LTU AI Adoption Risk Assessment

LA TROBE UNIVERSITY

"A computer can never be held accountable therefore a computer must never make a management decision."

– IBM training manual 1979

This AI Risk Assessment Tool is intended to help the University identify and manage the risks associated with AI-related projects — whether developed internally or involving external vendors. It applies to:

- the procurement/acquisition of all new AI Systems/Tools, including teaching tools, research initiatives, administrative systems, and experimental prototypes; and
- existing, University approved AI System/Tools for which new use cases are proposed.

The result of this assessment is also intended to help inform whether your project can proceed, what level of risk mitigation is needed, and whether third-party providers meet our procurement requirements.

The Tool is intended to be used and completed by the relevant business area seeking to adopt/use an AI System/Tool, together with assistance from Information Services who have the expertise to assist with the technical AI and related questions. All terms defined in the Glossary (see page 14) are highlighted in **bold** throughout the document.

What is 'Artificial Intelligence':

Artificial Intelligence is an umbrella term which describes the theory and development of computer systems which can infer, from the input it receives, how to generate predictions, recommendations, content or decisions. AI uses **Algorithms** which are a set of instructions or mathematical formula, which guide the computer in generating an output.

AI **Algorithms** are incapable of thinking and do not replace human judgement.

Why assessing risk matters:

While AI **Algorithms** can help streamline processes, improve efficiency and support innovation, new and evolving technologies can also introduce potential risk and harm. AI systems can amplify inequality, undermine democracy (e.g. via misinformation and disinformation) and threaten both privacy and security. As AI tools incorporating **Algorithms** become increasingly sophisticated and commonplace – including **Generative AI** models such as CoPilot and ChatGPT– it is more important than ever that these technologies are approached with due care and diligence, and with appropriate levels of human oversight and clear human accountability. Furthermore, for sophisticated AI tools provided by external vendors our understanding of risks may be limited, and it is important that risk mitigation is considered as part of the acquisition/contracting process.

It is also important that such technologies are used in manner that advances the aims of the University, including an environment that is sustainable, fosters critical thinking, promotes equitable access to education and furthers human knowledge.

The University's approach to AI should also align with our Cultural Qualities:

- **Accountable:** we apply clear lines of responsibility, accountability and instil transparency in our AI use.
- **Connected:** we engage across disciplines and functions, using cross-University collaboration to identify, manage and monitor issues and risks.
- **Innovative:** we apply a risk-based approach, classifying AI according to its level of risk and applying corresponding levels of oversight, review and controls to mitigate that risk.
- **Care:** we actively consider environmental sustainability and the impact of unwanted bias, data security and potential benefits in the use of AI. Although they may be supported by AI applications, significant decisions will be made by humans to promote a safe and inclusive community.

It is important that La Trobe University's adoption of AI aligns with its core purpose, community expectations and Australian regulatory standards, for example:

- **Australian Government:** This process is intended to align with the Voluntary AI Safety Standards released by the Australian Government in 2024, which outline key expectations for transparency, accountability, fairness, privacy, and human oversight.

- **Higher Education Compliance:** As a registered higher education provider, we must comply with the Tertiary Education Quality and Standards Agency (TEQSA) Higher Education Standards Framework 2021, including obligations related to academic integrity, data governance, and student wellbeing. In particular:
    - **Standard 1.4 – Ensuring integrity of learning and assessment** - Institutions must ensure that assessment methods are fair and effective. The use of AI tools that manipulate or deceive students could compromise the integrity of assessments.
    - **Standard 2.2 – Supporting equity and inclusion** - Providers are required to promote equity and diversity. AI systems that exploit vulnerabilities or discriminate against certain groups would be inconsistent with this standard.
    - **Standard 3.3 – Appropriateness of learning resources** - Learning resources, including AI tools, must be appropriate and support student learning. Tools that employ manipulative techniques could hinder learning and are therefore inappropriate.
    - **Standard 4.1 – Ethical research practices** - Research must be conducted ethically. The deployment of AI in research that manipulates participants or collects data without proper consent would violate this standard.
- **University Policy:** As noted in the University's Responsible AI Adoption Policy:
    - the University will adopt a risk-based approach to ensure that obligations and oversight are proportionate to the risk posed; and
    - an AI Risk Assessment must be completed for:
        - the procurement/acquisition of all new AI Systems/Tools, including teaching tools, research initiatives, administrative systems, and experimental prototypes; and
        - existing, University approved AI System/Tools for which new use cases are proposed.
- **Emerging Regulatory and Ethical Norms:** This assessment is also intended to reflect evolving international best practices, such as the EU AI Act, the OECD AI Principles, and risk management standards like ISO/IEC 23894:2023 (AI Risk Management).

