WaTech
Washington Technology Solutions

Implementing risk assessments for high-risk AI systems

December 2024

Table of contents

Purpose of the report

Executive Order 24-01 on Artificial Intelligence requires WaTech to produce guidance on risk assessments for the deployment of high-risk generative AI systems.

The EO states the risk assessments must leverage existing security and privacy assessment processes, and must include the following:

a. Information about the high-risk generative AI system, including whether the high-risk generative AI system is provided by a third party, the name and address of the third party, and relevant state agency.

b. The intended uses of the high-risk generative AI system.

c. Assessment of the fitness of the high-risk AI system for the intended purpose.

d. Assessment of impacted communities, benefits, harms, and mitigations of the high-risk AI system.

e. An evaluation of the potential harms of the high-risk generative AI system which may include harms to individuals and groups, discriminatory or unfair outcomes, deceptive practices, societal risks, privacy and cybersecurity considerations, and national security concerns.

f. An assessment of mitigations including but not limited to consideration of restricted uses and limitations on use, policies, deidentified data, and commercial terms; and

g. Information about the agency approach to generative AI governance that is consistent with the AI Risk Management framework published by the National Institute of Science and Technology.

Executive Summary

This report provides guidance on understanding AI risks, identifying high-risk systems and performing risk assessments for the deployment of high-risk generative AI (GenAI) systems. It also describes the work being done to operationalize risk identification and assessment by leveraging WaTech's existing security and privacy assessment processes. Agencies are encouraged to integrate the guidance into their workflows to ensure risk management that aligns with Washington's Responsible AI Principles and the NIST AI Risk Management Framework. By taking these steps, agencies can help ensure the safe, equitable and effective deployment of AI technologies while addressing potential harms and building trust with the communities they serve.

AI is a fast-growing technology that includes technologies like biometrics, natural language processing, and computer vision. GenAI is a subset of AI that creates content such as text, images, audio, and video. While it has the potential to transform services, it also brings challenges, such as risks to public safety, privacy, and trust in government. This report is designed to help state agencies take advantage of new technology that improves government performance and services by identifying, addressing and managing risks.

The way that AI systems are developed and implemented poses challenges for how risks can be identified and managed. Understanding these challenges, the specific types of risk posed by AI systems, and leveraging existing processes can all help agencies innovate responsibly and effectively manage risk.

Key takeaways

- **Roles in the AI lifecycle:** Risks may be introduced at three levels: in the model itself, in a system that integrates one or more models, and in specific uses of the system. Risks can be introduced at the model level based on design decisions or the data used to create the model. System risks include poor performance based on how the system is connected and implemented. Use case risks can include inappropriate use, which can be intentional or accidental. It also includes using a system for a purpose it is not well-suited for. State agencies typically participate at the system and/or use case levels and are better situated to manage risks that exist at those levels. It can be more difficult for agencies to identify and treat model risks that are already built into the AI.
- **Using principles to guide AI use:** Washington has already adopted responsible AI principles. Those principles should be used to help frame risk management activities.
- **Types of AI risks:** AI creates some challenges that are similar to traditional technology risks, but others are exacerbated or completely new. These risk types should be understood, and activities should be designed to capture risks unlikely to be considered by existing processes. Some of the most prevalent risks for state agencies include hallucinations, unfair bias, automation bias, data privacy, and information security.

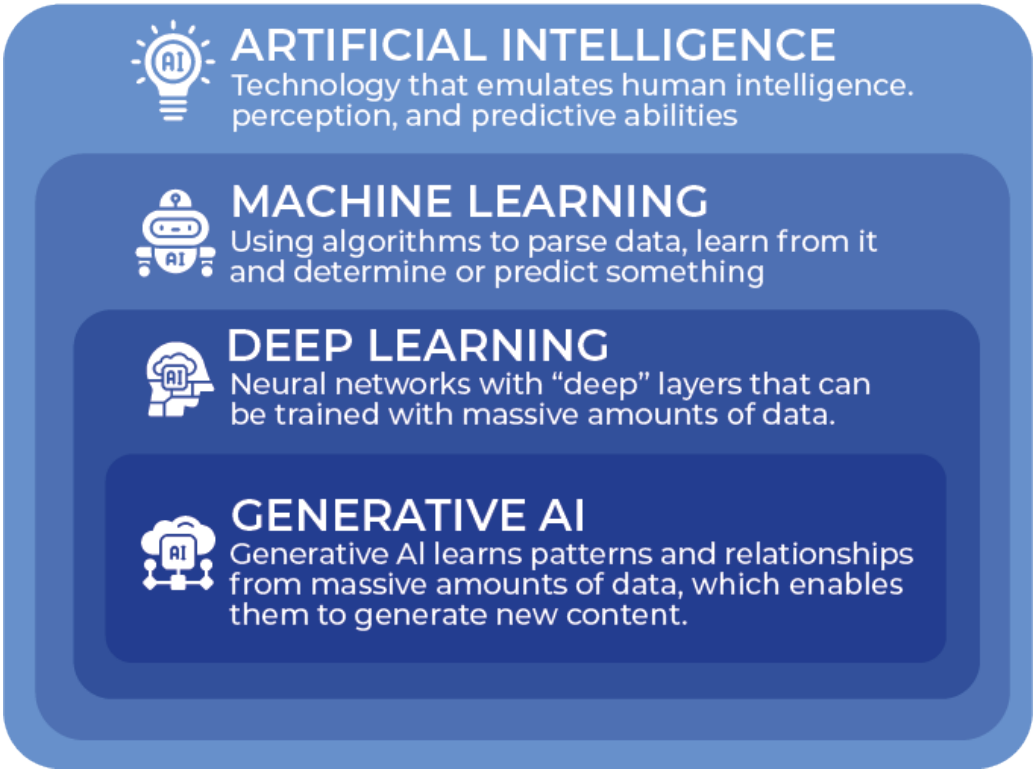
- **Factors affecting how risks can be managed:** The ability to manage risks can be influenced by where in the AI lifecycle the risk is introduced. The actual causes also vary and need to be treated in different ways. For example, they can be created by design decisions, training data, input data, or human action. The types of harms range significantly, too. Risks can affect individuals, groups, or society as a whole.
- **Defining high-risk AI systems:** High-risk AI systems are AI systems that pose a high risk of harm to individuals' health, safety or fundamental rights. Whether a system is high-risk depends on the magnitude of potential harm and the likelihood of that harm occurring. The types of harm can include direct impacts, like when an AI system is used to determine eligibility for benefits. But it can also include indirect impacts, like when an AI system is used to allocate resources in a way that impacts a particular group. Agencies should examine other categories of systems to treat as high risk based on their operating context.
- **Identifying high-risk AI systems:** Identifying high-risk systems requires new practices. Those new practices should be integrated into existing risk workflows whenever possible. This approach helps reduce redundancy, allows faster implementation, and improves the collaboration that is essential for reviewing AI systems. WaTech manages several risk processes that can be leveraged to integrate new AI risk practices, including information security risk assessments, security design reviews, privacy threshold analyses and privacy impact assessments. Those processes are well-suited to assess intended system uses at the time the system is first implemented and should be modified to require agencies to document AI system risk level during required security or privacy reviews. WaTech has created guidance to help agencies make risk level determinations. When a high-risk system is identified a complete risk assessment should be completed before implementation.

