System (ALS) for deployment within an online learning portal for primary students.

It practices the Human Centricity principles by engaging user research through involving its stakeholders during the design and development such as policymakers, academia, technical experts and its users. It had considered students individual learning readiness and personalised effective support needs for each student.

Deployers

User F is a Mathematics teacher who uses an AI-enabled ALS for online learning of the students in her classes.

She practices the Human Centricity principles by using the AI tool to analyse the questions in which students have mostly answered incorrectly and highlights which subject area she would need to focus on. She adapts her teaching materials to support the needs of each class and personalises revision materials for each student.



[8] It is recommended for organisations to undergo an additional safety review before any deployment of AI systems or tools which affects minors and vulnerable individuals.

Accountability and Integrity

There needs to be human accountability and control in the design, development, and deployment of AI systems. Organisations should be accountable for decisions made by AI systems and should act with integrity throughout the AI system lifecycle.

Organisations should ensure the proper functioning of AI systems and its compliance with applicable laws, internal AI governance policies and ethical principles. In the event of a malfunction or misuse of the AI system that results in negative outcomes, responsible individuals should act with integrity and implement mitigating actions to prevent similar incidents from happening in the future.



To facilitate the allocation of responsibilities, organisations should adopt clear reporting structures for internal governance and procedures for complaints handling, setting out clearly the different kinds of roles and responsibilities for those involved in the AI system lifecycle.[9]

AI systems should also be designed, developed, and deployed with integrity – any errors or unethical outcomes should at minimum be documented and corrected to prevent harm to users upon deployment.

Developers

Organisation H is a tech company in the financial sector which provides an AI Anti-Money laundering Filter Model that identifies predictive indicators of suspicious transactions to reduce the number of false positives generated by the non-AI system, thereby reducing the number of alerts that require manual review.

It practices the Accountability and Integrity principles by ensuring that it sets up a 'Responsible Data Use' framework where a Committee from various departments was appointed to oversee and govern it, which ensured that the AI model complies with legal, compliance, security and any data quality issues.

Deployers

User G is a Compliance officer for a financial institution who uses an AI system to alert him of suspicious transactions.

He practices the Accountability and Integrity principle through a human-in-the-loop approach by reviewing whether the patterns and suspicions of the AI system were valid. Upon review with the consumer, he confirms that there is no legitimate reason for the transactions to be suspicious and retrains the AI system to improve its assessment of suspicious transactions.

[9] This includes outsourced service providers or third-party vendors which have been engaged throughout the AI system lifecycle.

AI Governance Framework

Implementing effective AI governance requires a structured framework in order to balance innovation with risk management, legal compliance and ethical standards. This section of the Guide includes guidance on measures promoting the responsible use of AI that organisations should consider to adopt and is designed around the established methodology of people, process and technology.

07



People

The use of AI can bring significant risks and additional obligations to the organisation. To ensure effective, efficient and acceptable use of AI systems within the organisation, establishing an AI governance framework which includes a multi-level involvement and assurance within the organisation is essential. It is important for organisations to instill a culture of practicing accountability and integrity in designing, developing, deploying AI systems and using AI tools.

Organisations should detail a set of ethical and fair usage principles when developing or deploying AI products, services or systems. When establishing ethical principles for AI, organisations should review their existing corporate values in light of the latest developments on AI technologies and the Guiding Principles mentioned in Section 5. When establishing usage policies for AI, organisations should provide clearer guidance and transparency. It should also ensure that such systems have proper authorisation over the use of any copyrighted materials.

Organisations should take steps to promote adaptability to technological advancements among the employees through reskilling and upskilling programmes. The initiative should consider the opportunities arising from digital transformation which extends beyond enhancing technical skills such as AI governance, AI ethics and AI certification.[10]

[10] An example of an International Standard that promotes effective implementation of AI-related projects within organisations is the PMI Certified Professional in Managing AI (PMI-CPMAI) certification.



Level of human involvement in AI-augmented decision-making

Organisations should determine the level of human involvement in AI-augmented decision-making. This includes:

- Conducting relevant risk impact assessments to determine level of risk.
- Mitigating risks in order to build trust towards the acceptance and greater use of AI technologies.