What AI cannot be used for = Very High Risk use cases

There will be some AI use cases where the risks (whether legal, ethical, reputational) are too great that they outweigh any benefit(s) that may be derived.

As set out in the University's Responsible Adoption of AI Policy, there are some AI Systems/use cases that will generally be considered Very High Risk and if the risks cannot be adequately controlled to bring the risk level down, they will not be permitted at the University. Please see the **Responsible Adoption of AI Policy** for further information and examples.

How This Tool Works

Step 1

Information Services and Business Area to complete Sections 1 and Sections 2.

Step 2

After considering the information set out in Sections 1 and 2, Information Services will undertake a preliminary assessment as to the level of risk that the AI System and proposed use case may present.

Step 3

If the level of risk is determined to be very low or low, including consideration of relevant controls to be applied, then the project may be approved by the Chief Information Officer.

Step 4

If the level of risk is anything medium or higher, a more in-depth risk assessment questionnaire (Section 3) will need to be completed and the matter will be referred to the AI Ethics Committee for their advice.

Step 5

The AI Ethics Committee will consider the matter and, if endorsed, will refer the project to the Responsible Adoption of AI Committee (Chaired by the VC) for approval.

Step 6

The procurement process (if applicable) is then undertaken and, if minimum expected or recommended risk mitigation measures cannot be achieved with a vendor, the matter may be referred back to relevant Approver for further consideration.

Section 1: Project Information & Accountability

Preliminary Assessment - Acccountability

The following questions are intended to:

- clarify accountability for the project and ongoing review of the AI System; and
- help Information Services conduct a preliminary assessment to determine whether further information is required and/or the consideration by the University's Responsible Adoption of AI Committee and/or its subcommittees.

Stakeholder Details

Project Name	
Author	Person completing this form (generally a staff member from IS in consultation with the relevant business area and other stakeholders).
Project Sponsor & Accountable officer	LTU staff member ultimately accountable for the project and who will generally be a SEG member.
Project Manager (if applicable)	The person responsible for ensuring the Project is compliant from a legal compliance perspective and that all internal processes are followed.
Business/System Owner	The senior manager responsible for the AI System on an ongoing basis.
System Administrator	The person who manages access to an IT platform/system on an ongoing basis.
AI Technical Lead	The individual responsible for developing, customising, or technically implementing the Algorithm or AI System. Their input may be needed to complete technical sections of the assessment.

Project Summary

Add a brief project description/AI use case: In **Plain Language**, describe: • The problem the project/use case is addressing • How the AI System will assist • Intended value proposition, and key benefits • Expected project start and end dates • Any key milestones planned to be met	
Are there any alternative, non-AI tools or approaches that could achieve similar results?	◯ Yes ◯ No ◯ Unsure
What type of AI System is being proposed or used? See Glossary for definitions.	◯ Predictive analytics ◯ **Natural Language Processing** (NLP) ◯ **Large Language Model** (LLM) ◯ **Automated Decision-making** ◯ **Generative AI** ◯ Recommender system ◯ Computer vision ◯ Other (specify): ◯ Unsure (please describe if possible):

Who will use this AI System?	◯ Students ◯ Academic staff ◯ Professional/admin staff ◯ General public ◯ Research participants ◯ Other (specify):
What other purposes could this AI System or functionality be used for (either internally or externally, including by individuals looking to do harm)	Please explain:
Who could be impacted by the use of this AI System?	◯ Students ◯ Academic staff ◯ Professional/admin staff ◯ General public ◯ Research participants ◯ Other (specify):
Which of the following lifecycle stages best describes the current stage of your AI use case? *Assessments must be reviewed when use cases either move to a different stage of their lifecycle or significant changes occur to the scope, function or operational context of the use case. If in doubt, contact IS*	◯ Early experimentation ◯ Design, data and models ◯ Verification and validation ◯ Deployment and testing ◯ Operation and monitoring ◯ Retirement
Preliminary Procurement Questions	
Has the AI System been acquired but not yet deployed? e.g. as part of software for which the University already has a licence.	◯ Yes ◯ No ◯ Unsure
Will the AI System be developed or provided by a third-party vendor?	◯ Yes - if yes, are there identified or preferred providers: ◯ No - please provide further information (e.g. note the existing University AI System that will be used or, if one is to be developed, what is proposed)?

