Artificial intelligence governance additional definitions

Intent:
The intent of this policy is to establish a notion of appropriate use of Artificial Intelligence (AI) systems within the College; and to address security, privacy, trust, law and regulatory, ethical and social impacts of AI on the College and its stakeholders.

Plain Text Objectives of Artificial Intelligence Governance Policy:
The following objectives are plain text definitions of the objectives listed in the Artificial Intelligence Governance Policy. This includes *current and future technology or techniques* used to access AI systems on College Information Technology Resources (ITR).

Large Language Model (LLM)
A computational model notable for its ability to achieve general-purpose language generation and other natural language processing tasks. LLM's acquire their capabilities by learning statistical relationships from vast amounts of text data during a computationally intensive self-supervised and semi-supervised training processes.

Risk Classifications:

Unacceptable

- deploying **subliminal, manipulative, or deceptive techniques** to distort behaviour and impair informed decision-making, causing significant harm.
- **exploiting vulnerabilities** related to age, disability, or socio-economic circumstances to distort behaviour, causing significant harm.
- **biometric categorisation systems** inferring sensitive attributes (race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation), except labelling or filtering of lawfully acquired biometric datasets or when law enforcement categorises biometric data.
- **social scoring**, i.e., evaluating or classifying individuals or groups based on social behaviour or personal traits, causing detrimental or unfavourable treatment of those people.
- **assessing the risk of an individual committing criminal offenses** solely based on profiling or personality traits, except when used to augment human assessments based on objective, verifiable facts directly linked to criminal activity.
- **compiling facial recognition databases** by untargeted scraping of facial images from the internet or CCTV footage.
- **inferring emotions in workplaces or educational institutions**, except for medical or safety reasons.
- **'real-time' remote biometric identification (RBI) in publicly accessible spaces for law enforcement**, except when:

- searching for missing persons, abduction victims, and people who have been human trafficked or sexually exploited;
- preventing substantial and imminent threat to life, or foreseeable terrorist attack; or
- identifying suspects in serious crimes (e.g., murder, rape, armed robbery, narcotic and illegal weapons trafficking, organised crime, and environmental crime, etc.).

High

- used as a safety component or a product covered by EU laws
- those under the use cases (below), except if:
    - the AI system performs a narrow procedural task;
    - improves the result of a previously completed human activity;
    - detects decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment without proper human review; or
    - performs a preparatory task to an assessment relevant for the purpose of the use cases listed
- AI systems are always considered high-risk if it profiles individuals, i.e. automated processing of personal data to assess various aspects of a person's life, such as work performance, economic situation, health, preferences, interests, reliability, behaviour, location or movement.
- Use cases:
    - Non-banned biometrics: Remote biometric identification systems, excluding biometric verification that confirm a person is who they claim to be. Biometric categorization systems inferring sensitive or protected attributes or characteristics. Emotion recognition systems.
    - Critical infrastructure: Safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity.
    - Education and vocational training: AI systems determining access, admission or assignment to educational and vocational training institutions at all levels. Evaluating learning outcomes, including those used to steer the student's learning process. Assessing the appropriate level of education for an individual. Monitoring and detecting prohibited student behavior during tests.
    - Employment, workers management and access to self-employment: AI systems used for recruitment or selection, particularly targeted job ads, analyzing and filtering applications, and evaluating candidates. Promotion and termination of contracts, allocating tasks based on personality traits or characteristics and behavior, and monitoring and evaluating performance.
    - Access to and enjoyment of essential public and private services: AI systems used by public authorities for assessing eligibility to benefits and services, including their allocation, reduction, revocation, or recovery. Evaluating creditworthiness, except when detecting financial fraud. Evaluating and classifying emergency calls, including dispatch prioritizing of police, firefighters, medical aid and urgent patient triage services. Risk assessments and pricing in health and life insurance.

  - Law enforcement: AI systems used to assess an individual's risk of becoming a crime victim. Polygraphs. Evaluating evidence reliability during criminal investigations or prosecutions. Assessing an individual's risk of offending or re-offending not solely based on profiling or assessing personality traits or past criminal behavior. Profiling during criminal detections, investigations or prosecutions.
  - Migration, asylum and border control management: Polygraphs. Assessments of irregular migration or health risks. Examination of applications for asylum, visa and residence permits, and associated complaints related to eligibility. Detecting, recognizing or identifying individuals, except verifying travel documents.
- Administration of justice and democratic processes: AI systems used in researching and interpreting facts and applying the law to concrete facts or used in alternative dispute resolution. Influencing elections and referenda outcomes or voting behavior, excluding outputs that do not directly interact with people, like tools used to organize, optimize and structure political campaigns.