Human involvement based on level of risk

Human-in-the-loop

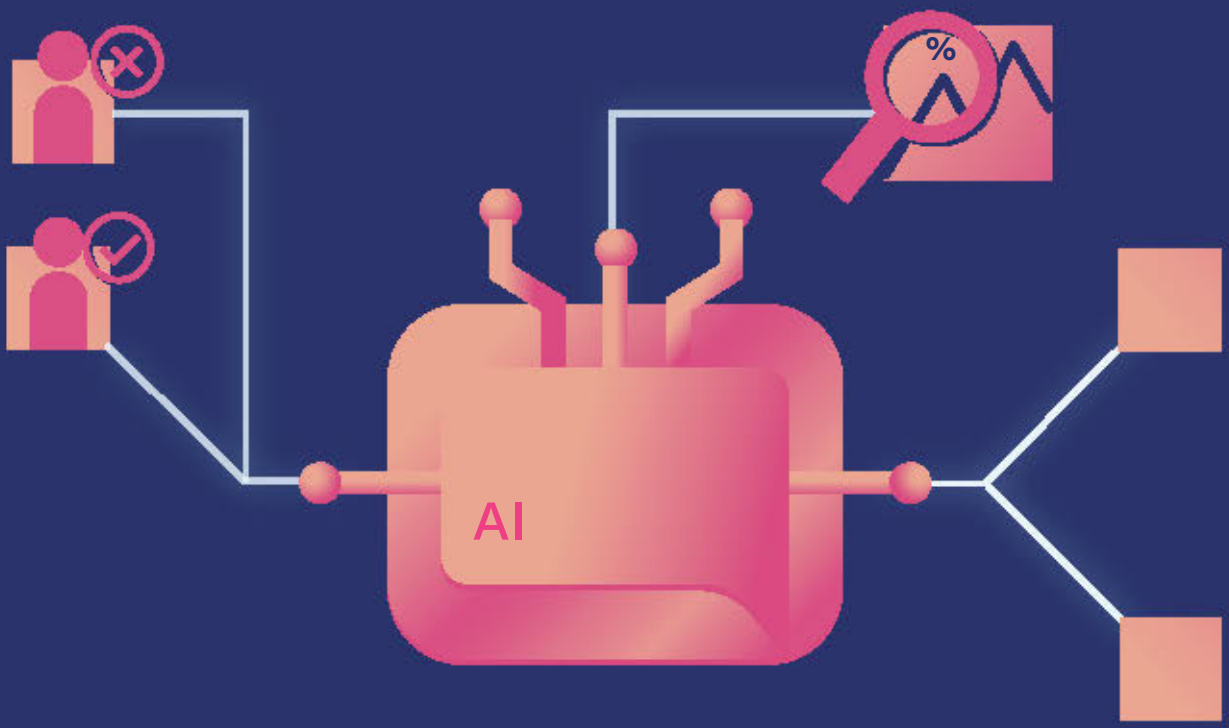
Organisation I uses AI for decision-making on its recruitment process. The AI system assists the organisation in identifying the suitability of potential candidates that meet the expected criteria of minimum entry requirements. In this approach, the AI system provides baseline recommendation or information for humans to weigh their decision.

Human-out-of-the-loop

Organisation J uses AI to enhance its digital marketing activities by learning about customers' purchases. The AI system creates customer profiles to understand their preferences and automatically recommends the products or services within similar categories or interests. In this approach, the AI system does not require any human oversight and has full control over its recommendations.

Human-over-the-loop

Organisation K uses AI to identify possible diagnosis of its patients and provides recommendations on potential medical treatments. Where it results in an undesirable outcome, humans would be able to alter the parameters to reach an acceptable outcome. In this approach, the AI system allows humans to take a supervisory role in its decision when it encounters challenges.



Process

Organisations should have in place appropriate internal governance frameworks that allow them to have oversight over AI technologies and tools.

Internal governance framework

- Develop standards, guidelines, tools and templates to help organisations design, develop, and deploy AI responsibly and ethically.
- Clearly lay out the roles and responsibilities of personnel involved in the responsible design, development and/or deployment of AI.
- Developing procedures for continuous monitoring and feedback to regularly assess and improve AI governance practices.



General principles of risk management that are integrated, structured and comprehensive should be considered.

Risk management should consider the whole system, with all its technologies and functionalities and its impact on the environment and relevant stakeholders.

Risk management framework should be integrated as part of the organisation's AI System framework.[11] Organisations should conduct risk-based assessments before starting any data collection and processing or modelling.

Organisations also should be ready to deploy measures to mitigate risks of unjust bias due to insufficiently representative training, testing and validation datasets. In addition, organisations should consider to ensure that AI systems are adaptable and allow for re-evaluative assessments due to the complexities of risk scenarios and ever-changing environments in the era of technological advancements.



[11] An example of an International Standard that specifies requirements for managing risks throught the AI lifecycle is the National Institute of Standards and Technology (NIST) AI Risk Management Framework.

Risk mitigation measures to support the governance of Generative AI (GenAI)

Content Verification to address mistakes and disinformation

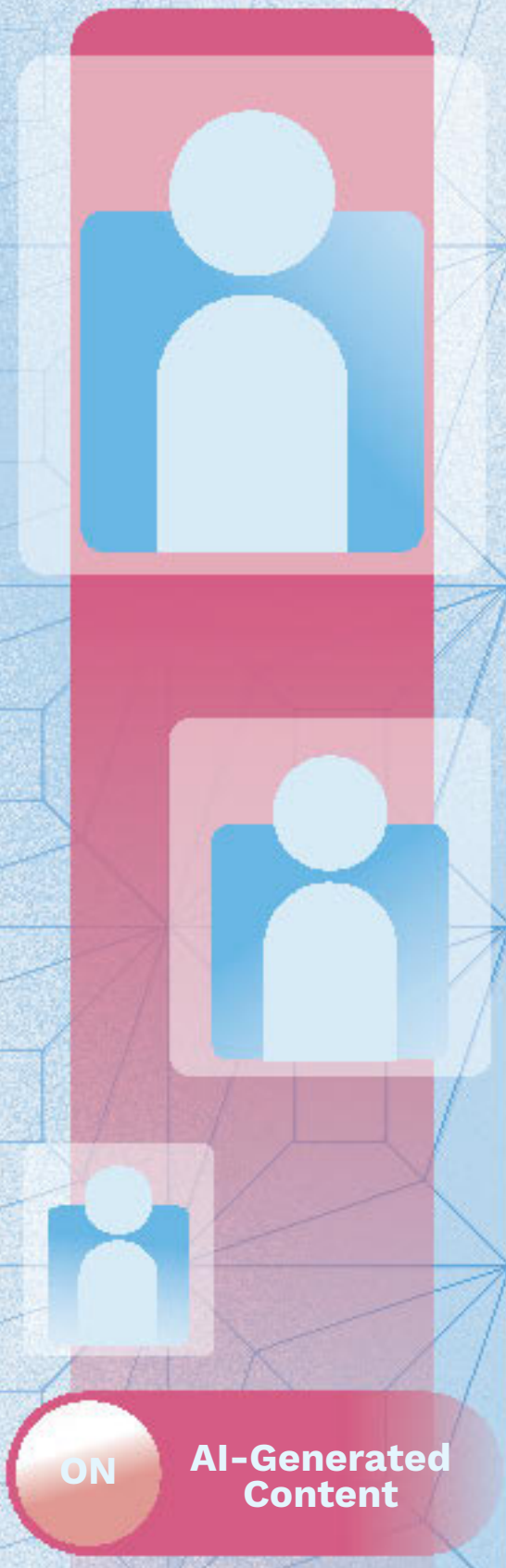
GenAI service providers should estimate the accuracy of information prior to displaying it to users alongside the content requested. Such services should also be able to cite sources indicating where the information was acquired. If the model cannot produce a response with a high level of accuracy and identify its source(s), the output should be replaced with a pre-defined message alerting users that the information generated may not be accurate.