Best/Worst Case Scenarios

Describe the best-case scenario(s) that could realistically arise from using the algorithm and the potential benefit(s)

Who benefits and how? For instance, will it: • deliver a better-quality existing service or outcome) and, if so, to who and how? • reduce processing or delivery times and, if so, to who and how? • generate financial efficiencies or savings and, if so, to who and how? • deliver a new service or outcome (particularly it cannot be done without AI) and, if so, to who and how? • enable future innovations to existing services or new services or outcomes and, and, if so, to who and how? *For each benefit, state how likely it is that the benefit will be realised in practice.*	Potential benefit description Please add extra benefits as required **Benefit 1:** **Benefit 2:**	Likelihood **Benefit 1:** ◯ Very unlikely or N/A ◯ Unlikely ◯ Mid-range 50/50 ◯ Yes ◯ Likely ◯ Very likely **Benefit 2:** ◯ Very unlikely or N/A ◯ Unlikely ◯ Mid-range 50/50 ◯ Yes ◯ Likely ◯ Very likely
What are the likely challenges or hurdles to achieving the best-case scenario?		
Describe the worst-case scenario(s) that could realistically arise from using this AI System, and the potential consequences/harm(s)		
Who might miss out or be disadvantaged in this scenario? For instance, consider the risks of: • typical AI errors (e.g.: bias, hallucinations, inaccuracies etc) and who may be disadvantaged and how • unauthorised use of health or **Sensitive Personal Information**, who might be impacted and how • environmental harms or harms to the broader community, who may be impacted and how • Potential for bias in the system, the impact of a right, privilege or entitlement; who is likely to be impacted by this and how? • Reputational risk (e.g. if something goes wrong), who may be impacted and how	Potential disadvantage description Please add extra disadvantages as required **Disadvantage 1:** **Disadvantage 2:**	Likelihood **Disadvantage 1:** ◯ Very unlikely or N/A ◯ Unlikely ◯ Mid-range 50/50 ◯ Yes ◯ Likely ◯ Very likely **Disadvantage 2:** ◯ Very unlikely or N/A ◯ Unlikely ◯ Mid-range 50/50 ◯ Yes ◯ Likely ◯ Very likely
Who at LTU is likely to be held accountable to rectify and solve the consequences/harm?		
What is the likely process to rectify the consequences/harm?		

Section 2: Preliminary Risk Screening Questions

Preliminary Risk Screening Questions	
The following questions are intended to determine whether: • a more in-depth Risk Assessment Questionnaire is required for higher-risk initiatives to fully identify potential impacts and controls.	
Will the AI System be used for or as part of any **Automated Decision Making**?	◯ Yes ◯ No If yes, please explain in further detail:
Will the AI System/Tool be used (or is it capable of being used) to automate, aid, replace or inform any decisions that are likely to have a **Material Impact** on any individuals, communities or other groups, including Aboriginal or Torres Strait Islanders?	◯ Yes ◯ No If yes, please explain in further detail:
Will this project/tool will use or generate **Personal Financial Information**, **Sensitive**, **Health** or **Children's Information** as part of generating the output?	◯ Yes ◯ No
Is the AI System being used as part of a research project?	◯ Yes ◯ No If yes, has approval been provided by one of the Research Ethics Committees? Ethics approval reference:
The use of AI be hidden from the users/potentially impacted individuals.	◯ Yes ◯ No If no, please explain how **Users** or potentially impacted individuals will be made aware that AI is being used?:
The use of AI be hidden from the users/potentially impacted individuals.	◯ Yes ◯ No
Has the IS Architecture/Security team been consulted so that a detailed architecture security assessment of the AI model can be undertaken?	◯ Yes ◯ No **IF YES:**
How has the IS Architecture/Security team assessed the AI System?	Please attach/add IS Architecture/Security advice/analysis

Will a third party be able to access or use the Production or Output Data?	◯ Yes ◯ No
Can the service/outcome be delivered as efficiently/effectively without the use of AI? *NB – there is an environmental cost as well as other risks associated with the use of AI.*	◯ Yes ◯ No ◯ Unsure
Confirm that the project does not involve a very high risk (i.e. prohibited) System/use case *See the section 'What AI cannot be used' above for further information about prohibited tools/use cases. If the project involves a prohibited use/ tool – please speak to IS and do not proceed further.*	◯ Yes , I confirm that the project does not involve a very high risk (i.e. prohibited) AI System/use case: ◯ Unsure

IMPORTANT **Preliminary Risk Assessment – Following completion of the above, Information Services should be able to undertake a Preliminary Risk Assessment and determine the overall risks.**

Section 3: In-depth Risk Assessment Questionnaire

Elevated Risk Screening Questions

The following screening questions are intended to help fully assess whether the proposed AI tool or project poses an elevated ethical or societal risk to the university.