  Agencies also often experience AI incidentally, where it is implemented by a vendor as part of a solution and not central to the intended use. And AI is being implemented in existing systems that have already been reviewed. Risk assessment processes will need to be further modified to account for uses that are not central to the intended use, functionality that is added after a system has been implemented, and new uses that are not consistent with the original uses reviewed.
- **Assessing high-risk AI systems:** WaTech and members of the AI Community of Practice Risk Subcommittee are developing standard tools to be used to assess systems that have been identified as high-risk AI systems. These tools are based on Washington's responsible AI principles and key aspects of the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (NIST AI RMF). Areas of focus will include ensuring the systems are fit for the intended purpose, assessing risk to communities including the risk of unfair bias, and monitoring performance and risk management activities.
- **Implementing risk identification and assessment processes:** AI technology is rapidly evolving. Risk assessment processes should be quickly implemented to enable innovation and help agencies avoid unnecessary risks. This will be an iterative exercise and risk

assessment processes will be continually improved through experience and collaboration. In the short term, existing processes and templates will be modified to identify AI risk levels. They will then be operationalized through formal policy requirements, and ultimately risk assessment processes should be further integrated and automated. Throughout this process, WaTech will lead outreach and training, collaborative improvement on processes and tools, and consideration of available resources to implement AI risk assessment activities.

Generative AI Overview

Generative AI (GenAI) is a type of artificial intelligence that creates original content, such as text, images, audio, or video, based on users prompts. GenAI creates new outputs by analyzing patterns in large data sets using advanced machine learning and deep learning techniques. This makes it valuable for applications like writing assistance, content summarization, coding support, chatbots, language translation, and multimedia generation.



These functions are not mutually exclusive. For example, language translation can be provided in an AI chatbot or content summarization can be provided in audio format.

GenAI is distinct from other types of AI, such as:

- **Expert systems** apply known facts and expert rules to mimic decisions that a human expert might make in a specific subject area. For example, a medical diagnosis system that helps healthcare professionals make diagnoses and recommend treatments based on systems and healthcare data.
- **Computer vision** includes systems that interpret and understand visual information like images or videos. Examples include object detection, image recognition and motion tracking. It is used in applications like autonomous vehicles, medical imaging, and augmented reality.

- **Biometrics** are a person's unique physical and behavioral characteristics that can be used to identify a person or verify their identity. Artificial intelligence can be used to analyze biometrics, including physical confirmation such as facial recognition, retina scanning or voice recognition, as well as behavioral traits such as keystroke or gait analysis.
- **Natural language processing** includes systems used to understand, interpret and generate human language. Common examples include speech recognition, translation, and sentiment analysis.

For additional information about GenAI, please see WaTech's Generative Artificial Intelligence Report.

Roles in the AI lifecycle

When analyzing AI systems, it is important to recognize that risks may exist at:

- The model level (the trained algorithm itself). Model risks can be based on design decisions during development, such as how guardrails are implemented to prevent generating harmful responses. Another model risk is using non-representative training data that leads to bias in system outputs.
- The system level (the model(s) together with implementation infrastructure like user interfaces). Examples of system risks include cascading failures based on connected systems, or having insufficient filters to flag potentially inappropriate outputs for human review.
- Use case level (specific uses of the system). Use risks include inappropriate uses. They can be accidental or intentional. Another risk is choosing a system that is not well-suited for the intended use.

Different actors have different levels of control and visibility into model and system development. As used in this report, providers are organizations that create models and systems, deployers are organizations that put them into operation, and users are the individuals who interact with the systems.

State agencies usually act as deployers and/or users. They may have limited visibility into design decisions, and it can be more difficult to identify model risks. Once identified, it can be difficult for agencies to effectively treat risks that are already built in. Agencies typically have a greater ability to assess systems and have significant control over uses.

