Artificial Intelligence Policy

Effective Date:
January 06, 2025

Category:
Business

Scheduled Review:
September 30, 2025

Supersedes:
ITP-BUS012, ITP-BUS014

1. Authority

Executive Order 2016-06, Enterprise Information Technology Governance

2. Purpose

This Information Technology Policy (ITP) establishes basic guidelines for the use and integration of artificial intelligence (AI) technologies and capabilities into Commonwealth business and decision processes by Commonwealth Employees and Contracted Resources (hereinafter referred to as "users").

3. Scope

This policy applies to all offices, departments, boards, commissions, and councils under the Governor's jurisdiction and any other entity connecting to the Commonwealth Network (hereinafter referred to as "agencies").

Third-party vendors, licensors, contractors, or suppliers shall meet the policy requirements of this policy as outlined herein.

4. Policy

For definitions found within this document, refer to the *IT Policy Glossary*

4.1 Key Considerations

Before agencies begin an AI initiative, they shall evaluate thefollowing key considerations:

4.1.1 Decision Making

Agencies shall evaluate how much decision-making they allow the AI solution to make. Material decisions which result in a programmatic or financial outcome should be carefully evaluated as to avoid unintended consequences if transferred to an AI solution. (Refer to the *Artificial Intelligence Guideline* for further details.)

4.1.2 Data Sets

AI solutions will rely on a pre-determined data set to complete the processes they have been assigned. Data sets identified should be reviewed to ensure quality and integrity of the data sets to minimize Overfitting and Underfitting due to Bias and Variance errors in the model that could alter the solution's decision- making patterns.

4.1.3 Methodology

Development and maintenance of an AI solution is different than a traditional IT system. AI solutions need regular and ongoing review to ensure the algorithms and models used yield the best possible result and not producing unintended consequences as changes in the data or business processes occur. As a result, business and technical decision makers need to apply a machine learning test- and-learn mentality to establish successful data analysis and determine the best model to use.

4.1.4 Audit

Agencies shall ensure they understand the algorithm and model established for the AI solutions decision making patterns. Sample data used to test and validate the algorithm shall be retained in the event an audit takes place.

Note: AI solutions that are used to formulate decisions regarding the topics listed below will be subject to audits:

- Direct or indirect material financial interests or transactions
- Administrative policy and program changes
- Benefits eligibility and determinations
- Life-changing
- Health, safety, and welfare of citizens or Commonwealth employees.

4.1.5 Disclosure

When a customer is interacting with an AI solution on behalf of an agency, the solution shall disclose to the customer that they are interacting with an AI solution

4.2 Readiness Assessment

Prior to the adoption of any AI enabled solutions, agencies shall conduct a readiness assessment and have a general understanding of AI. (Refer to the *Artificial Intelligence Assessment Tool* and *Artificial Intelligence Guideline* for further details.)

4.3 Governance

Proper governance of AI solutions is required prior to deployment of any AI solutions into the enterprise. Governance and appropriate oversight mitigate risks associated with emerging technologies. Agencies shall use existing governance bodies to ensure overall impact to business and technology operations are not negatively impacted by the integration of AI solutions.

Governing bodies are responsible for the continuous monitoring and outcomes of AI solutions to ensure alignment with business and technology strategic objectives. Governing bodies are recommended to provide oversight for the following:

- Examine the social, economic, and legal impacts of AI adoption on the workforce, citizens, and business operations.
- Determining the conditions and constraints in which supervised and unsupervised techniques will be used for training AI and algorithmic decision- makingsystems.
- Legal reviews required for use of third-party AI services, contracts, licenses, agreements, and specific use cases of AI solutions with potential impacts to workforce.

Note: it is important to understand the potential liabilities with intellectual property and data ownership associated with third party entities. Completion of the Cloud Use Case Review process for cloud-based AI solution (refer to the *IT Vendor Risk Management Policy.)*

