Address Security & Privacy Risks for Generative AI

Artificial intelligence risk mitigation starts with an acceptable use policy.

INFO~TECH RESEARCH GROUP

Analyst Perspective

Generative AI needs an acceptable use policy

When it comes to using generative AI (Gen AI), the benefits are tangible, but the risks are plentiful, and the tactics to address those risks directly are few.

Most risks associated with Gen AI are data-related, meaning that effective AI security depends on existing maturity elsewhere in your security program. But if your data security controls are a bit lacking, that doesn't mean you're out of luck.

The good news is that the greatest and most common risks of using Gen AI can be addressed with an *acceptable use policy*. This should be top priority when considering how your organization might incorporate Gen AI into its business processes.

Some future-state planning will also help you determine which other parts of your security program require upgrades to further reduce AI-related risks. What exactly must be improved, however, will depend on your specific use case.

Logan Rohde
Senior Research Analyst, Security & Privacy
Info-Tech Research Group

Executive Summary

Your Challenge

- Governing enterprise use of Gen AI to maximize benefits and minimize risks
- Protecting data confidentiality and integrity when using Gen AI systems
- Responding to an ever-shifting threat landscape

Organizations seeking to become early adopters of Gen AI may not have security teams presently equipped to mitigate the risks. Some may need to retroactively apply governance if unauthorized AI use is already happening.

Common Obstacles

- Uncertainty assessing risks of new technology
- Difficulty implementing governance
- Immaturity of existing data security controls

Unfamiliarity with Gen AI may create confusion about how to assess and mitigate risks, especially when determining how the technology can be used. Given the additional need for strong data security controls to address Gen AI risks, it can be difficult to know which issue must be addressed first.

Info-Tech's Approach

- Determine which risks apply to your Gen AI use cases
- Draft an AI security policy to address those risks
- Plan to address necessary improvements to data security posture

In most cases, Gen AI presents novel versions of familiar data security risks, meaning that most organizations only need to improve or expand existing controls rather than create new ones.

Info-Tech Insight

Start with what you can control. Using Gen AI carries significant data security risks. By determining which risks apply to you and implementing governance to address them, you can limit the most pressing issues with using Gen AI and plan to address larger, systemic issues with your data security program.

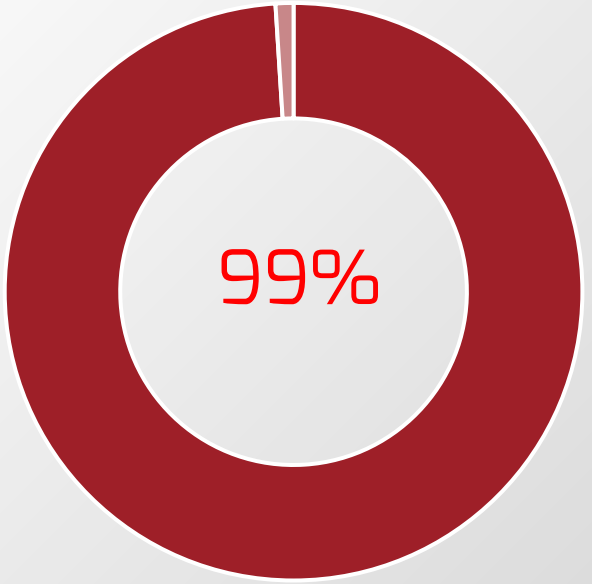
Your challenge

This research is designed to help organizations who need to:

- Evaluate risks associated with enterprise use of Gen AI.
- Assess suitability of existing security controls to mitigate risks associated with Gen AI.
- Communicate risks to the business and end users.
- Determine acceptable use criteria for Gen AI.

Implement Gen AI governance now – even if you don't plan to use it. Without an official policy, end users won't know the organization's stance. Moreover, the technology can make it easier for bad actors to execute various types of cyberattacks, and such risks should be communicated throughout the organization.

IT leaders who believe their organization is not presently equipped to leverage Gen AI:



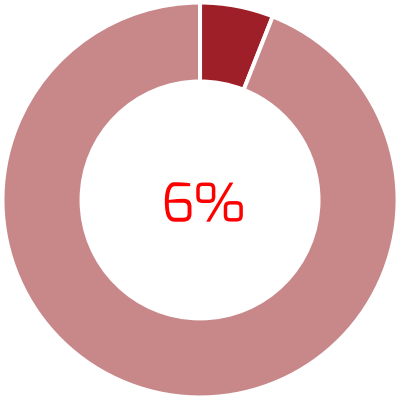
Source: Salesforce, 2023

Common obstacles

These barriers make this challenge difficult to address for many organizations:

- The novelty of Gen AI leaves many security and IT leaders unsure about how to evaluate the associated risks.
- What many miss about these risks, however, is that most are new versions of familiar data security risks that can be mitigated by defining acceptable use and necessary security controls to support governance of Gen AI.
- Assessing risk and defining acceptable use are the first key steps to Gen AI security improvement. Organizations must also re-evaluate their data security controls and plan necessary improvements to further mitigate risks associated with enterprise use of Gen AI.