What principles guide state AI use?

Starting with clear principles helps evaluate AI risks and make sure AI uses align with Washington state values. The National Institute of Standards and Technology (NIST) has developed an Artificial Intelligence Risk Management Framework (AI RMF) to help organizations manage AI risks. As part of the AI RMF, NIST enunciated principles for Trustworthy AI. Washington state has

adopted responsible AI principles that are based on NIST's Trustworthy AI principles and Washington State Agency Privacy Principles:

- **Safe, secure, and resilient:** AI should be used with safety and security in mind, minimizing potential harm and ensuring that systems are reliable, resilient, and controllable by humans. AI technology used by state agencies should not endanger human life, health, property, or the environment.
- **Valid and reliable:** Agencies should ensure AI use produces accurate and valid outputs and demonstrates the reliability of system performance.
- **Fairness, inclusion, and non-discrimination:** AI applications must be developed and utilized to support and uplift communities, particularly those historically marginalized. Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination.
- **Privacy and data protection:** AI use should respect user privacy, ensure data protection, and comply with relevant privacy regulations and standards. Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-enhancing AI should safeguard human autonomy and identity where appropriate.
- **Accountability and responsibility:** As public stewards, agencies should use AI responsibly and be held accountable for the performance, impact, and consequences of its use in agency work.
- **Transparency and auditability:** Acting transparently and creating a record of AI processes can build trust and foster collective learning. Transparency reflects the extent to which information about an AI system and its outputs is available to the individuals interacting with the system. Transparency answers "what happened" in the system.
- **Explainable and interpretable:** Agencies should ensure AI use in the system can be explained, meaning that "how" a decision was made by the system can be understood. Interpretability of a system means an agency can answer the "why" for a decision made by the system, and its meaning or context to the user.
- **Public purpose and social benefit:** The use of AI should support the state's work in delivering better and more equitable services and outcomes to its residents.

What risks does GenAI create?

GenAI poses some risks that are similar to traditional technology risks. Other risks may be:

- **Exacerbated by GenAI:** For example, mis- or dis-information campaigns are not new. But GenAI vastly increases the ease, speed, and scale at which false information can be created and distributed.
- **Entirely unique:** The risk of increasing access to biological weapons is not typically associated with commercially available technology. But rapidly improving GenAI models

increase the risk of bad actors achieving weapons capabilities that were not previously possible without significant scientific training and expertise.

These examples highlight one unusual characteristic of GenAI risks – the risks do not inherently decrease as the technology's capabilities increase. Safety controls and guardrails can be expected to improve over time, but there is no guarantee safety measures will improve at the same rate as the capacity to cause harm accidentally or intentionally.

To further examine GenAI risks and the steps organizations can take to protect against them, NIST published Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (GenAI Profile) in July 2024. The NIST GenAI Profile is an implementation of the NIST AI RMF specifically intended to help organizations implement the RMF functions, categories, and subcategories for GenAI.

As part of the GenAI Profile, NIST identified 12 categories of risks exacerbated by or unique to GenAI. These risks vary among many dimensions. For example:

- **Stage of the AI lifecycle:** *"Risks can arise during design, development, deployment, operation, and/or decommissioning" – GenAI Profile, p.2*. The practical impact is that some risks can be treated by deployers or users at a use case level, while others are best treated during model design and development. Deployers or users often have limited visibility into design and development features that create risks.
- **Scope and scale:** The types of impacts and harms range significantly. Risks may exist at a model level or only at a specific use case level. They may cause individual (e.g., a recommendation for self-harm), group (e.g., harmful bias against a particular class of people), or societal (e.g., environmental) harms. They may occur immediately, or only accumulate over time.
- **Source of risk:** Risks can arise from many different causes. This significantly impacts the best ways to address them. For example, risks may be created from design or development decisions, training data, or other inputs from the user. Many risks come from human action, including intentional abuse or misuse, and unintentional but inappropriate use.

A short description of each of the 12 risks identified in the GenAI Profile is included below, together with a crosswalk to the most relevant Washington State Responsible AI Principles and Washington State Agency Privacy Principles for each risk.

Chemical, biological, radiological, or nuclear (CBRN) information or capabilities

- "Eased access to or synthesis of materially nefarious information or design capabilities related to chemical, biological, radiological, or nuclear (CBRN) weapons or other dangerous materials or agents." NIST GenAI Profile, p.4.
    - Washington State Responsible AI Principles: Safe, Secure, and Resilient; Explainable and Interpretable.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use.

Confabulation

- "The production of confidently stated but erroneous or false content (known colloquially as "hallucinations" or "fabrications") by which users may be misled or deceived." NIST GenAI Profile, p.4.
    - Washington State Responsible AI Principles: Valid and reliable; Fairness, inclusion, and non-discrimination, Explainable and Interpretable.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Transparency & accountability.

Dangerous, violent, or hateful content

- "Eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct illegal activities. Includes difficulty controlling public exposure to hateful and disparaging or stereotyping content." NIST GenAI Profile, p.4.
    - Washington State Responsible AI Principles: Safe, Secure and Resilient.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use.

Data privacy

- "Impacts due to leakage and unauthorized use, disclosure, or de-anonymization of biometric, health, location, or other personally identifiable information or sensitive data." NIST GenAI Profile, p.4.
- Privacy risks include using personal information to train models without transparency or consent, revealing personal information from training data during operation, and the ability to combine information to associate sensitive information with particular individuals at a scale that would not otherwise be feasible.
    - Washington State Responsible AI Principles: Safe, Secure and Resilient; Privacy and Data Protection; Accountability and Responsibility; Transparency and Auditability.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Data minimization; Purpose limitation; Transparency & accountability; Due diligence; Individual participation; Security.