Date	
PRIVACY – The university must preserve public trust by ensuring secure, transparent and data use that is compliant with privacy law obligations.	
Will personal information be used or generated in the Project/AI System. If yes, please attach/link to a copy of the Privacy Impact Checklist (PIC).	◯ Yes ◯ No (Please proceed to security questions) ◯ If yes, is PIC attached? **IF YES:**
What type of **Personal Information** will be used or generated as part of the **Output**?	◯ **Personal Information** ◯ **Sensitive Personal** information ◯ **Health Information** ◯ Children's information ◯ Financial information ◯ **Biometric Information/Templates**
Will this project/tool will use or generate **Aboriginal and Torres Strait Islander data** as part of generating the output?	◯ Yes ◯ No
Will the tool generate an opinion or infer information about an individual?	◯ Yes ◯ No
SECURITY – The university must preserve public trust by ensuring secure data.	
Will this AI System interact with any other existing LTU systems?	◯ Yes ◯ No If yes, which LTU systems:
Will the AI Model be trained on LTU data?	◯ Yes ◯ No ◯ Unsure **IF YES OR UNSURE:**
Will the AI Model be used by third parties?	◯ Yes ◯ No
Does the system provide the ability to host the AI model in an LTU data centre?	◯ Yes ◯ No **FOLLOWING QUESTIONS MANDATORY:**
Is the AI Model hosted in Australia?	◯ Yes ◯ No
Where is the AI Model hosted?	Please state:

What records will be generated? *Consider the nature of the Outputs and who will be able to access them.*		
When and how will the records be deleted? *Please refer to the Records Management Policy and Records Team for advice on records storage, relevant retention periods and record destruction.*		
How will the **training data**, **production Data** and **output data** be stored and secured? *NB – include information regarding location, including county where it will be stored, security arrangements, access controls and data retention period.*	**Training Data:** **Production Data:** **Output Data:**	
FAIRNESS - Use of ai must be: fair, not perpetuate bias or inequality by using non-representative/poor data, have its performance monitored/overseen via human intervention and use rigorous data governance.		
Will the project/tool be used to automate, aid, replace or inform any operational or policy decisions that may have a **Material Impact** on any individuals?	◯ Yes ◯ No **IF YES:**	
Is the **Algorithm Training Data** materially different to the **Production Data**?	◯ Yes ◯ No ◯ Unsure	
	Training Data	**Production Data**
Describe the data used.	Describe the **Training Data** used to create the underlying **Algorithm**(s) used in the **AI model** (if known):	Describe the **Production Data** used (if known) to create the **Output Data**
What is source of the data and how was it obtained? *NB - For each data type, describe: the data source (including country) who originally collected it, how and why?*	Describe the data source (if known):	Describe the data source:
Was each data type obtained legally (e.g. with an individuals' consent for this purpose)?	◯ Yes ◯ No ◯ Unsure	◯ Yes ◯ No ◯ Unsure

Is the data relevant, sufficient and representative? When answering this question, consider 1. What and who is represented in the data? Do the data sets accurately represent the populations to which the **Algorithm** will be applied? 2. When was the data collected and for how long will it be relevant and accurate? 3. Is there enough of it to achieve the purpose?	◯ Yes ◯ No If no, what steps are proposed to mitigate risks with over/under representation:	◯ Yes ◯ No If no, what steps are proposed to mitigate risks with over/under representation:
Is the data sufficiently accurate and reliable to support the purpose?	◯ Yes ◯ No ◯ Unsure	◯ Yes ◯ No ◯ Unsure
Will proxy variables (i.e. synthetic data): be used? *If yes, what is the rationale for using them, their potential limitations and how they might affect the **Algorithm's** outputs. What steps have been taken to mitigate potential bias risks associated with them?*	◯ Yes ◯ No ◯ Unsure If yes, please provide further information:	Not applicable
Describe measures to ensure data and **Algorithm** reliability, integrity and security, including: • access controls, • monitoring; and • auditing.		
Describe how potential bias, discrimination and/or other unfair, unintended or unexpected outcomes (together "unfair outcomes") could be present in the data, the **Algorithm** and the wider AI System/project. For each unfair outcome, describe the nature of the harm, who might be harmed, how, when and why		
Are there performance testing mechanisms in place for the lifecycle of the AI System? What is the frequency of the testing/ review and who is responsible?	◯ Yes ◯ No ◯ Unsure	
Estimate the extent to which Users might rely on the recommendation of the **Output Data** instead of their own judgement? How will this risk be mitigated/ addressed?	◯ 1 (Very unlikely) ◯ 2 ◯ 3 ◯ 4 ◯ 5 (Very likely) Describe proposed mitigation:	