Don't fall behind on Gen AI risk management

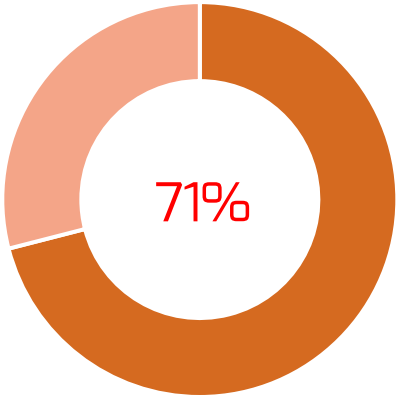


Organizations with a dedicated risk assessment team for Gen AI

Source: KPMG, 2023

IT leaders who believe Gen AI will introduce new data security risks

Source: KPMG, 2023



Key risk types for Gen AI

Data security and privacy

- The greatest risk associated with using Gen AI is a loss of data confidentiality and integrity from inputting sensitive data into the AI system or using unverified outputs from it.

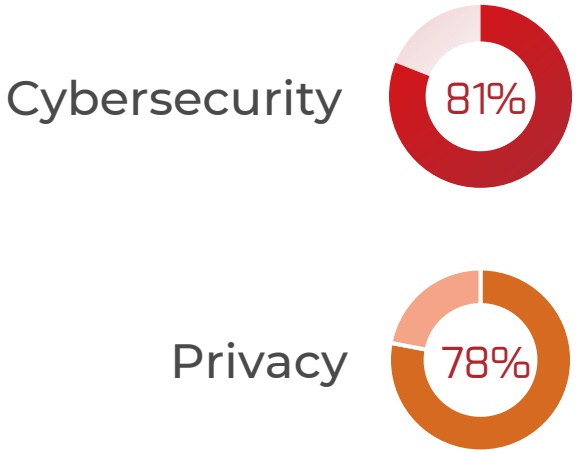
Data confidentiality

- Care must be taken when choosing whether to enter a given data type into an AI system. This is especially true in a publicly available system, which is likely to incorporate that information into its training data.
- Problems may still arise in a private model, particularly if the AI model is trained using personal identifiable information (PII) or personal health information (PHI), as such information may appear in a Gen AI output.

Data integrity

- Data integrity risk comes from repeatedly using unverified Gen AI outputs. A single output with faulty data may not cause much trouble, but if these low-quality outputs are added to databases, they may compromise the integrity of your records over time.

Top-of-mind Gen AI concerns for IT leaders



Source: KPMG, 2023

AI model versus AI system

AI model

An algorithm used to interpret, assess, and respond to data sets, based on the training it has received.

AI system

The infrastructure that uses the AI model to produce an output based on interpretations and decisions made by the algorithm.

Sources: TechTarget, 2023; NIST, 2023

Info-Tech Insight

The terms *AI model* and *AI system* are sometimes used interchangeably, but they refer to two closely related things. In many cases, the tactics to secure Gen AI will overlap, but sometimes additional security controls may be required for either the model or the system.

Public versus private AI

Public AI system

- Uses a publicly available AI system that benefits from multiple users worldwide entering data that can be used to further train the AI model.
- Carries significant risk of accidental data exposure and low-quality outputs compromising data integrity.
- Risk of attack on the AI system is owned by the vendor rather than the user.

Private AI system

- Private system used only within the organization that owns it.
- Data confidentiality and privacy risks are fewer, but still exist (e.g. PII used in training data).
- Outputs should be verified for quality before being used in business processes to prevent data integrity issues.
- Owner of the system assumes the attack risks.

Attacks on Gen AI

Input attacks

- Using knowledge of how the AI model has been trained, an input is entered that causes it to malfunction (e.g. misinterpret a risk as something benign).
- Often preceded by data exfiltration attack to learn how model works or what data it has been trained with.
- Data confidentiality should always be protected.

Data poisoning

- Training data is tampered with to corrupt AI model integrity.
- AI system data should be audited regularly to ensure data integrity has not been compromised.
- Data resiliency best practices should be followed (e.g. backups and recovery time objective [RTO] and recovery point objective [RPO] testing).

"Because few developers of machine learning models and AI systems focus on adversarial attacks and using red teams to test their designs, finding ways to cause AI/ML systems to fail is fairly easy."