- Legal requirements regarding transparency and disclaimers for public engagement and use of AI systems. As well as the RTE that will obligate Commonwealth agencies to explain the purpose of an algorithm and the kind of data it uses when making automated decisions. This includes third-party AI solutions. Agencies shall validate (understand) the functionality of third-party AI algorithms and how the data collected and utilized is managed by the third-party solution. The following elements shall be captured for any AI solution to satisfactorily comply with an RTE request:
    - Technical/design details of the AI system and algorithms
    - How the AI system was trained (including personnel and documentation)
    - How the AI system works (i.e. what are the inputs and outputs)
    - Data sources (documentation of all data sources)
    - Audit Logging (refer to the *IT Vendor Risk Management Policy* for audit logging requirements)
    - Change Management details and documentation that impact the AI system algorithms (i.e. decisions, inputs, outputs)
    - Testing and validation results
    - Timeframe documentation (captures time periods of testing, validations, governance approval, deployment, and other critical milestones the of AI solution)
    - Human elements (any personnel information that will assist in the RTE request)
- Decisions made by AI that have legal, financial, human resource, legislative, organizational, or regulatory impact must include a human verification process.
- Evaluate and authorize AI: technology architecture frameworks, software, platforms, libraries, software as a service (SaaS), platform as a service (PaaS), and infrastructure, and relevant tools that can be properly integrated and supported in our IT ecosystem and securely interface with our back-end services/systems.
- Protocols and procedures for assessing and handling inquiries or accidental events

regarding AI system anomalies with priority given for decisions that have potential implications for public safety or perceived workforce/labor discriminatory practices.

4.4 Data Management

Machine learning algorithms learn from data. It is critical to subject them to the right data for the problem to be solved. Even if there is a reliable and relevant data source, it is imperative to develop proper methods for data evaluations and preparedness to make sure that it is in a useful state, scale, format, composition, and representative to the problem being solved.

- Institute proper data and information management controls, procedures, and processes for data set selection, evaluation, and preparation for use with AI solutions.
- Data availability, quality, and integrity are critical for AI systems. AI systems should not be trained with data that is biased, inaccurate, incomplete, or misleading. All AI training shall be vetted through the appropriate governing processes.
- Create procedures for properly parsing data sources used with AI systems models into multiple randomized data sets consisting of training, cross-validation, and test data.
- AI systems should have access to and use only what data sources they need.
- Establish data validation procedures and processes to select, analyze, clean, and certify the quality and integrity of the data sources that will be used for AI automation solutions.
- Institute processes and procedures for preprocessing and transforming the selected data set to format, clean, sample, decompose, and aggregate the data to ensure alignment with the model and the problem being solved.

4.5 Model Testing and Validation

Machine algorithms are complex and requires expertise and practical experience in determining and implementing the best machine learning algorithms to solve the problem and form accurate outcomes. Equally important is the proper testing and validation of the model to determine the degree of Underfitting, Overfitting, and errors related to Bias and Variance. Modeling and testing methods shall be established to:

- Use AI measurement methods (accuracy, recall, and precision metrics) to evaluate each model’s performance and to choose the best model to solve the problem and produce the best results.
- Use multiple randomized data sets consisting of training, cross-validation, and test data to determine the best model and minimize potential Underfitting and Overfitting resulting from Bias and Variance errors.
- Define, validate, and document execution of hand-off criteria as to when judgment

and decisions from an AI system are transitioned to a human.

4.6 Security

Safety and security must be considered regarding full disclosure and transparency of machine designs, algorithmic models, and decisions. The following shall be considered for designing all AI systems:

- Evaluate the level of risk that AI systems are exploited by malicious actors and determine appropriate risk controls.
- Establish controls to prevent adversarial learning to include attacks that try to influence the training data of spam filters or systems for abnormal network traffic detection, designed to mislead the learning algorithm for subsequent exploitation.
- AI systems vulnerability scanning methods and techniques need to be enhanced for the discovery and categorization of security vulnerabilities or other design flaws and appropriate mitigation or resolution requirements to address known vulnerabilities. (Refer to the *eCommerce Policy.*)
- Expand incident management procedures and processes for proper handling of AI systems cybersecurity attacks or security findings to those who are in the best position to fix the problem. (Refer to the *IT Security Incident Reporting Policy.)*
- AI systems are required to comply with all Commonwealth IT Policies, Management Directives, and any other applicable regulations.
- AI systems should only collect, use, share, and store data in accordance with privacy and personal data laws and best practices.
- Establishing AI solution risk profiles based on a set of criteria to categorize and regulate the degree of oversight, review, controls, testing, documentation, and validation required pre and post deployment of AI solutions into our business and technical ecosystems. The NIST AI Risk Management Framework is currently being developed. An initial draft is available at: https://www.nist.gov/itl/ai-risk- management-framework

If using Generative AI, users may only use a Commonwealth approved Public Generative AI tool. Public Generative AI tools that receive approval are available for use only in accordance with this policy. There are many Public Generative AI tools that offer different strengths and weaknesses. The Commonwealth will continue to evaluate and may approve additional Public Generative AI tools that may be of value to users.