Will a user/impacted individual have the ability to view and/or challenge the output.	◯ Yes ◯ No Please explain:
Has or will guidance/training be provided to help Users understand how to responsibly use the AI Tool/ Project, including in relation to checking the accuracy of Output Data and minimising risks.	◯ Yes ◯ No If yes, please briefly describe:
Has or will guidance/training be provided to help Users understand how to responsibly use the AI Tool/ Project, including in relation to checking the accuracy of Output Data and minimising risks.	◯ Yes ◯ No If yes, please briefly describe:
Will **Aboriginal and Torres Strait Islander Data** be used in the **Training Data** or Production Data?	◯ Yes ◯ No (If no, proceed to *Transparency* section) **IF YES:**
Were Aboriginal and Torres Strait Islander people/voices involved in the co-design of the **algorithm**.	◯ Yes If yes, what issues (if any) were raised and how will they be addressed? ◯ No If not, why not?:
How will principles of **Aboriginal and Torres Strait Islander data sovereignty** (i.e. the right to own, control, access and possess **Aboriginal and Torres Strait Islander Data**) be implemented/maintained?	
TRANSPARENCY – The use of ai must be transparent, freedom of information act compliant, secure and ethical, and allow concerns to be raised and addressed.	
Is there a communication plan in relation to those users/potentially impacted that includes: a. Transparency about when a user is interacting with AI, not a person b. How and why AI is being used, what decisions are being made by it, who it will impact and how c. What **Personal Information** is involved d. How Training and Production Data has been collected, secured and stored e. Specifying explicitly if any **Automated Decision Making** exists in this project.	◯ Yes ◯ No

Is it possible to explain in layperson's terms how the **Algorithm** used to train the **AI Model** produces its outputs, including the logic behind its design and how it makes any decisions?	◯ Yes ◯ No
ACCOUNTABILITY – Identifiable university officers remain accountable for decision-making in relation to the use case and application of ai	
Who will be responsible for the AI Outputs, including monitoring the performance of the system, appeal/redress process?	
How will errors be communicated to those impacted and more widely and who will be responsible?	
COMMUNITY BENEFIT – The pros of using ai to the community must outweigh the cons. Any use of ai must also prioritise benefits to the broader community, minimise harm and maximise benefit.	
Is this AI system delivering a new or existing service?	◯ New ◯ Existing
Have those who are most likely impacted been consulted?	◯ Yes ◯ No If yes – provide details:
Have you considered the environmental impacts, including energy/ Green House Gas emissions, water use and/or land use and resources?	◯ Yes ◯ No If yes – provide details:
Is there sufficient human oversight of the AI system?	Please describe:
What processes have you got in place to measure the environmental impacts, including energy/Green House Gas emissions, water use and/ or land use and resources?	
What steps will you take to reduce the environmental impact, and how will you measure their effectiveness?	

Glossary

Aboriginal and Torres Strait Islander Data – digital or digitisable data information or knowledge that is about, form or connected to Aboriginal and Torres Strait Islander people. It includes data about population, place, culture and environment.

Aboriginal and Torres Strait Islander Data Sovereignty – the inherent rights and interests that Aboriginal and Torres Strait Islander people have in relation to the collection, ownership and application of Aboriginal and Torres Strait Islander Data.

Algorithm - A set of instructions/formula applied to a dataset that guide a computer in performing specific tasks or s olving problems.[1] Algorithms are used to train/create **AI Models** (they are essentially the underlying logic by which the **AI Model** operates).

Artificial Intelligence or Artificial Intelligence System – A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

Artificial Intelligence Model - An **AI Model** is the output of an algorithm or algorithms that has been applied to a data set. An **AI Model** is a core component of an AI system used to make inferences from inputs to produce outputs. Prior to deployment, an AI System is typically built by combining one or more **AI Model** developed manually or automatically (e.g., with reasoning and decision-making algorithms) based on machine and/or human inputs/data.