– Robert Lemos, Technology Journalist and Researcher, Lemos Associates LLC, in Dark Reading

Attacks on Gen AI

Weaponization of AI model

- The AI system is compromised via malicious code used to distribute it throughout the organization (e.g. ransomware attack).
- Access to knowledge of AI model, system, and training data should be on a need-to-know basis.
- Unverified code should not be incorporated into AI system.

Sponging

- A series of difficult-to-process inputs are entered into the AI model to slow down its processing speed and increase energy consumption (similar to denial-of-service [DoS] attack).
- The AI system should be designed with a failure threshold to prevent excessive energy consumption.

AI-assisted cyberattacks

Regardless of whether an AI system is public or private, we must all contend with the risk that someone else will use Gen AI to facilitate a familiar cyberattack.

Phishing

- Gen AI can be used to create convincing phishing emails, not just with text, but with images, sound, and video (i.e. deepfake).

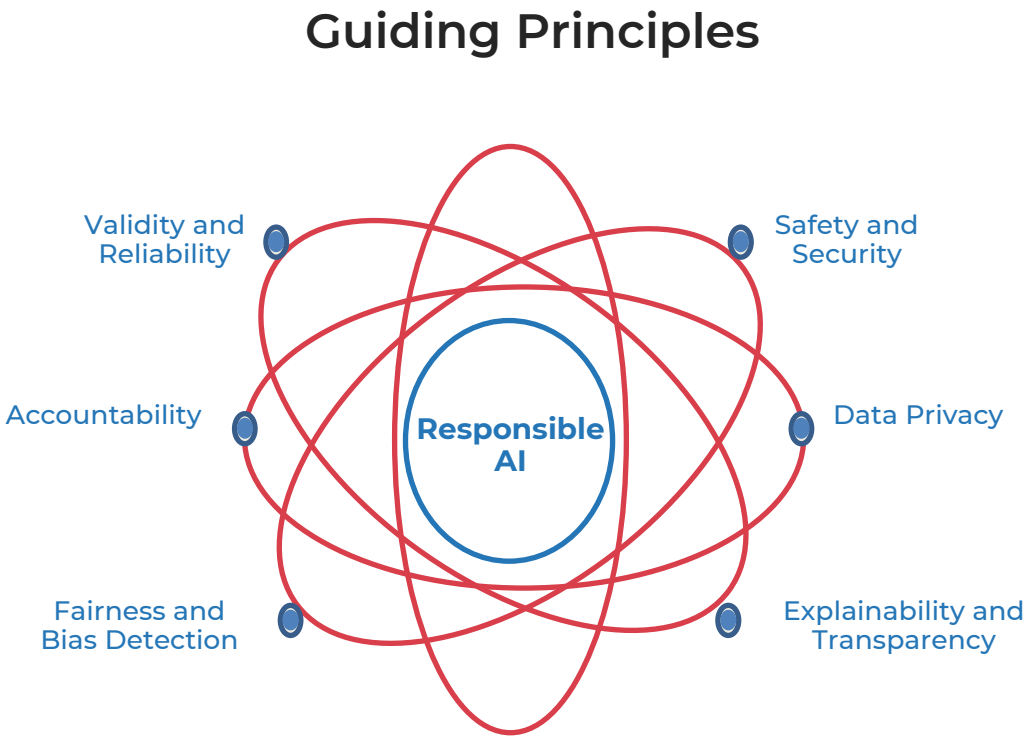
Malware

- Gen AI chatbots are programmed not to help people carry out illegal activities. However, the way the question is asked can influence the chatbot's willingness to comply.

"AI generation provides a novel extension of the entire attack surface, introducing new attack vectors that hackers may exploit. Through generative AI, attackers may generate new and complex types of malware, phishing schemes and other cyber dangers that can avoid conventional protection measures."

– Terrance Jackson, Chief Security Advisor, Microsoft, in "Exploring," Forbes, 2023

Guiding principles of responsible AI



Principle #1 – Privacy

Individual data privacy must be respected.

- Do you understand the organization's privacy obligations?

Principle #2 – Fairness and Bias Detection

Unbiased data will be used to produce fair predictions.

- Are the uses of the application represented in your testing data?

Principle #3 – Explainability and Transparency

Decisions or predictions should be explainable.

- Can you communicate how the model behaves in nontechnical terms?

Principle #4 – Safety and Security

The system needs to be secure, safe to use, and robust.

- Are there unintended consequences to others?

Principle #5 – Validity and Reliability

Monitoring of the data and the model needs to be planned.

- How will the model's performance be maintained?

Principle #6 – Accountability

A person or organization must take responsibility for any decisions that are made using the model.

- Has a risk assessment been performed?

Principle #n – Custom

Add principles that address compliance or are customized for the organization/industry.

Gen AI essentials

1. AI suitability test

Before committing to Gen AI deployment, make sure the benefits outweigh the risks and that there is a specific advantage to using Gen AI as part of a business process.