Environmental impacts

- "Impacts due to high compute resource utilization in training or operating [GenAI] Models, and related outcomes that may adversely impact ecosystems." NIST GenAI Profile, p.4.
- Developing and operating GenAI is resource-intensive and can require large energy consumption and cause significant carbon emissions.
    - Washington State Responsible AI Principles: Safe, Secure and Resilient; Accountability and Responsibility.

Harmful bias or homogenization

- "Amplification and exacerbation of historical, societal, and systemic biases; performance disparities between sub-groups or languages, possibly due to non-representative training data, that result in discrimination, amplification of biases, or incorrect presumptions about performance; underserved homogeneity that skews system or model outputs, which may be erroneous, lead to ill-founded decision-making, or amplify harmful biases." NIST GenAI Profile, p.4.
    - Washington State Responsible AI Principles: Valid and Reliable; Fairness, Inclusion and Non-discrimination.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Individual participation.

Human-AI configuration

- "Arrangements of or interactions between a human and an AI system which can result in the human inappropriately anthropomorphizing [GenAI] systems or experiencing algorithmic aversion, automation bias, over-reliance, or emotional entanglement with [GenAI] systems." NIST GenAI Profile, p.4.
- Automation bias is a potentially significant risk that happens when, over time, humans experience increasing reliability of AI systems and begin to rely on system outputs more than they should. This can exacerbate other risks as humans fail to scrutinize outputs and identify performance issues like unfair bias or inaccurate results.
    - Washington State Responsible AI Principles: Safe, Secure, and Resilient; Valid and Reliable; Fairness, Inclusion, and Non-discrimination; Privacy and Data Protection; Accountability and Responsibility; Explainable and Interpretable.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Transparency & accountability.

Information integrity

- "Lowered barrier to entry to generate and support the exchange and consumption of content which may not distinguish fact from opinion or fiction or acknowledge uncertainties or could be leveraged for large-scale dis- and mis-information campaigns." NIST GenAI Profile, p.4.
    - Washington State Responsible AI Principles: Accountability and Responsibility; Safe, Secure, and Resilient; Explainable and Interpretable.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Purpose limitation; Individual participation.

Information security

- "Lowered barriers for offensive cyber capabilities, including via automated discovery and exploitation of vulnerabilities to ease hacking, malware, phishing, offensive cyber operations, or other cyberattacks; increased attack surface for targeted cyberattacks, which may compromise a system's availability or the confidentiality or integrity of training data, code, or model weights." NIST GenAI Profile, p.4.
    - Washington State Responsible AI Principles: Safe, Secure, and Resilient; Valid and Reliable; Privacy and Data Protection.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Transparency & accountability; Security.

Intellectual property

- "Eased production or replication of alleged copyrighted, trademarked, or licensed content without authorization (possibly in situations which do not fall under fair use); eased exposure of trade secrets; or plagiarism or illegal replication. " NIST GenAI Profile, p.5.
    - Washington State Responsible AI Principles: Accountability and Responsibility; Fairness, Inclusion, and Non-discrimination; Privacy and Data Protection.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Data minimization; Purpose limitation.

Obscene, degrading, and/or abusive content

- "Eased production of and access to obscene, degrading, and/or abusive imagery which can cause harm, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults." NIST GenAI Profile, p.5.
    - Washington State Responsible AI Principles: Safe, Secure, and Resilient; Fairness, Inclusion, and Non-Discrimination; Privacy and Data Protection.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Individual participation.

Value chain and component integration

- "Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not processed and cleaned due to increased automation from [GenAI]; improper supplier vetting across the AI lifecycle; or other issues that diminish transparency or accountability for downstream users." NIST GenAI Profile, p.5.
- Systems often involve many third party components that may not be fully understood or vetted. As systems grow in complexity and interconnectedness, it can be difficult to identify the cause of particular issues

- Washington State Responsible AI Principles: Safe, Secure, and Resilient; Valid and Reliable; Fairness, inclusion, and Non-discrimination; Privacy and Data Protection; Accountability and Responsibility; Explainable and Interpretable.
- Washington State Agency Privacy Principles: Transparency & accountability; Due diligence.

Additional risks not explicitly articulated by the NIST GenAI Profile include:

Non-transparency

- It is difficult or impossible for individuals to know if an AI system is being used, how that system operates, and the impacts of the systems on individuals, groups, or society.
  - Washington State Responsible AI Principles: Fairness, Inclusion, and Non-discrimination; Privacy and Data Protection; Accountability and Responsibility; Transparency and Auditability.
  - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Transparency & accountability.

Lack of explainability

- Outputs are very difficult to explain in clear and concise language that would be understandable to those auditing the system or those potentially impacted by their use. This risk may be especially prevalent when a tool is procured through a third-party vendor. Even the developers of models may not know why or how particular outputs are provided due to the size and complexity of the models.
  - Washington State Responsible AI Principles: Valid and Reliable; Accountability and Responsibility; Explainable and Interpretable.
  - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Purpose limitation; Transparency & accountability.

Lack of accountability

- Individuals who are affected by AI outputs may not have the ability to meaningfully challenge a system's decisions.
  - Washington State Responsible AI Principles: Fairness, Inclusion, and Non-Discrimination; Accountability and Responsibility; Public Purpose and Social Benefit.
  - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Transparency & accountability; Individual participation.