Automated Decision Making - Automated decision-making (ADM) in AI refers to the use of computer systems, algorithms, and **AI Model** to make decisions or perform actions with minimal or no human intervention.

Biometric Information - Biometric characteristics can be physiological features (for example, a person's fingerprint, iris, face or hand geometry), or behavioural attributes (such as a person's gait, signature, or keystroke pattern) [OVIC]

Biometric Template - Is a digital template where key features of the biometric characteristic are extracted and processed to create the template, which is stored in a database for future use (e.g. for comparing biometric information to the template at a later date). [OVIC]

Computer Vision - is a field of AI that trains computers to interpret and comprehend the visual world, analysing images and videos. Computer Vision algorithms analyse images and videos for tasks like object detection, face recognition, and self-driving cars.

Foundation Model - Foundation models, sometimes called general purpose AI Systems, provide a basis for future modification and fine-tuning for specific tasks. They are trained on vast amounts of data at scale, including images, text, audio, video and other data types like 3D models, with less emphasis on data quality so that they can be adapted to a wide range of downstream tasks.

Generative AI - Generative AI, often referred to as Gen AI, is an emerging field within AI that creates new content such as text, images, voice, video, and code by learning from data patterns. Examples include, Microsoft CoPilot and ChatGPT.

Health Information - health information has the meaning set out in the Health Records Act 2001 (Vic). In summary, health information is **Personal Information**: about the physical, mental or psychological health or disability of an individual; about an individual's expressed wishes regarding the future provision of health services to them; about a health service provided, or to be provided, to an individual; collected to provide a health service; about an individual collected in connection with organ or body substance donation; or that is genetic information in a form which is or could be predictive of the health of the individual or of their descendants.

Large Language Model LLMs are a subset of **Gen AI** model that specialises in generating synthetic text. LLMs are trained on huge sets of data, often from the internet. LLMs are built on machine learning.

Material Impact – in this context, it means something that could reasonably be expected to affect the rights, opportunities or access to critical resources or services of individuals, communities or other groups in a real and potentially negative or harmful way, or that could similarly influence a decision-making process with public effect. For example, decisions that may impact a person's access to education, employment etc.

Machine Learning (ML) is a subset of AI that involves the development of algorithms and statistical models that are intended to enable an AI system to improve over time without being explicitly programmed.

Machine Learning Algorithm: is an algorithm used to determine parameters of a Machine Learning Model from data according to given criteria. ML algorithms are trained on data to make predictions or decisions. For example, machine learning algorithms influence consumer buying decisions by making recommendations based on their purchase history

1. ISO (International Organization for Standardization), AS ISO/IEC 42001:2023, Information technology – Artificial intelligence – Management system, ISO 2023

Machine Learning Model: mathematical construct trained with an algorithm to generate an inference or prediction based on input data or information.

- Supervised MLM involves making predictions through a constant or stable variable. Algorithms in this format can take known data and its subsequent answers to set parameters for predictions within new datasets. This method allows for accurate predictions of new data because of the constructive parameters assembled in previously studied data.
- Unsupervised MLM – The unsupervised learning method involves rough exploration of data to create a foundational understanding of patterns and constructs that are intrinsically buried in the data. In order to draw insights and inferences out of data with no predetermined parameters, such as in the supervised learning method, this algorithm explores datasets with the goal of discovering discrete or even hidden patterns. This method is often used across industries and a popular choice for market research and studies of genetic sequencing.

Natural Language Processing Natural language processing (NLP) is a field of **Artificial Intelligence** (AI) that deals with the ability of computer systems to understand and generate human language. NLP algorithms are used to analyse text, comprehend, converse with users and perform tasks like language translation, sentiment analysis, and question answering.

Output Data – is the new data an AI tool creates based on Production Data.

Personal Information – Personal Information is any information or an opinion about an individual who is identifiable or whose identity may be reasonably ascertainable and includes photos [PDPA].

Predictive Analytics involves using statistical analysis and machine (ML) to identify patterns, anticipate behaviours and forecast upcoming events. Organisations use predictive analytic AI to predict potential future outcomes, causation, risk exposure and more.

Sensitive Information – Sensitive Personal Information is a subset of **Personal Information** and includes personal information that is about a person's: racial or ethnic origin; political opinions; membership of a political association; religious beliefs or affiliations; philosophical beliefs; membership of a professional or trade association; membership of a trade union; sexual preferences or practices; or criminal record [PDPA].