2. Gen AI risk mapping

Risks will emerge depending on use and therefore will vary somewhat between organizations. Determining which ones apply to you will affect how you govern Gen AI use.

3. Gen AI security policy

A policy detailing required security protocols and acceptable use for Gen AI is the most immediate step all organizations must take to deploy Gen AI securely.

4. Data security improvement plan

Enterprise use of Gen AI carries significant risks to data security. If any current controls are insufficient to account for Gen AI risks, a plan should be in place to close those gaps.

Manage Security and Privacy Risks for Generative AI

INFO-TECH RESEARCH GROUP

#ITRG

Determining acceptable use is the first step

Unfamiliarity with generative AI can cause some confusion about how to assess and mitigate risks, especially when determining what the technology can and cannot be used for. There is also a need for strong data security controls to address generative AI risks, and it can be difficult to know which issue to address first.

RISK MITIGATION ESSENTIALS

Upholding Responsible AI Principles

- AI Suitability Test
- AI Security Policy
- Generative AI Risk Map
- Data Security Improvement Plan

Public AI System

RISK SCENARIOS	RISK CATEGORIES	ATTACKS ON GENERATIVE AI
Data Exposure	Data Confidentiality	Weaponization of AI Model
		Data Exfiltration → Input Attack
Bad Data	Data Integrity	Data Poisoning
		Sponge Attack

Private AI System

Start with what you can control

Using generative AI carries significant data security risks. But by determining which risks apply to you and implementing governance to address them, you can limit the most pressing issues with using generative AI and plan to address larger, systemic issues with your data security program.

Determining acceptable use is the first step

Start with what you can control

Using Gen AI carries significant data security risks. By determining which risks apply to you and implementing governance to address them, you can limit the most pressing issues with using Gen AI and plan to address larger, systemic issues with your data security program.

Look for problems before getting invested

While Gen AI opens many possibilities, some risks will be difficult to address. For example, if your proposed use case requires sensitive data to be entered into a public AI system to produce an output for use in your supply chain, it will be virtually impossible to mitigate such risks effectively.

Build a strong perimeter

AI security is still in its early stages and best practices are still being determined. Until more specific controls and techniques are developed, the best course of action is to use a robust data security program to make your sensitive data as difficult to access as possible, and to monitor for intrusions.

Risk likelihood just went up

The use of Gen AI to facilitate cyberattacks doesn't fundamentally change the nature of the risk. But because Gen AI makes the process easier, we should account for this in our risk assessments.

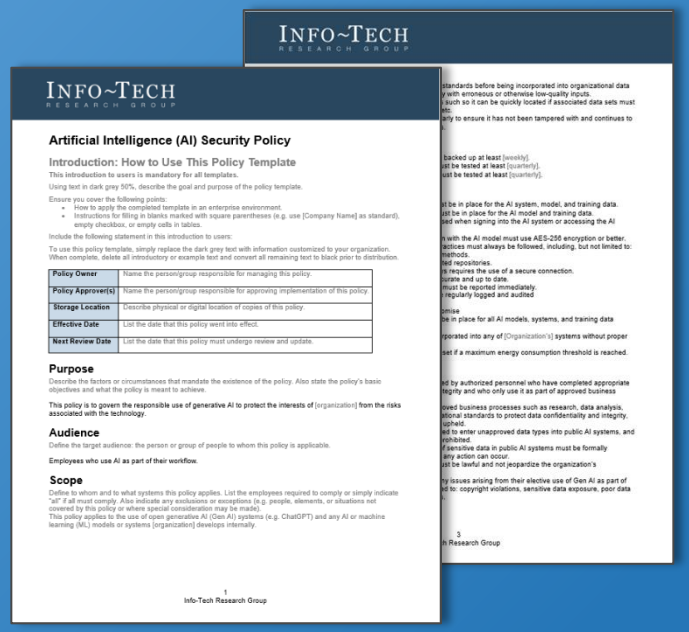
Watch for overlap

There will usually be both an input and an output component when using Gen AI, which means both risk factors are present, but one may be dominant. Therefore, both inputs and outputs should receive sign-off before use to limit data confidentiality and integrity risks.

Key deliverable:

AI Security Policy Template

Set standards for data confidentiality and integrity, acceptable use, and technical IT controls.



INFO-TECH RESEARCH GROUP

Artificial Intelligence (AI) Security Policy

Introduction: How to Use This Policy Template

This introduction to users is mandatory for all templates.

Using text in dark grey 50%, describe the goal and purpose of the policy template.

Ensure you cover the following points:

- How to apply the completed template in an enterprise environment.
- Instructions for filling in blanks marked with square parentheses (e.g. use [Company Name] as standard), empty checkbox, or empty cells in tables.