Threats to legitimacy and public trust

- Use of AI systems may undermine the legitimacy and public trust of governmental entities when such entities delegate decision-making responsibility to unaccountable and nontransparent systems.
    - Washington State Responsible AI Principles: Accountability and Responsibility; Public Purpose and Social Benefit.
    - Washington State Agency Privacy Principles: Lawful, fair & responsible use; Transparency & accountability.

What is a high-risk AI system?

Within Washington, the definition of high-risk is included in Executive Order 24-01 and expanded on in this report. At a minimum, agencies should identify high-risk systems using this definition and guidance. They should also consider identifying additional categories of high-risk systems at the agency level. This includes considering each agency's specific operating context and priorities. For example, an agency may want to treat a system as high-risk based if it has an impact on mission critical functions, even if it doesn't otherwise meet the definition in this report. Or a negative outcome may be so significant that an agency wants to treat a system as high-risk even if the likelihood of that outcome occurring is very low. Requirements from other jurisdictions, including the European Union and Colorado, can be a helpful resource to help identify high-risk systems or consider additional categories of high-risk systems.

High-risk systems in Washington

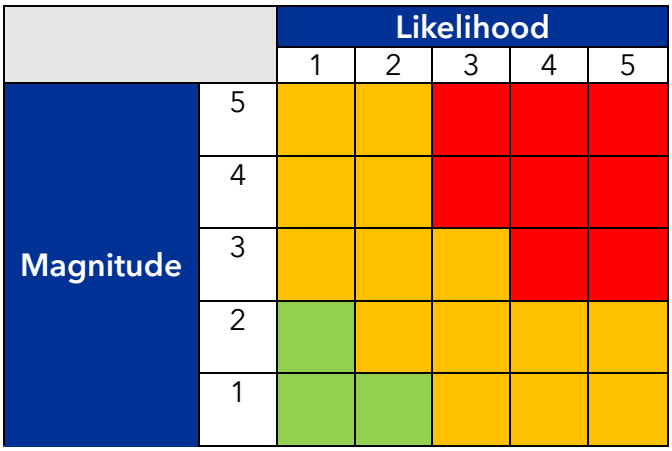
"'High-Risk Generative AI System' means systems using GenAI technology that creates a high risk to natural persons' health and safety or fundamental rights." Exec. Order 24-01, 1.b, 2024.

This definition can be readily adapted to encompass all AI, including non-generative AI systems: A high-risk AI system is a system using AI technology that creates a high risk to natural persons' health, safety or fundamental rights.

These risks include:

- Direct impacts, such as when an AI system is used to determine eligibility for benefits.
- Indirect impacts, such as when an AI system is used to allocate resources in a way that impacts the people in a particular community.

Whether a system creates a high risk is dependent on (1) the magnitude of an impact to natural persons' health, safety or fundamental rights and (2) the likelihood of that impact occurring. This can be operationalized using a risk matrix like the one below.



Magnitude	Likelihood
1 – Negligible. No foreseeable direct or indirect impact to natural persons.	**1 – Remote or improbable**. Very low chance of occurring.
2 – Low. Any impact to natural persons is very unlikely to impact health, safety or fundamental rights.	**2 – Unlikely**. Low chance of occurring.
3 – Moderate. Some impact to natural persons that may include indirect impact to health, safety or fundamental rights.	**3 – Possible**. Moderate chance of occurring.
4 – Significant. Major effect causing substantial harm or disruption to health, safety or fundamental rights. May include direct impact in individual circumstances, or indirect, systemic impacts.	**4 – Likely**. High chance of occurring.
5 – Severe or catastrophic. Extreme impact resulting in serious harm, injury or violation of fundamental rights.	**5 – Probable**. Very high chance of occurring.

Potential risk factors that impact the magnitude or likelihood of a particular risk include at least:

- **The intended use and operating context:** What are the outputs and who is intended to benefit? Who might be adversely impacted? Who are the users and what is the role of humans in reviewing system outputs?
- **Data characteristics:** Will the system use confidential or personal information? Is necessary data available and high-quality?
- **System characteristics and safeguards:** Is there evidence of biased, discriminatory, inaccurate or otherwise unreliable outputs? What measures are in place to address unreliable outputs? How explainable is the system? How transparent is system implementation?

Using this matrix, a system in the green area is low risk, a system in the yellow area is moderate risk, and a system in the red area is high risk. For example, a GenAI tool used solely for internal communications about agency logistics with human review of every output would be low risk because it does not have a foreseeable impact to people and any impact has a remote chance of happening. A system with the potential for biased outputs that is used to make decisions that impact fundamental rights without human review would be high risk, because the impacts would be significant and are likely to occur.

An additional resource, Artificial Intelligence Risk Level Guidance, is attached as Appendix A.

Other risk-based approaches

There are few significant regulatory regimes that apply specifically to artificial intelligence. The most prominent examples are the European Union AI Act (EU AI Act) in Europe and the Colorado AI Act in the United States. Although Washington state agencies are unlikely to be subject to these laws, they provide context and examples for agencies to consider as they implement AI risk management strategies.

European Union AI Act

The EU AI Act regulates three primary categories of AI systems: prohibited systems, high-risk systems, and systems with transparency requirements only. General purpose AI models are also identified as a separate category of AI systems requiring different considerations. Different requirements apply for each category, and the requirements vary depending on an AI actor's role. For example, an AI developer has different requirements than an organization deploying an AI system created by someone else.

The EU AI Act lists eight categories of AI uses that create an unacceptable level of risk and are therefore prohibited.