Training Data – refers to the dataset used to teach a model by identifying patterns and relationships within it

Production Data – also referred to as 'input data,' refers to the real-world data that the trained model is applied to once deployed (e.g. the data a user inputs to generate an output).

Users – means any natural or legal person, public authority, agency or other body using an AI system.

Information Services to Complete

Section 4: AI Risk Findings and Mitigation Strategies

Part A – Detailed Risk Assessment

This section provides further detail of the potential risks and controls identified for the Project and allocates a risk rating for each risk before the application of possible controls (the "Inherent Risk") and again following consideration of the application of the proposed/additional controls (the "Residual Risk"). A sample entry is provided by way of example only.

Project Name		Date Risk Assessment Completed	

Risk Domain	Risk Title & Description	Causes	Consequence	Inherent Risk	Proposed / Additional Controls	Residual Risk
				Likelihood / consequence		Likelihood / consequence
For example: (Fairness.)	For example: (An algorithm to be used in an AI System as part of the recruitment process to screen for preferred candidates was trained on data about the agency's current and previous employees, creating a risk that it could learn to prioritise similar types of people, excluding those of a different ethnicity, gender or physical ability.)	For example: (The use of training data from current and previous employees only, which risks reproducing historic / societal biases in the context of employment offers.) (Lack of clear governance and human oversight of the algorithm, including little to no bias or "fairness" testing and insufficient monitoring of bias over the algorithm's lifecycle.)	For example: (Discrimination, Exclusion from employment andLoss of opportunity.)	For example: (Very high, Almost, Certain and Major.)	For example: (Re-train the model using appropriate, sufficiently representative training data.) (Implement appropriate testing and review mechanisms across the algorithm lifecycle to check for potential biased results.) (Adopt processes outside the AI System to encourage a broad cross-section of job seekers, such as making sure job descriptions are appropriately worded to reduce unconscious bias.)	For example: (Medium, Possible and Moderate.)

Part B – Overall Risk Profile

Overall, we identified *(Insert number)* potential harms in the Project. The following Residual Risk ratings are based on the risk assessment criteria in Part C

- Very High *(Insert number)*
- High *(Insert number)*
- Medium *(Insert number)*
- Low *(Insert number)*
- Very Low *(Insert number)*

Those potential harms set out in detail above are summarised below.

The following is a summary of potential risks to the University:

- ◯ Negative impact on service delivery or operational performance
- ◯ Significant negative media coverage
- ◯ Compliance breaches
- ◯ Regulatory involvement or intervention, including corrective actions or penalties for non-compliance
- ◯ Legal costs
- ◯ Reputation damage – the University
- ◯ Reputation damage – Government
- ◯ Loss of public confidence and social licence
- ◯ Possible independent investigation commissioned
- ◯ Major recovery action required to restore credibility
- ◯ Financial repercussions

Part C – Understanding the Risk Levels – Risk Classification Guide

Table 1 – Harm assessment matrix

		Consequences				
		Minimal	**Minor**	**Moderate**	**Significant**	**Catastrophic**
Likelihood	**Almost Certain**	Low	Medium	High	Very High	Very High
	Likely	Low	Medium	Medium	High	Very High
	Possible	Very Low	Medium	Medium	Medium	High
	Unlikely	Very Low	Low	Medium	Medium	Medium
	Rare	Very Low	Very Low	Very Low	Low	Low

Table 2 – Likelihood rating & criteria

Likelihood of occurrence	**Criteria:** probability in 12 month period
Almost Certain	> 80%
Likely	60 - 80%
Possible	40 - 60%
Unlikely	20 - 40%
Rare	< 20%

Table 3 – Impact rating & criteria

Consequence of harm – if AI systems or applications malfunction or produce unintended outcomes (non-exhaustive & by way of illustration only)	
Catastrophic	**Catastrophic, irreversible consequences for individuals, societies, or the environment, for example:** • excluded from accessing public services; • significant breach of human rights; • significant loss of rights or freedoms; • significant negative impacts on specific groups; • financial distress; • significant loss of agency or autonomy; • serious or long-term psychological or physical damage, • death (including suicide); • significant public health implications.
Significant	**Significant consequences and considerable damage even if able to corrected/reversed:** • inability to access services; • serious breach of human rights; loss of rights or freedoms; • material negative impacts on specific groups; financial loss; loss of employment; • loss of agency or autonomy; • negative impacts on mental and/or physical health; • public health implications
Moderate	**Significant inconvenience, moderate disruption/harm even when able to be corrected/reversed:** • difficulty accessing services; breach of charter rights or anti-discrimination legislation; • negative impacts on rights or freedoms; • disadvantages to specific groups; extra costs; • employment changes; • anxiety, confusion, • stress; minor physical ailments
Minor	**Inconvenience, though can be easily corrected/reversed without harm/damage:** • harder to access services; • difficulty asserting rights or freedoms; • inconvenience, • frustration or irritation;
Minimal	**Minor inconveniences or no impact:** • The potential consequences of an AI system in this category are either non-existent or so insignificant that they can be safely disregarded