Include the following statement in this introduction to users:

To use this policy template, simply replace the dark grey text with information customized to your organization. When complete, delete all introductory or example text and convert all remaining text to black prior to distribution.

Policy Owner	Name the person/group responsible for managing this policy.
Policy Approver(s)	Name the person/group responsible for approving implementation of this policy.
Storage Location	Describe physical or digital location of copies of this policy.
Effective Date	List the date that this policy went into effect.
Next Review Date	List the date that this policy must undergo review and update.

Purpose

Describe the factors or circumstances that mandate the existence of the policy. Also state the policy's basic objectives and what the policy is meant to achieve.

This policy is to govern the responsible use of generative AI to protect the interests of [organization] from the risks associated with the technology.

Audience

Define the target audience: the person or group of people to whom this policy is applicable.

Employees who use AI as part of their workflow.

Scope

Define to whom and to what systems this policy applies. List the employees required to comply or simply indicate "all" if all must comply. Also indicate any exclusions or exceptions (e.g. people, elements, or situations not covered by this policy or where special consideration may be made).

This policy applies to the use of open generative AI (Gen AI) systems (e.g. ChatGPT) and any AI or machine learning (ML) models or systems [organization] develops internally.

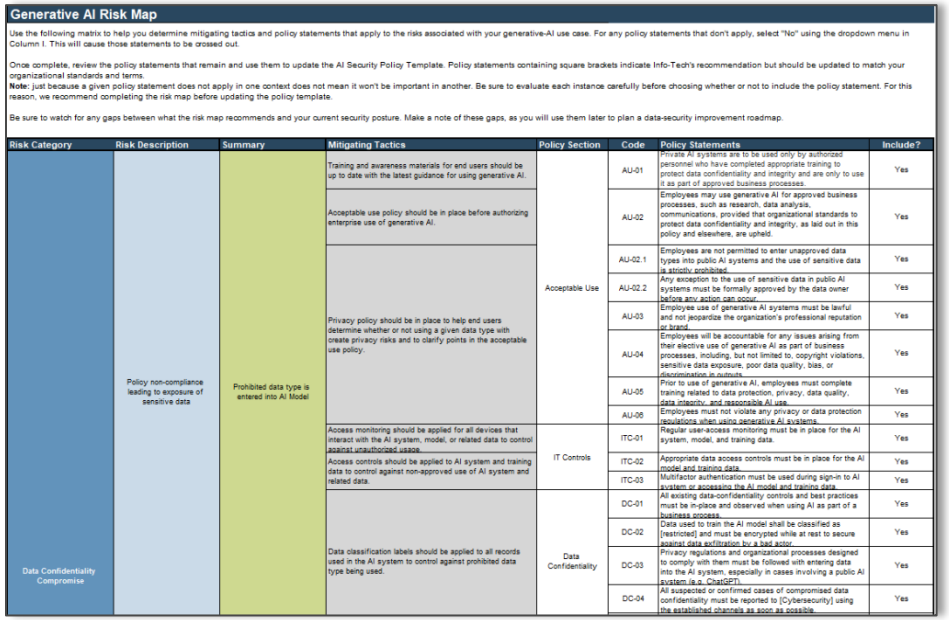
1
Info-Tech Research Group

Blueprint deliverables

Each step of this blueprint is accompanied by supporting deliverables to help you accomplish your goals.

Generative AI Risk Map

Determine the risks associated with your Gen AI use case and the applicable policy statements



Generative AI Risk Map

Use the following matrix to help you determine mitigating tactics and policy statements that apply to the risks associated with your generative-AI use case. For any policy statements that don't apply, select "No" using the dropdown menu in Column I. This will cause those statements to be crossed out.

Once complete, review the policy statements that remain and use them to update the AI Security Policy Template. Policy statements containing square brackets indicate Info-Tech's recommendation but should be updated to match your organizational standards and terms.
Note: just because a given policy statement does not apply in one context does not mean it won't be important in another. Be sure to evaluate each instance carefully before choosing whether or not to include the policy statement. For this reason, we recommend completing the risk map before updating the policy template.

Be sure to watch for any gaps between what the risk map recommends and your current security posture. Make a note of these gaps, as you will use them later to plan a data-security improvement roadmap.