Content Labelling to address malicious deepfakes

GenAI service providers should embed identifiable watermarks in its outputs. This measure helps individuals and security services detect and flag deepfakes and AI-synthesized content without affecting legitimate use. GenAI services should analyse generated content for any inappropriate content as well as the use of public figures' faces and voice samples.

IP Licensing and Due Diligence to address possible infringement

GenAI developers should obtain licenses for any IP included in training data, preventing indiscriminate content scraping. Customers should perform their own due diligence to confirm whether AI models were trained with protected content before using them.





User Data Control to address privacy and confidentially

GenAI service providers should obtain users consent in the use of their data for AI model training purposes. They should also provide an option to remove all prompt history on request in accordance with relevant laws and regulations, such as the Personal Data Protection Order (PDPO), 2025.

Validating outputs to address biases

Users should adopt a practice of regularly verifying and validating AI–generated code and other content, especially in critical applications where accuracy is paramount. Implementing thorough testing and validation processes can help identify and rectify any inconsistencies or errors in the outputs.

Prompt-Injection Safeguards

Implement controls to detect and mitigate prompt-injection attacks, including prompt sanitization, context isolation, and output validation, to reduce the risk of malicious manipulation of model behavior via user inputs.

Technology

Building a resilient system is the cornerstone of building a robust AI system and thereby contributes to building trust in an AI system.

Testing, validation and verification considering the complexity and criticality of the system is crucial. This should not only include the core components such as the AI model, data sets and business logic but also the underlying systems and infrastructure and more importantly to test the system in entirety. Organisations should employ techniques to test applications and create dedicated environments for testing. Creating a digital twin might be a good idea for AI systems used in critical functions.

Organisations may consider using internationally recognised AI governance standards,[12] testing framework[13] and software toolkit that validates the performance of AI systems against a set of internationally recognised principles through standardised tests and is consistent with international AI governance frameworks. Organisations should implement safeguards to help employees adapt to an AI-augmented work environment.



[12] An example of an International Standard that specifies requirements for establishing, implementing, maintaining and continually improving AI Management Systems (AIMS) within organisations is the ISO/IEC 42001:2023.

[13] An example of regional testing framework is Singapore's AI Verify – which is a single integrated toolkit that operates within enterprise environment that can perform technical tests on common supervised learning classification and regression models for most tabular and image datasets.