4.6.1 Account Creation

Generative AI tools often require that users enter an email address to register and create an account. Users, who are utilizing an approved Public Generative AI tool for Commonwealth business purposes, shall use their Commonwealth e-mail address for registration and account creation purposes.

Once created, the account associated with a user's Commonwealth e-mail address shall be used solely for Commonwealth business purposes. Personal use of Public Generative AI from an account using a Commonwealth e-mail is prohibited.

Upon completion of the registration and the account creation process, users shall opt-out of data sharing and disable the chat history within the Public Generative AI system. If unable to opt-out, the user must contact the Office of Administration, Office of Information Technology (OA IT) prior to using the Public Generative AI system.

4.7 Risk Assessment

Generative AI is a versatile technology that can be used for a variety of purposes. Like any other tool, different use cases create different risks and rewards. While Generative AI tools are relatively new, many of the risks are the same as common internet or software-based tools. Examples of common risks with Generative AI tools are:

- Sharing private or confidential information in a Generative AI prompt.
- Generative AI outputs that are inaccurate or misleading in communications to the public or relying on inaccurate or misleading outputs to inform agency programs or policies.
- Reinforcing existing Bias in work products due to Bias in Generative AI outputs.
- Copyright infringement.

Because the use of Public Generative AI tools can pose significant risk depending on the information or data input into the tool, proper governance of such tools is required. When assessing whether to use an approved Public Generative AI tool, users should consider if the use case is high or low risk and high or low impact.

High risk uses should be approached with additional review and governance and avoided when their impact is also minimal. Use cases that are low risk and high impact are potential opportunities to use Public Generative AI. Examples of such use cases are:

- **High risk/low impact (avoid)**: Using Generative AI to draft an external facing communication that includes sensitive information for citizens and would have taken minimal time to write manually. Copying and pasting that output for use with minimal review.
- **Low risk/high impact**: Using Generative AI to compare a new and old version of a publicly available policy and asking the Generative AI tool to identify which sections have been modified, then, confirming the nature of these changes manually.

Generative AI & Coding: Generative AI tools can be an excellent coding resource with high impact. Public Generative AI tools used for coding purposes should be approached with caution, and special attention should be paid to risk assessment. Subject matter experience is required to properly validate Generative AI outputs, and this requirement is particularly necessary for coding use cases. Users must be cautious not to include production code or proprietary information in prompts, must assess vulnerabilities in code outputs, and must keep in mind that assessment of these outputs may require technical knowledge. Use of Generative AI in coding may result in more bugs or flaws in programs since it may gather the code from flawed sources.

4.8 Accountability, Review, and Verification

When using any Generative AI for Commonwealth business purposes, **the user is**

accountable for any Generative AI outputs and must review and verify all associated output content.

Qualifications to verify and review outputs: For a user to be able to review and verify outputs adequately, the user must have experience in the relevant topic area. For example, a software engineer may be able to verify the quality of code generated in a coding language in which the engineer specializes but may not be able to verify if a contract is legally sound without the requisite legal training.

Generative AI content should not be assumed to be accurate. At a minimum, users should review the output for:

- **Bias:** Since the data used to train Generative AI is vast, from a variety of sources, and not always vetted, outputs may contain inaccurate assumptions or stereotypes regarding certain individuals or communities.
- **Dated Information:** The data used to train Generative AI may have a fixed cutoff date, meaning any output generated will not reflect information available after a certain cutoff date.
- **Inaccurate Information:** Generative AI relies on Training Data. Training Data is vast and not always consistent or accurate. Inaccuracies in the Training Data may be included in the output generated by the Generative AI system.

Inaccurate output can also be generated regardless of the Training Data. The Generative AI system may produce a confident response that appears plausible; however, the response is fabricated and divorced from reality (sometimes referred to as "hallucinations"). In one recent example, a user doing legal research using Generative AI was provided several court decisions, and the decisions provided by the Generative AI system turned out to be non-existent and completely fabricated.