Prohibited Uses			
Subliminal, manipulative, or deceptive techniques	Exploiting vulnerabilities	Biometric categorization	Social scoring
Crime risk prediction	Facial recognition databases from untargeted scraping	Emotion detection in school or workplace setting	Real-time remote biometric identification

The EU AI Act also includes a specific list of high-risk AI systems, with different requirements for providers and deployers. Similar to the definition in Washington, these specific examples are not high-risk if they do not pose a significant risk of harm to the health, safety or fundamental rights of natural persons.

High-risk systems			
Other biometrics	Managing critical infrastructure	Determining access to education or assessing performance	Employment purposes, including candidate or work evaluation
Determining access to essential private or public services and benefits	Law enforcement uses other than prohibited uses	Immigration management and border control	Administration of justice and democratic processes

Colorado AI Act

In May 2024, Colorado became the first state to pass a law that directly regulates AI. It is scheduled to go into effect in February 2026. Like the EU AI Act, it creates different requirements for providers (which it calls developers) and deployers. And like both the EU AI Act and the risk assessments described in this report, it is focused on high-risk systems. But it uses a slightly different definition of high-risk AI system than either the EU AI Act or Washington state.

"High-risk artificial intelligence system" is defined as any AI system that makes, or is a substantial factor in making, a consequential decision. And a consequential decision is a decision that has a significant effect on:

- Education enrollment or an education opportunity.
- Employment or an employment opportunity.
- A financial or lending service.
- An essential government service.
- Health-care services.
- Housing.
- Insurance.
- A legal service.

This definition is focused on the type of decision that is being made, rather than whether there is a high risk of harm occurring. This is a helpful example for Washington agencies to consider in identifying high-risk AI systems in at least two ways. First, it includes a specific list of domains that could be used as examples of the term "fundamental rights" in the Washington state definition. Second, because AI is a nascent technology and rapidly evolving, the likelihood and magnitude of risks can be difficult to quantify. Focusing on whether a harm would be significant if it did occur can be a simpler exercise than calculating the likelihood of it occurring.

What processes can help identify high-risk AI systems?

Consistently identifying high-risk AI systems will require new or modified practices including processes, policies, standards, and training. Whenever possible these new practices should be integrated into existing workflows rather than created as new, standalone requirements.

Key benefits to this approach include:

- **Reduced redundancy:** Integrating into existing processes helps align evaluations with established frameworks. Work that is repeated in each process can be consolidated or re-used, and AI risks can be considered alongside other concerns such as privacy, cybersecurity and compliance.
- **Faster implementation and organizational readiness:** The time required to adopt AI risk assessment processes is accelerated by using procedures and workflows that teams are already familiar with. Instead of going through the time-consuming process of developing and communicating entirely new processes, existing processes can be expanded and adapted to address specific AI concerns.
- **Improved collaboration:** In addition to technical risks, AI presents societal risks like bias and privacy violations. Technical and societal risks should be considered together and having integrated processes helps ensure risks and opportunities for mitigation are considered holistically.
- **Future improvements:** Bringing processes together now will make future process improvements easier. For example, automation and rule-based logic can be used to make these processes one cohesive workflow rather than integrated but largely parallel processes. This will further reduce redundancy and unnecessary work, freeing time to focus on identifying and managing the highest priority risks.

WaTech manages security and privacy risk assessment processes that should be leveraged to implement new AI risk assessment activities. These include risk assessments required by the Risk Assessment Standard, security design reviews (SDRs) required by the Security Assessment and Authority Policy, and privacy assessments required by the Privacy and Data Protection Policy.

Risk assessment standard

As required by the Risk Assessment Standard, agencies must conduct information security risk assessments (ISRAs) in six situations:

- Prior to the acquisition of an information system, cloud service, or managed service which will store, process, or transmit category 3 or category 4 data.
- When an existing agency-controlled information system undergoes a significant change in technology or use. Examples include significant software upgrades, changes in hosting platforms or vendors, or changes in the data categorization or volume of records stored, processed, or transmitted by the system.

- At least once every three years for all agency-controlled information systems that store, process, or transmit category 3 or category 4 data.
- Annually for information systems the agency deems to be business essential.
- Prior to the sharing of category 3 or category 4 data as with agencies and/or vendors. See SEC-08 Data Sharing Policy and SEC-08-01-S Data Classification Standard for details.
- When a security patch is not applied.

Security assessment and authorization policy

As required by the Security Assessment and Authority Policy, an SDR is required in three situations:

- A new agency IT implementation that includes at least one of the following conditions:
    - A. Agency-managed Cloud services – SaaS, PaaS, and IaaS.
    - B. Vendor-managed Cloud or dedicated hosting.
    - C. Internet available services hosted on-premises.
    - D. If required by the agency security program or policies.
- The WaTech SDR team assesses IT implementations under oversight and determines whether a WaTech SDR is required for the proposed technological solution(s). See the PM-01 IT Investments - Approval and Oversight Policy.
- The agency is planning significant changes for a solution previously reviewed and approved by the SDR team. See SEC-05 Change Management Policy.

Privacy and data protection policy

At a minimum, a privacy threshold analysis (PTA) is required for projects that process personal information at the time an SDR is opened. The PTA is a brief, high-level description of the technology, the data involved, and how it will be used, shared or otherwise processed. When the PTA indicates the potential for significant privacy risks or privacy harms, agencies must complete a more comprehensive privacy impact assessment (PIA).

The PTA/PIA process is integrated into the SDR process. When an SDR is opened, the submitting agency indicates whether the project involves processing personal information. If it does, a PTA is automatically assigned and the SDR will not be closed until the PTA is submitted and accepted. Agencies are encouraged to begin working on PTAs prior to opening SDRs for projects that are known to involve personal information.