Use Cases on AI Governance and Ethics in Brunei Darussalam

08

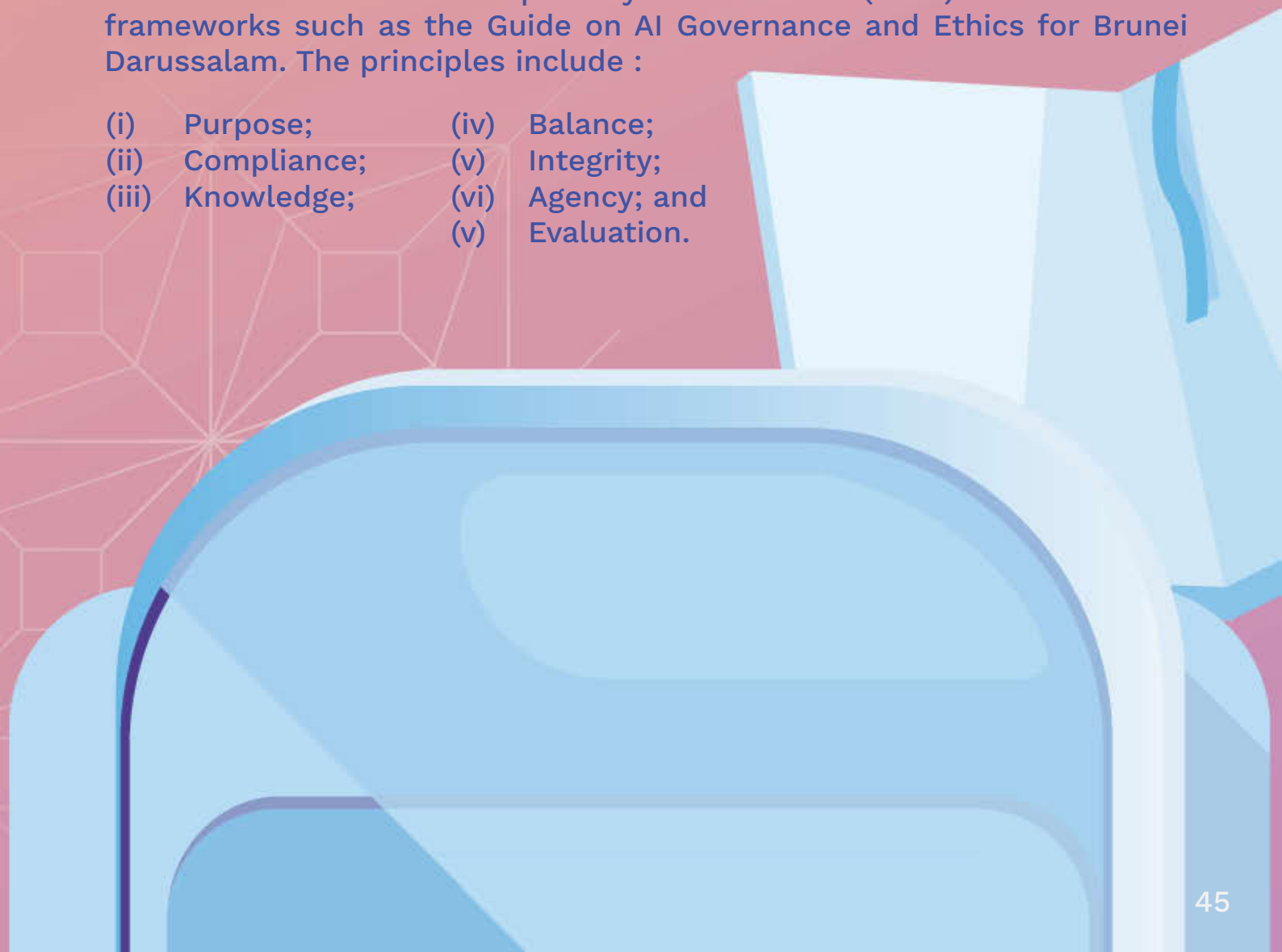
Establishing sector-specific guidelines promoting the responsible use of AI

The Generative AI - Guidance for Education, Handbook Edition by Ministry of Education

The Ministry of Education (MOE) conducted a survey on Core Education, consisting of school leaders and teachers which indicated multiple uses of AI in Education which extended beyond generic experimentation and towards more purposeful and embedded practices. It can be seen that uses of AI showed beneficial use through the delivery of creative tasks for students, brainstorming instructional planning and automating flows on administrative work.

By September 2025, MOE published the "Generative AI - Guidance for Education, Handbook Edition" with the purpose of ensuring the responsible and ethical use of generative AI by the educational institution leaders, administrators, educators, learners, parents and guardians. The Guide encourages seven principles for ethical AI in Education which is grounded by national values. It aligns with international frameworks such as the UNESCO AI Competency Framework (2024) and national frameworks such as the Guide on AI Governance and Ethics for Brunei Darussalam. The principles include :

(i) Purpose;
(ii) Compliance;
(iii) Knowledge;
(iv) Balance;
(v) Integrity;
(vi) Agency; and
(v) Evaluation.





Developing internal policy for the responsible use of AI in academia

Policy and Guidelines on Generative AI Book by Universiti of Brunei Darussalam

The Universiti of Brunei Darussalam (UBD) acknowledged the use of Generative AI tools which have the potential to enhance the effectiveness of educational procedures, customise learning experiences for students and promote the acquisition of modern teaching and learning skills. It is also imperative that principles and practices of academic integrity remain to be the moral code of higher education.

In 2023, UBD developed a “Policy and Guidelines on Generative Artificial Intelligence (GenAI)” book with the purpose of ensuring both educators and students are given clear guidelines on harnessing AI for teaching and learning, while upholding ethical and responsible standards which aligns with the principles in the policy.

Developing internal guidelines and implementing responsible use of AI in industry

AI Governance Guidelines for Responsible LLM use by Brunei Shell Petroleum

As Large Language Models (LLMs) became increasingly accessible and widely adopted, Brunei Shell Petroleum (BSP) recognised the need to establish clear governance guidelines to ensure employees could leverage these powerful AI tools responsibly while maintaining security, confidentiality, and regulatory compliance.