Risk Category	Risk Description	Summary	Mitigating Tactics	Policy Section	Code	Policy Statements	Include?
Data Confidentiality Compromise	Policy non-compliance leading to exposure of sensitive data	Prohibited data type is entered into AI Model	Training and awareness materials for end users should be up to date with the latest guidance for using generative AI.	Acceptable Use	AU-01	Private AI systems are to be used only by authorized personnel who have completed appropriate training to protect data confidentiality and integrity and are only to use it as part of approved business processes.	Yes
			Acceptable use policy should be in place before authorizing enterprise use of generative AI.		AU-02	Employees may use generative AI for approved business processes, such as research, data analysis, communications, provided that organizational standards to protect data confidentiality and integrity, as laid out in this policy and elsewhere, are upheld.	Yes
			Privacy policy should be in place to help end users determine whether or not using a given data type with create privacy risks and to clarify points in the acceptable use policy.		AU-02.1	Employees are not permitted to enter unapproved data types into public AI systems and the use of sensitive data is strictly prohibited.	Yes
					AU-02.2	Any exception to the use of sensitive data in public AI systems must be formally approved by the data owner before any action can occur.	Yes
					AU-03	Employee use of generative AI systems must be lawful and not jeopardize the organization's professional reputation or brand.	Yes
					AU-04	Employees will be accountable for any issues arising from their elective use of generative AI as part of business processes, including, but not limited to, copyright violations, sensitive data exposure, poor data quality, bias, or discrimination in outputs.	Yes
					AU-05	Prior to use of generative AI, employees must complete training related to data protection, privacy, data quality, data integrity, and responsible AI use.	Yes
					AU-06	Employees must not violate any privacy or data protection regulations when using generative AI systems.	Yes
			Access monitoring should be applied for all devices that interact with the AI system, model, or related data to control against unauthorized usage.	IT Controls	ITC-01	Regular user-access monitoring must be in place for the AI system, model, and training data.	Yes
			Access controls should be applied to AI system and training data to control against non-approved use of AI system and related data.		ITC-02	Appropriate data access controls must be in place for the AI model and training data.	Yes
					ITC-03	Multifactor authentication must be used during sign-in to AI system or accessing the AI model and training data.	Yes
			Data classification labels should be applied to all records used in the AI system to control against prohibited data type being used.	Data Confidentiality	DC-01	All existing data-confidentiality controls and best practices must be in-place and observed when using AI as part of a business process.	Yes
					DC-02	Data used to train the AI model shall be classified as [restricted] and must be encrypted while at rest to secure against data exfiltration by a bad actor.	Yes
					DC-03	Privacy regulations and organizational processes designed to comply with them must be followed with entering data into the AI system, especially in cases involving a public AI system (e.g. ChatGPT).	Yes
					DC-04	All suspected or confirmed cases of compromised data confidentiality must be reported to [Cybersecurity] using the established channels as soon as possible.	Yes

Blueprint benefits

IT Benefits

- Improved understanding of AI-related risks and how they apply to your use cases
- Lower risks associated with near-term use of Gen AI by setting acceptable use standards
- Long-term risks associated with AI addressed via long-term security planning
- Reduced data loss incidents related to use of public Gen AI systems

Business Benefits

- Increased productivity via Gen AI
- Lower regulatory risks related to the use of Gen AI
- Defined standards for how AI can be used, avoiding additional legal risks related to copyright infringement

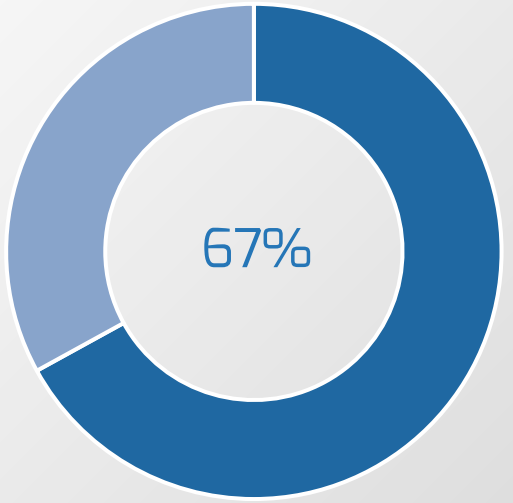
Measure the value of this blueprint

Expedite your policy and lower risk

Work to complete	Average time to complete	Info-Tech method	Time saved
Write Gen AI security policy	8 days – research risks, determine requirements, interview stakeholders, draft and revise policy	0.5 days	7.5 days

Improvement metrics	Outcome
Reduced risk of data confidentiality compromise	50% reduction of risk over a one-year period
Reduced risk of data integrity compromise	
Reduced risk of input attack	
Reduced risk of data poisoning	
Reduced risk of sponge attack	

IT leaders prioritizing Gen AI in the next 18 months



Source: Salesforce, 2023

Determine Gen AI suitability

Some things are better done the old-fashioned way

Before determining the specific security and privacy risks associated with your desired use for Gen AI (and how to address them), consider whether AI is the best way to achieve your goals.