Limited

Certain AI systems intended to interact with natural persons or to generate content may pose specific risks of impersonation or deception, irrespective of whether they qualify as high-risk AI systems or not. Users must be made aware that they interact with chatbots. Deployers of AI systems that generate or manipulate image, audio or video content (i.e. deep fakes), must disclose that the content has been artificially generated or manipulated except in very limited cases (e.g. when it is used to prevent criminal offences). Providers of AI systems that generate large quantities of synthetic content must implement sufficiently reliable, interoperable, effective and robust techniques and methods (such as watermarks) to enable marking and detection that the output has been generated or manipulated by an AI system and not a human.

Minimal

Systems presenting minimal risk for people (e.g. spam filters) will not be subject to further obligations beyond currently applicable legislation (e.g., GDPR).

General-purpose AI model or system (GPAI):

Model

An AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications. *This does not cover AI models that are used before release on the market for research, development and prototyping activities.*

System

An AI system which is based on a general-purpose AI model, that has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems.

GPAI systems may be used as high-risk AI systems or integrated into them. GPAI system providers should cooperate with such high-risk AI system providers to enable the latter's compliance.

Data Poisoning:

AI data poisoning is a type of cyber attack where an attacker deliberately manipulates the data used to train or influence an AI system, aiming to corrupt its outputs or degrade its performance. This can involve introducing subtly altered or entirely fabricated data points into a training dataset, with the intent to cause the AI to learn incorrect patterns, make errors, or exhibit biased behavior. Data poisoning targets the learning process itself, potentially compromising the integrity of the AI without direct interference with its code or operation.

Data Tuning:

AI tuning, also known as model tuning or hyperparameter optimization, involves adjusting the parameters that govern the learning process of an AI model to improve its performance. These parameters, which are not learned directly from the data but are set prior to training, can significantly influence how well an AI model trains and generalizes to new data. AI tuning aims to find the optimal set of these parameters to enhance the model's accuracy, efficiency, and effectiveness in solving specific tasks.

Data Bias:

AI data bias refers to the presence of prejudiced assumptions or partialities within the training data used for machine learning models, leading these models to systematically and unfairly discriminate against certain individuals or groups. This bias can result from non-representative or incomplete data samples, historical inequalities, or flawed data collection methods, and it often manifests in the model's decisions, predictions, or behavior, perpetuating or amplifying existing societal biases.

Deception/Trust Rot:

AI deception or trust rot refers to the erosion of trust in AI systems caused by instances where these systems intentionally or inadvertently deceive users. This can occur through the generation of misleading, inaccurate, or biased outputs, or when AI behaves in unpredictable or unexplainable ways. Trust rot undermines confidence in AI technologies, impacting their reliability and the willingness of users to adopt and interact with these systems.

Hallucination:

AI hallucination refers to instances where an AI system generates false or fabricated information in its outputs, despite being presented with accurate data. This phenomenon typically occurs in generative AI models, such as those used for text or image creation, where the AI might produce outputs that are unconnected to or inconsistent with the input data or known facts. Hallucinations can be a result of model overfitting, lack of sufficient training data, or errors in the model's learning process.

Data sheet

Accountable officer
Executive Responsible Team member responsible for Artificial Intelligence

Responsible officer
Associate Director, ITS Security

Approval
Executive

Contact area
ITS Security

Relevant dates

Approved	Executive:
Effective	January 2, 2025
Next review	October 2025
Modification history	
Verified By	Office of the President, December 2024

Associated policy(ies)
Acceptable Use of Information Technology Resources Policy (300-2-4)
Cloud Computing Policy (300-2-15)
Enterprise Risk Management Policy (600-1-4)
Information Security, and Identity Management Policy (300-2-11)
Privacy and Access Policy (300-2-10)

Directly related guideline(s) (if any)
- Control Objectives for Information and related Technology (CoBIT)

Related legislation
Bill C-27 Artificial Intelligence and Data Act (Canada)
EU Artificial Intelligence Act

Attachments (optional)