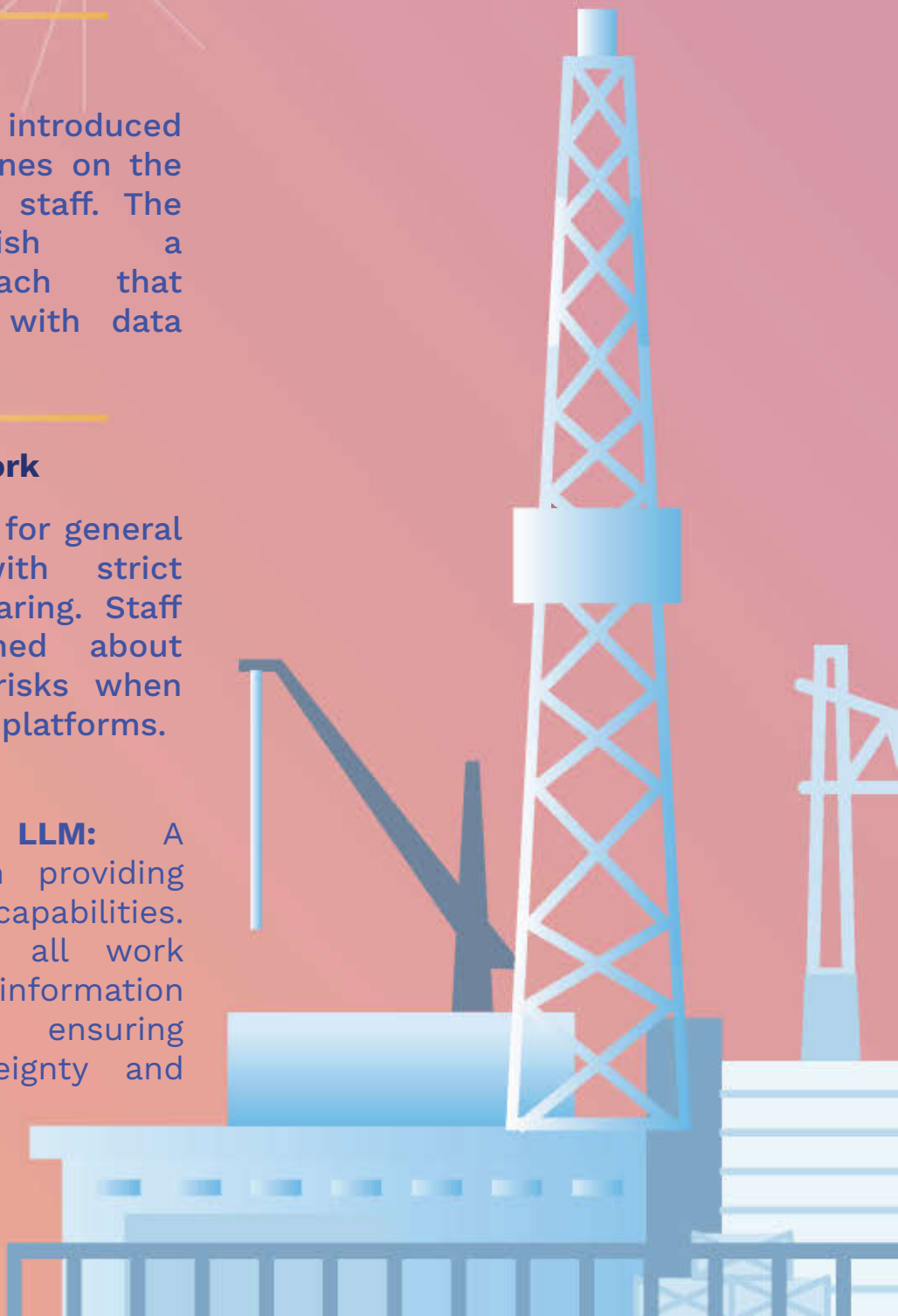
Implementation

In April 2025, BSP introduced comprehensive "Guidelines on the Usage of LLMs to all staff. The guidelines establish a dual-platform approach that balances accessibility with data protection.

Dual-Platform Framework

Public LLMs: Permitted for general productivity tasks with strict limitations on data sharing. Staff are explicitly cautioned about potential compliance risks when using externally-hosted platforms.

BSP On-Premises LLM: A locally-hosted solution providing secure, controlled AI capabilities. Made mandatory for all work involving confidential information and personal data, ensuring complete data sovereignty and security.