Info-Tech's AI Suitability Test

1. What are the benefits of using Gen AI for this purpose?
2. Does the intended purpose involve entering sensitive data into the AI system?
3. Does the intended purpose incorporate Gen AI outputs into business processes or the supply chain?
4. How severe would the impact be if sensitive data were exposed?
5. How severe would the impact be if a faulty output were used?
6. Will a public or private AI system be used?
7. What alternatives exist to achieve the same goal and what drawbacks do they have?
8. Considering your answers to the above questions, how suitable is AI for the proposed purpose?

Sources: Belfer Center, 2019; "Data Privacy," Forbes, 2023

"[T]he outcomes of … AI suitability tests need not be binary. They can … suggest a target level of AI reliance on the spectrum between full autonomy and full human control. "

– Marcus Comiter, Capability Delivery Directorate at DoD Joint Artificial Intelligence Center

Activity

On tab 2 of the *Generative AI Risk Map*, complete the AI suitability test to determine the extent to which you can rely on AI for your use case.

AI suitability test

AI Suitability Test

Complete the following questionnaire to help determine the extent to which your use case can rely on AI. The final result will be self-determined, but answering these questions will help you to see areas where your use of AI may require additional oversight or where certain parts of the use case are best executed by human labor.

	Questions	Response
1	What are the benefits of using generative AI for this purpose?	
2	Does the intended purpose involve entering sensitive data into the AI system?	
3	Does the intended purpose incorporate generative AI outputs into a business processes or the supply chain?	

Download *Generative AI Risk Map*

Info-Tech Insight

Look for problems before getting invested. While Gen AI opens many possibilities, some risks will be difficult to address. For example, if your proposed use case requires sensitive data to be entered into a public AI system to produce an output for use in your supply chain, it will be virtually impossible to mitigate such risks effectively.

Assess risks for Gen AI

Risks depend on the use case

- Exactly which risk factors apply, and to what extent, will depend on your Gen AI use case, with the biggest variables being whether you're inputting data, using an output from the system, and whether the system is public or private.
- For example, asking the system to organize a data input so that you can use the output carries a lower data confidentiality risk in a private system than in a public one because the information isn't shared beyond the organization's AI system.
- However, another possible use case is asking a public AI system to generate a data set by compiling industry statistics, which carries virtually no input-related risk, but has significant data quality/integrity risk because the system may have used unknown or even fictitious sources.

"All AI models generate text based on training data and the input they receive. Companies may not have complete control over the output, which could potentially expose sensitive or inappropriate content during conversations. Information inadvertently included in a conversation with a Gen AI presents a risk of disclosure to unauthorized parties."