Table 4 – Harm severity & implications

Catastrophic	Highest priority to mitigate; this level of harm cannot be accepted.
High	High priority; this level of harm cannot be accepted without appropriate controls and a formal risk acceptance process.
Medium	Moderate priority; this level of harm cannot be accepted without appropriate controls and some form of risk acceptance process.
Low	Low priority; this level of harm can be accepted.
Very Low	Very low priority; this level of harm can be accepted.

Table 5 – Risk escalation & reporting

This table sets out who must be informed and has authority to accept risk based on its severity.

Very High	Not permitted
High 3	Responsible Adoption of AI Committee
Medium 2	Responsible Adoption of AI Committee
Low 1	Chief Information Officer
Very Low	Chief Information Officer

Table 6 – Risk Domains

This table sets out the relevant risk domain.

Risk Domains
> 80%
60 - 80%
40 - 60%
20 - 40%
< 20%

Section 5: Assessment and Approvals

Author sign-off:

I confirm this AI Project Risk Assessment has been prepared as a result of input from an appropriate multi-disciplinary group involved in the AIA process for this Project and that it accurately represents the identified potential harms and risks and the existing and/or proposed controls considered sufficient to mitigate them.			
Name and Position:			
Signature:		**Date:**	

Reviewer / decision maker:

I confirm that I understand the potential harms and risks of the Project and the extent to which the proposed controls will be sufficient to mitigate them.			
Assessment outcome:	*Approve Or Unable to Approve Or Approve Subject to Conditions – Please Detail*		
Name and Position:			
Signature:		**Date:**	

Executive Sponsor:

As the person accountable for this Project, I confirm I understand the potential harms and risks and the sufficiency of the proposed controls. I (Do / do not) support the Project proceeding as proposed in this document.			
Name and Position:			
Comments:			
Signature:		**Date:**	

Section 6: Procurement Focused Questions

Externally Acquired AI Products & Systems

Governance committees are groups within our organisation that focus on ensuring good governance practices, primarily by overseeing the performance, effectiveness and compliance to existing processes, policies and ways of working; and who may develop or oversee new technologies or ways of working.

Purpose/benefits	
What specific benefits or outcomes is the AI System designed to deliver?	
Can the vendor provide both qualitative and quantitative analyses in respect of the benefits?	
What metrics are used to measure success? What trade-offs were made, and how are outliers handled?	
Data Relevance and Data Governance	
How was the AI Model or Models developed? What approach was chosen, and why?	
Can the vendor explain the System's logic and share details about training and validation?	
What **Training Data** was used? Describe: • the data source (including country) • who originally collected it? • how was it obtained? • was it obtained legally? • why was it selected?	
Is the data sufficiently accurate and reliable to support the Purpose? When answering this question, consider: • When was the Data collected and for how long will it be relevant and accurate? • Is there enough of it to achieve the Purpose? • What and who is represented in the Data? Do the data sets accurately represent the populations to which the Algorithm will be applied? • How will risks associated with over or under representation be addressed?	◯ Yes ◯ No
What data governance and management practices does the vendor have in place?	

Describe how inaccuracies in the data will be identified and addressed and data quality will be maintained on an ongoing basis.	
Describe potential limitations or biases in the Training Data.	
Will the vendor have access to the Production Data or the Output Data?	◯ Yes ◯ No If yes, describe why that access required and how such access will be managed?
Risk Management	
What potential harms or risks could arise? Are there mitigation strategies and plans for continuous monitoring?	
What measures ensure robustness, privacy, and protection against attacks? Has third-party code or APIs been evaluated for quality?	
Does the AI System technically allow for the automatic recording of events ('logs') over the lifetime of the AI System.	
Technical Documentation and Instructions for Use	
Is user training provided?	◯ Yes ◯ No
What fallback mechanisms are in place for unexpected issues?	

Date procurement questions completed	
Author	