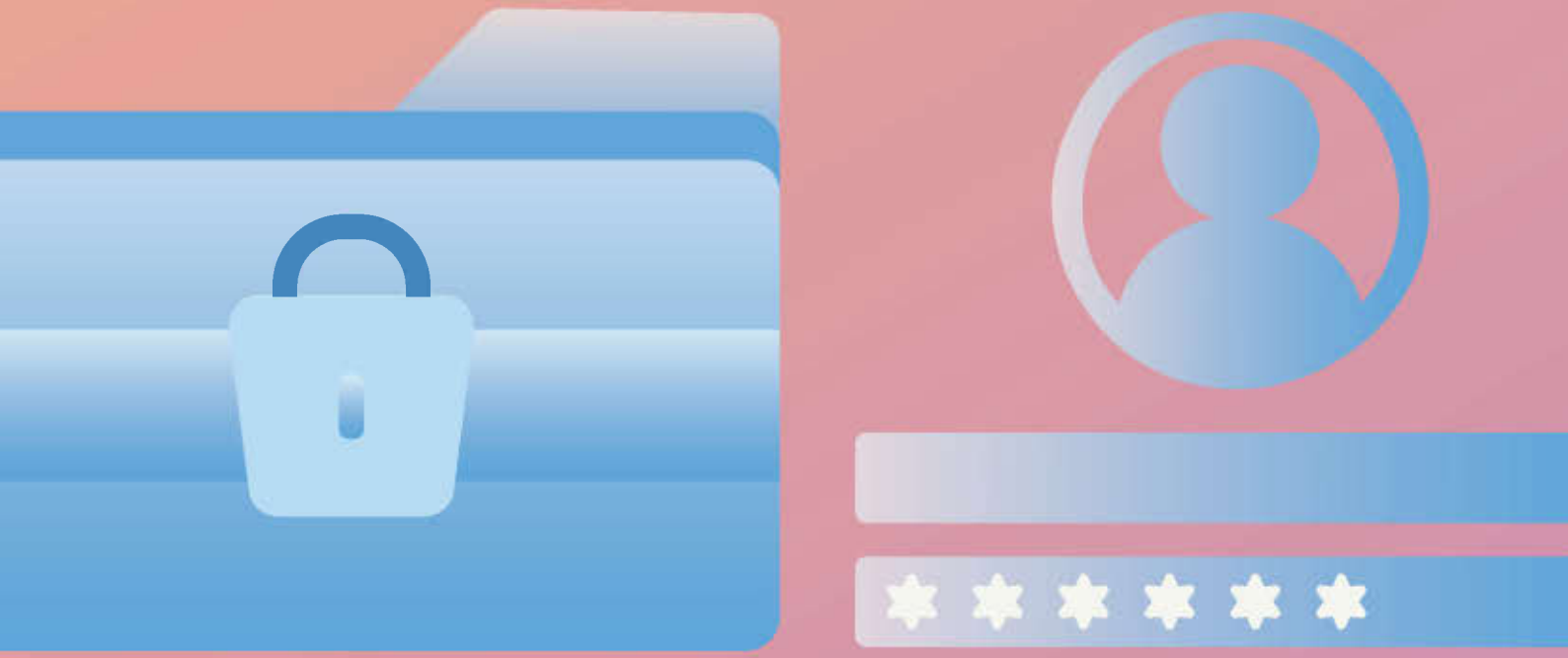


Key Governance Principles Applied

Data Protection & Data Governance: The guidelines enforce a clear classification system distinguishing between data suitable for public LLMs versus sensitive information requiring the on-premises solution, aligning with BSP's data protection obligations.

Transparency & Explainability: Staff are educated on how LLMs operate, their capabilities, and critically, their limitations, enabling informed decision making about when and how to use these tools.

Accountability & Integrity: The framework establishes clear acceptable use policies and best practices, including prompt crafting techniques, output verification requirements and explicit human oversight obligations to mitigate AI risks.



Human Centricity: Guidelines emphasize human judgment as paramount, requiring staff to verify AI-generated information and maintain oversight over all AI-assisted work, particularly in critical applications.

Fairness and Equity: Staff are trained to recognize potential biases in AI outputs and implement appropriate checks to ensure fair and equitable outcomes.

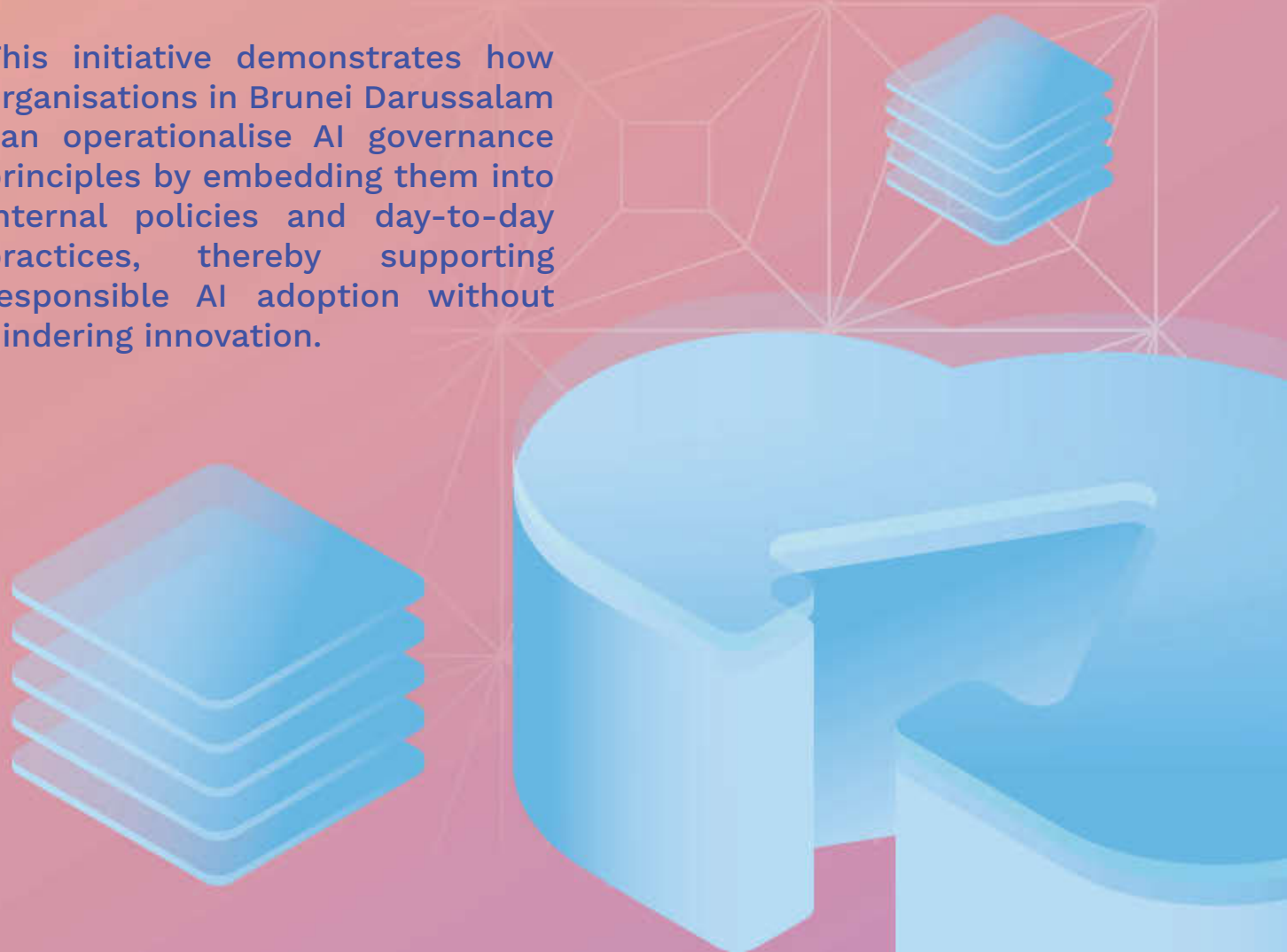
The guidelines enable BSP to harness AI innovation while maintaining its commitment to responsible technology use, data security, and compliance. By providing both accessible public tools and secure on-premises alternatives, BSP empowers employees to work efficiently while protecting sensitive information and upholding the organisation's values and regulatory obligations.