Challenges to identifying high-risk AI systems

The three risk processes described above should be modified to require agencies to document the identified risk level any time an AI system undergoes an ISRA or SDR. There are additional challenges to using these processes to identify high-risk AI systems that will require further modifications.

The processes include requirements to either periodically review assessments or re-visit when there are significant changes. But they are primarily aimed at the initial implementation of new technologies. That is not always consistent with how AI is operationalized.

As described in the GenAI Intelligence Report, state agencies commonly experience AI in three ways:

- **Intentional AI** refers to intentional adoption of AI-enabled technology solutions.
- **Incidental AI** refers to new or existing technologies with AI enabled within them, but AI is not their primary purpose or the reason the technology is being used.
- **Third party AI** refers to vendors using AI to perform services.

For all three types, the AI's original functionality and intended uses may be known at the time the technology is implemented or a relationship with a new vendor is implemented. But AI functionality is also added or modified later. Similarly, existing functionality may be used in a new way.

These two dimensions – (1) how an agency is experiencing AI and (2) shifting capabilities or uses – pose different challenges for first identifying the use of AI and then identifying high-risk uses.

<table>
<tr><th></th><th>Intentional AI</th><th>Incidental AI</th><th>Third-party AI</th></tr>
<tr><td rowspan="2">Original functionality</td><td rowspan="2">Existing ISRA, SDR, and PTA/PIA processes are well-equipped to identify these use cases.</td><td>Existing ISRA, SDR, and PTA/PIA processes can be used to identify these use cases.</td><td>Existing ISRA, SDR, and PTA/PIA processes are not well-equipped to capture AI uses that are not part of a technology implementation.</td></tr>
<tr><td>Processes and templates can be modified to capture appropriate information.</td><td>Procurement requirements and contact terms can help identify these uses.</td></tr>
</table>

Changes to capability or use	Existing ISRA, SDR and PTA/PIA processes can be used to identify these use cases.	Existing ISRA, SDR and PTA/PIA processes can be used to identify these use cases.	Contract terms can help identify these uses.
	Additional training and periodic review will help ensure these uses are identified.	Additional training and periodic review will help ensure these uses are identified.	Existing ISRA, SDR, and PTA/PIA processes are not well-equipped to capture AI uses that are not part of a technology implementation.
		Contract terms can help identify these uses.	

In addition to requiring risk level identification during any AI system ISRA or SDR, the additional changes described above should be taken risk identification and assessment capabilities., A phased plan for how these considerations can be implemented across Washington state agencies is included in the "Risk identification and assessment implementation" section of this report.

How should high-risk systems be assessed?

Whenever an AI system is identified as high-risk, a complete AI risk assessment should be completed prior to implementation. WaTech is developing an AI Risk Assessment (AIRA) template that can be used to perform this risk assessment. The template is organized around Washington's Responsible AI principles and includes citations to relevant NIST AI RMF sections to ensure consistently with that framework. *See Executive Order 24-01, Section 8.g*.

For each section, agencies should identify risks, identify steps that have been taken to address those risks, and establish commitments to how risks should be measured and monitored on an ongoing basis.

The tool will include 10 sections:

- **Section 1 – System identification and contacts:** This section identifies the project, the relevant agency contacts, and the vendors involved. *See Executive Order 24-01, Section 8.a*.
- **Section 2 – System information and operating context:** This section gathers basic information about the system and operating context, including its intended purpose and what policy or legal requirements apply. *See Executive Order 24-01, Section 8.b*.

- **Section 3 – Public purpose and social benefit:** This section addresses who is intended to benefit from the system, how success will be measured, and why the AI system is the preferred option. *See Executive Order 24-01, Section 8.c.*
- **Section 4 – Safe, secure and resilient:** This section considers the role of humans in system development and operation, system functionality to prevent harm, and security vulnerabilities. *See Executive Order 24-01, Sections 8.d, 8.e and 8.f.*
- **Section 5 – Valid and reliable:** This section collects the specific performance metrics to ensure outputs are accurate and meet intended system purpose, and how unintended or inappropriate use will be avoided. *See Executive Order 24-01, Sections 8.c, 8.f and 8.g.*
- **Section 6 – Fairness, inclusion, and non-discrimination:** This section explores people or communities that may be adversely affected, how the system will be evaluated for biased outputs, and what controls are in place to avoid creating or reinforcing unfair bias. *See Executive Order 24-01, Sections 8.d and 8.e.*
- **Section 7 – Privacy and data protection:** This section ensures that privacy impacts have been thoroughly considered, and that appropriate data is used to train and operate the system. *See Executive Order 24-01, Sections 8.e and 8.f.*
- **Section 8 – Accountability and responsibility:** This section defines how feedback on system performance will be gathered and considered, and AI governance requirements for system vendors. *See Executive Order 24-01, Sections 8.e and 8.f.*
- **Section 9 – Transparency and auditability:** This section explains the way system use, including limitations and intended uses, is explained to users. It also addresses logging system outputs. *See Executive Order 24-01, Sections 8.e and 8.f.*
- **Section 10 – Explainable and interpretable:** This section documents what resources are available to address lack of explainability and interpretability. *See Executive Order 24-01, Sections 8.e and 8.f.*

Risk identification and assessment implementation

Through the AI Community of Practice and its subcommittees WaTech has been facilitating development of policies, processes, and tools to implement AI risk identification, assessment and management activities. The table below includes the planned approach that WaTech and agencies across the enterprise should take to implement these activities:

SHORT-TERM Next 3 months	MID-TERM 3 months – 1 year	LONG TERM 1+year
WaTech: Distribute initial version of AI risk assessment template	**WaTech:** Establish enterprise AI policy requirements.	**WaTech:** Consolidate processes to gather information one-time as part of cohesive workflow.
Agencies: Modify risk assessment Information Security Risk Assessment templates to include identification of AI risk level	**Agencies:** Operationalize AI policy requirements.	**WaTech:** Implement a software platform to manage assessment activities in an integrated, streamlined manner.
	WaTech and Agencies: Evaluate long-term resourcing plan to identify risks and responsibly adopt innovative AI technologies.	
WaTech: Modify Security Design Review process to require identification of AI risk level.	**WaTech and Agencies:** Iterative improvement of AI risk assessment processes and tools.	
	Agencies: Modify procurement and contracting practices as described in the reports required by EO 24-01, sections 3 and 6.	
WaTech: Outreach and training on risk assessment processes, in addition to the training described in the report required by EO 24-01, section 4.		

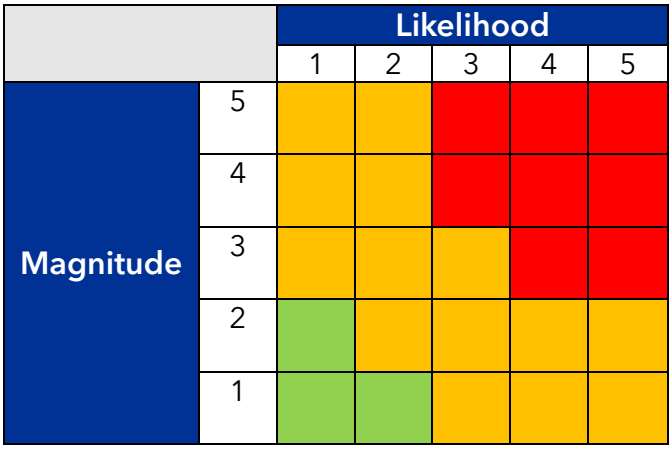
Contact

For questions about this report, please contact Angela Kleis, WaTech Director of Policy & External Affairs: angela.kleis@watech.wa.gov.

Appendix A: Artificial Intelligence Risk Level Guidance

Agencies must determine whether AI-enabled systems are high-risk and conduct a complete Artificial Intelligence Risk Assessment (AIRA) prior to implementing a high-risk AI system. This guidance describes when an AI system is "high-risk," and includes considerations to help make that determination.

The information in this document specifically addresses the minimum requirements for AI systems that pose a high risk to health, safety or fundamental rights. Agencies may prioritize other types of risks posed by AI systems. For example, an AI system could pose a high reputational or operational risk even if it does not impact health, safety or fundamental rights. Similarly, agencies may have different risk tolerances based on their operating context. Agencies are encouraged to consider additional types of risk that justify conducting an AIRA.



Magnitude	Likelihood
1 – Negligible. No foreseeable direct or indirect impact to natural persons.	**1 – Remote or improbable**. Very low chance of occurring.
2 – Low. Any impact to natural persons is very unlikely to impact health, safety or fundamental rights.	**2 – Unlikely**. Low chance of occurring.
3 – Moderate. Some impact to natural persons that may include indirect impact to health, safety or fundamental rights.	**3 – Possible**. Moderate chance of occurring.
4 – Significant. Major effect causing substantial harm or disruption to health, safety or fundamental rights. May include direct impact in individual circumstances, or indirect, systemic impacts.	**4 – Likely**. High chance of occurring.
5 – Severe or catastrophic. Extreme impact resulting in serious harm, injury or violation of fundamental rights.	**5 – Probable**. Very high chance of occurring.

What is a high-risk AI system?

A high-risk AI system is a system using AI technology that creates a high risk to natural persons' health, safety or fundamental rights. These risks include:

- Direct impacts, such as when an AI system is used to determine eligibility for benefits, and
- Indirect impacts, such as when an AI system is used to allocate resources in a way that impacts the people in a particular community.

Whether a system creates a high risk is dependent on (1) the magnitude of an impact to natural persons' health and safety or fundamental rights and (2) the likelihood of that impact occurring.

Potential risk factors.

When determining whether an AI system is high-risk, consider at least:

- The intended use and operating context
- Data characteristics
- Systems characteristics and safeguards

The intended use and operating context

- What is the intended use of system outputs and who will benefit?
- Who might be adversely impacted and how? Consider at least the workforce, customers/residents, and vulnerable or underserved populations.
- What is the role of humans in reviewing system outputs? For example, do they have little to no review, do humans review in defined circumstances, or do humans verify, modify, or review all outputs?
- Who are the intended users? Is it intended for internal or external use? Is it intended for broad use or a small group of people?
- What laws apply to agency use of the system?
- What alternatives were considered and why is this the preferred option?

Data characteristics

- What is the highest data classification level for included information? Consider both information used for training and information processed during operation, including intended prompts, inputs, fine-tuning, augmentation or any other system customization.
- What type of information is involved? Will the system process personal information?

System characteristics and safeguards

- Is there evidence of biased, discriminatory, inaccurate or otherwise unreliable outputs? What measures are in place to address unreliable outputs?
- Are there built-in filters or monitoring to detect harmful outputs, manipulation, or misuse?
- How explainable is the system? Do users know the system uses AI? Does the system provide sources or otherwise help identify to users why it is producing a particular output?
- How transparent is system implementation? Are users trained or notified of appropriate use and system limitations?