- **Inappropriate Content:** If Training Data contains inappropriate content, the inappropriate content could appear in the Generative AI output.
- **Intellectual Property:** Generative AI tools continually ingest publicly available information for training purposes including information that may be subject to copyright. Copyrighted information could be inappropriately included in any output generated by the Generative AI system, creating intellectual property risks.
- **Confidential, Non-Public Information:** Since the data used to train Generative AI is vast, from a variety of sources, and not always vetted, outputs may contain confidential, non-public information.

4.9 Disclosure

Users shall be transparent about their use of Generative AI and must disclose to customers, residents, visitors, and industry when Generative AI has been used to generate content that may be public facing or shared externally. Generative AI use must be disclosed even if it was only used to generate a portion of the content. The disclosure shall be prominently displayed and include an indication that the content was generated either entirely or in part by Generative AI and identify the Generative AI system and version that was used.

Example: "ChatGPT-3.5 was used in the creation of this document."

4.10 Data Privacy

Users should not have any expectation of privacy when interacting with Public Generative AI tools. Any data or information that users would not include in public facing documents or emails should never be entered into any Public Generative AI tools.

4.11 Prohibited Uses

4.11.1 Production Code and Proprietary Information

While it is acceptable to use Public Generative AI tools to modify or interpret code or other content, **under no circumstances should production code or proprietary information be used in prompts.**

4.11.2 Non-Text Outputs

Users shall not utilize Public Generative AI for non-text-based outputs. Most Generative AI platforms can generate images, video, audio, or other types of content. However, the risks related to inadvertently including other's intellectual property or generating offensive content are significantly higher and more difficult to detect than with text-based outputs. Structured data, numbers, code, and different languages are acceptable outputs from Generative AI only so long as the output is properly reviewed and verified by users with the appropriate expertise.

4.11.3 Private and Sensitive Data

No class C data, as defined in the *Data Classification Policy*, may be input into any Public Generative AI prompt, tool, or system. This includes, but is not limited to:

- Sensitive Security Information
- Personal Identifiable Information (PII)
- Protected Health Information (PHI)
- Regulated Data – Such as data from or regulated by:
  - Social Security Administration (SSA)
  - Internal Revenue Service (IRS)
  - Centers for Medicare & Medicaid Services (CMS)
  - Criminal Justice Information (CJI)
  - Criminal History Record Information Act (CHRIA)
  - Family Educational Rights and Privacy Act (FERPA)
  - Payment Card Industry Data Security Standard (PCI DSS)
- Confidential or Non-Public Information
- Privileged Information
- Prerequisite-Required Information

Additionally, any non-public records or information that would be considered privileged or exempt from access under the *Right-to-Know Law (RTKL), 65 P.S. §§ 67.101, et seq.* may not be input into any Public Generative AI prompt, tool, or system.

4.11.4 Decision Making

Generative AI outputs are not to be used to make decisions for or on behalf of employees. Employees may use Generative AI outputs to inform a larger decision-making process, but ultimately the Commonwealth employee or official remains the final decision maker. Users must review and verify all output produced with the assistance of Generative AI. The user will be accountable for any decision-making based upon such output. Generative AI cannot make reliable subjective or value-based judgments and may not be used for such purposes.

- For example, do not use generative AI to make final decisions that affect employment.

4.12 Acceptable Uses

Examples of acceptable uses of Public Generative AI include:

- Drafting a job posting or job description
- Summarizing or paraphrasing a writing
- Taking a technical answer to a question and rewriting it in customer-friendly language
- Creating an outline for a memo or other communication
- Brainstorming icebreakers for a meeting

The examples provided above assume that the Generative AI tool has been approved for use; only public, non-confidential data is involved; and proper review and verification is completed as outlined in section 6.3 of this policy. This is not a comprehensive list of the permitted uses, but rather illustrates some common lower risk use cases.

5. Contact

Questions or comments may be directed via email to OA, IT Policy.

6. Exception from Policy

In the event an agency chooses to seek an exception from this policy, a request for a policy exception shall be submitted via the IT Policy Governance Process. Refer to *IT Policy Governance Policy* for guidance.

7. Revision History

This chart contains a history of this publication's revisions. Redline documents outline the revisions and are available to Commonwealth users only during the drafting process.

Version	Date	Purpose of Revision
Original	01/06/2025	Base Document