This framework demonstrates how organisations can proactively develop internal governance structures that align with national AI governance principles while addressing sector-specific requirements and risk profiles.

Responsible Use of AI within Organisational Policies by Perbadanan Tabung Amanah Islam Brunei

In December 2025, Perbadanan Tabung Amanah Islam Brunei (TAIB) introduced provisions on the responsible use of Artificial Intelligence (AI) within its Acceptable Use Policy (AUP). The policy establishes guidelines to ensure that AI technologies are used responsibly, securely, and ethically within the organisation, while protecting the confidentiality of sensitive information and ensuring compliance with applicable laws and organisational values.

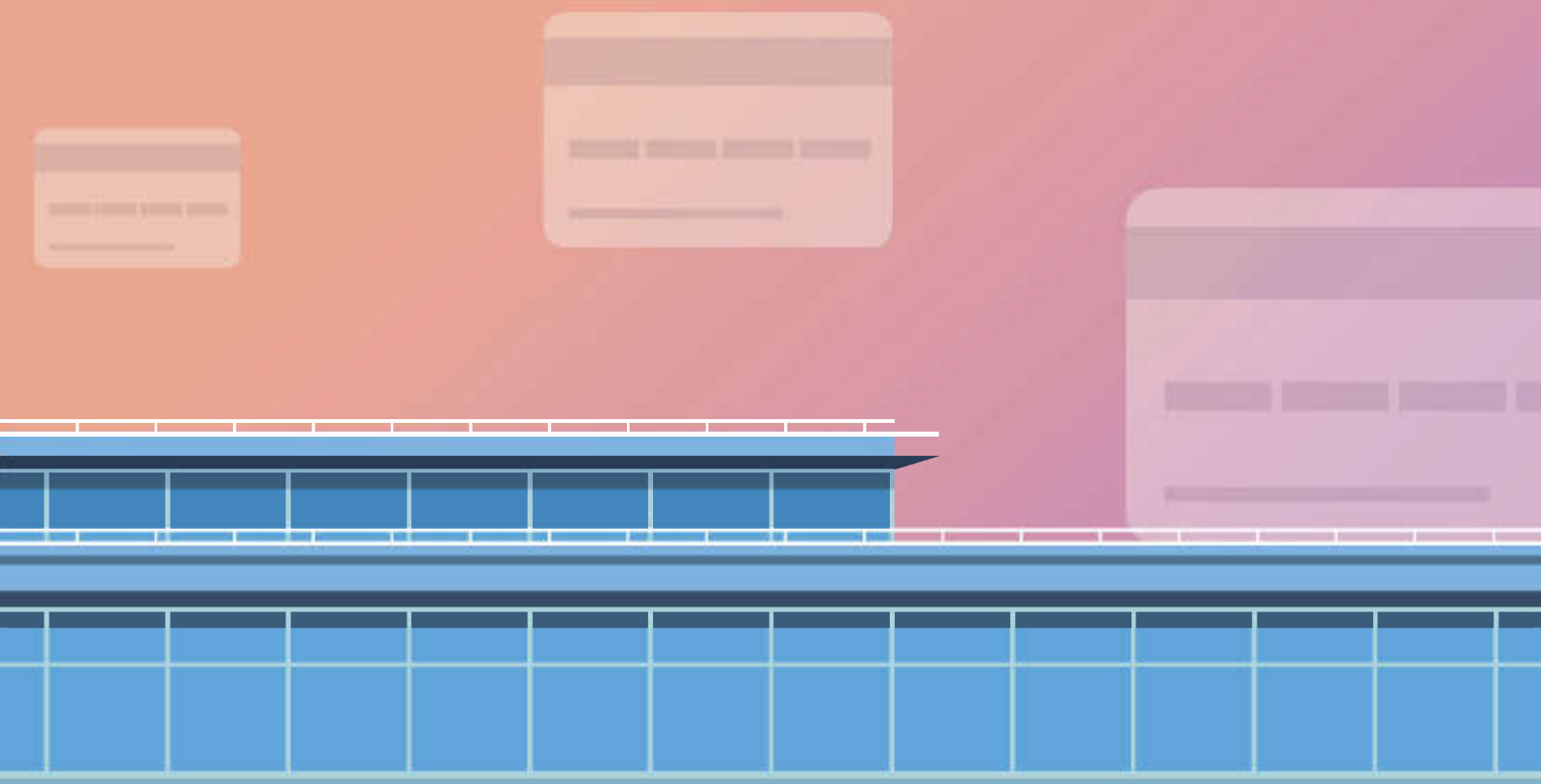
This initiative demonstrates how organisations in Brunei Darussalam can operationalise AI governance principles by embedding them into internal policies and day-to-day practices, thereby supporting responsible AI adoption without hindering innovation.



Emmi, Generative AI-powered assistant by Baiduri Bank

Emmi was introduced on Baiduri Bank's website in 2020 as part of its digital transformation journey. Initially built on Declarative AI, Emmi handled simple, structured customer queries using natural language processing (NLP), providing quick and accurate responses to common banking questions.

In 2025, Emmi was upgraded to a Generative AI-powered assistant, enabling it to understand complex, conversational queries and deliver more natural, context-aware responses. Despite its advanced capabilities, Emmi operates with a human-in-the-loop approach, with its knowledge base and behaviour created and monitored by Baiduri Bank, and it does not reference external articles or sources. When Emmi cannot interpret a customer's intent or question, it seamlessly hands off the conversation to a live agent, ensuring uninterrupted service and customer satisfaction.



To maintain ethical and secure AI practices, Emmi incorporates guardrails to detect and manage questions of malicious intent or sensitive nature, preventing inappropriate or harmful responses. As an added layer, filters based on local guidelines such as language awareness, cultural sensitivity are also implemented to ensure that responses are professional and culturally appropriate. Additionally, Emmi provides analysis and reports on customer interactions, creating a continuous feedback loop that helps Baiduri Bank better understand customer needs and refine its services.

Unified AI Governance Gateway for AI Solutions by EVYD Technology

EVYD Technology (EVYD) implements a unified Responsible AI Governance gateway across all its healthcare AI solutions, anchored on the following principles:

Safety & Accuracy:
Information-verification layers, safety filters controls to minimise misinformation and harmful outputs.

Accountability & Traceability:
Audit logs for data usage and model outputs, bias monitoring, fairness checks and continuous model-performance auditing.

Transparency & Trust:
Clear user disclosures on AI usage and culturally aligned communication which is consistent with national norms and MIB values.

Human Oversight:
Human- in- the- loop mechanisms ensuring clinicians or authorised personnel remain as the final decision-makers.

These governance measures ensure that EVYD's AI systems remain safe, reliable, ethical and aligned with Brunei Darussalam's AI governance principles within the Guide.



Cultivating a culture of responsible AI, where integrity meets innovation at Royal Brunei

Royal Brunei Airlines (RB) has begun its journey towards building a responsible and structured approach to Artificial Intelligence through employee surveys and the development of an AI Governance Framework. It envisions to enhance efficiency and innovation while upholding the values and principles of Melayu Islam Beraja (MIB).

In September 2025, RB conducted its first AI Readiness Survey to assess how AI is currently being used and perceived across different divisions within RB. This was followed by a second-layer survey aimed at understanding how employees apply AI in their daily work and identifying the types of support or guidance that they may need moving forward.

The findings from these surveys have guided the development of RB's AI Governance Framework. This framework aims to ensure that AI is used responsibly, ethically, and securely, while upholding RB's integrity, values, and commitment to service excellence. The AI Governance Framework will not only be about setting the controls within RB, it also aims to drive meaningful improvement and future readiness across five key areas such as:

(i) People;
(ii) Operations;
(iii) Technology;
(iv) Customer Experience;
(v) Safety and Security.



To maintain accountability, RB ensures there is appropriate review, approval and standardisation of any new AI tools through its own IT Steering Committee (ITSC). This ensures all AI initiatives remain aligned with the organisation's strategic priorities, governance requirements, and long-term digital roadmap.

AI Governance and Ethics Working Group

Members of AI Governance and Ethics Working Group are representatives from the following organisations:

1. Ministry of Transport and Infocommunications (MTIC)
2. Ministry of Health (MOH)
3. Ministry of Finance and Economy (MOFE)
4. Ministry of Education (MOE)
5. Ministry of Defence (MINDEF)
6. Ministry of Development (MOD)
7. Ministry of Culture, Youth and Sports (MCYS)
8. Dewan Bahasa dan Pustaka
9. E-Government National Centre (EGNC)
10. Cyber Security Brunei (CSB)
11. Authority for Infocommunications Technology Industry of Brunei Darussalam (AITI)
12. Brunei Darussalam Central Bank (BDCB)
13. Safety, Health and and Environment National Authority (SHENA)
14. Universiti Brunei Darussalam (UBD)
15. Universiti Teknologi Brunei (UTB)
16. Universiti Islam Sultan Sharif Ali (UNISSA)
17. Baiduri Bank
18. Bank Islam Brunei Darussalam (BIBD)
19. Brunei LNG Sendirian Berhad (BLNG)
20. Brunei Shell Petroleum Co. Sdn Bhd (BSP)
21. Datastream Digital Sdn Bhd (DST)
22. Dynamik Technologies Sdn Bhd (Dynamik)
23. EVYD Technology Sdn Bhd (EVYD)
24. Huawei Technologies (B) Sdn Bhd (Huawei)
25. Imagine Sdn Bhd (Imagine)
26. I.T. Protective Security Services Sdn Bhd (ITPSS)
27. Microsoft
28. Perbadanan Tabung Amanah Islam Brunei (TAIB)
29. Progresif Sdn Bhd (Progresif)
30. Royal Brunei Airlines Sdn Bhd (RB)
31. Unified National Networks Sdn Bhd (UNN)

COPYRIGHT NOTICE
© AITI, 2026. This document is the property of the Authority for Infocommunications Technology Industry of Brunei Darussalam ("AITI"), a body corporate with perpetual succession with its address at B13 and B14, Simpang 32-5, Jalan Berakas, Kampung Anggerek Desa, Brunei Darussalam. It must not be copied, used or reproduced for any other purpose other than for which it is supplied, without the expressed written consent of AITI.

DISCLAIMER
The information contained in this document does not constitute legal advice and should not be treated as such. AITI disclaims any responsibility or liability for any use or misuse of this document by any person and makes no representation or warranty, express or implied, as to the accuracy or sustainability of the information to any third party.