– Eric Schmitt, Global Chief Information Security Officer, Sedgwick

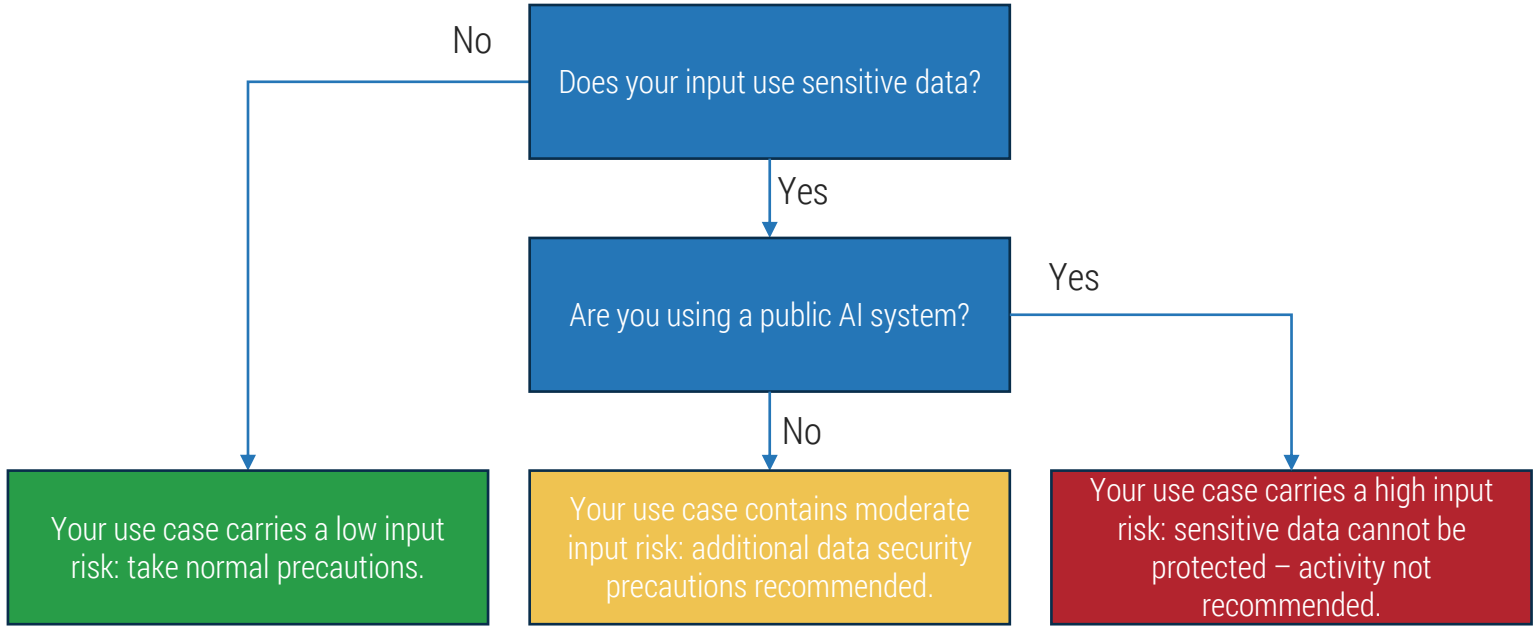
Info-Tech Insight

Watch for overlap. There will usually be both an input and an output component when using Gen AI, which means both risk factors are present, but one may be dominant. Therefore, both inputs and outputs should receive sign-off before use to limit data confidentiality and integrity risks.

Input risks

Risks depend on the use case

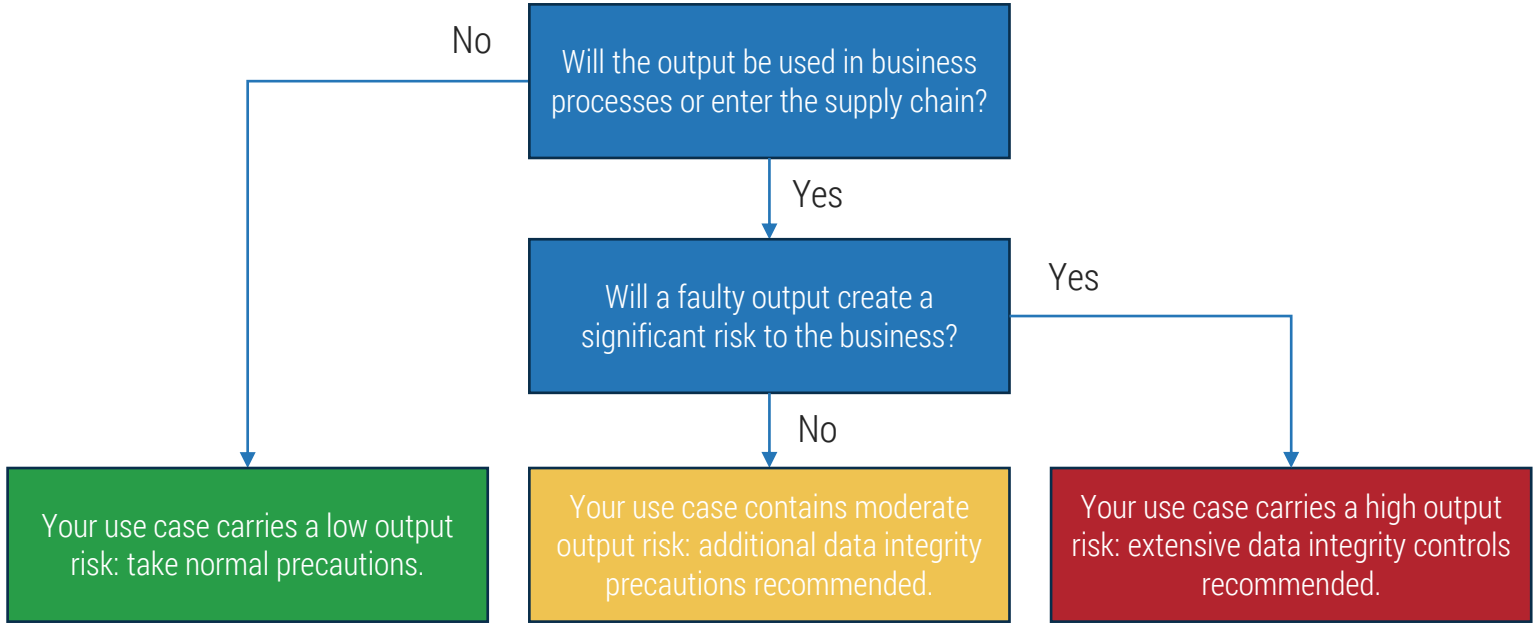
If your use case involves entering data into an AI system, the greatest risk is exposure of sensitive data, resulting in a data confidentiality compromise that may break privacy regulations or place intellectual property or trade secrets at risk.



Output risks

Risks depend on the use case

The greatest risk involved with Gen AI output is that it may be low-quality. If not verified (and corrected), this data may suffer from bias, inaccuracies, or other issues that may degrade data quality, eventually leading to data integrity loss from the number of errors it contains.

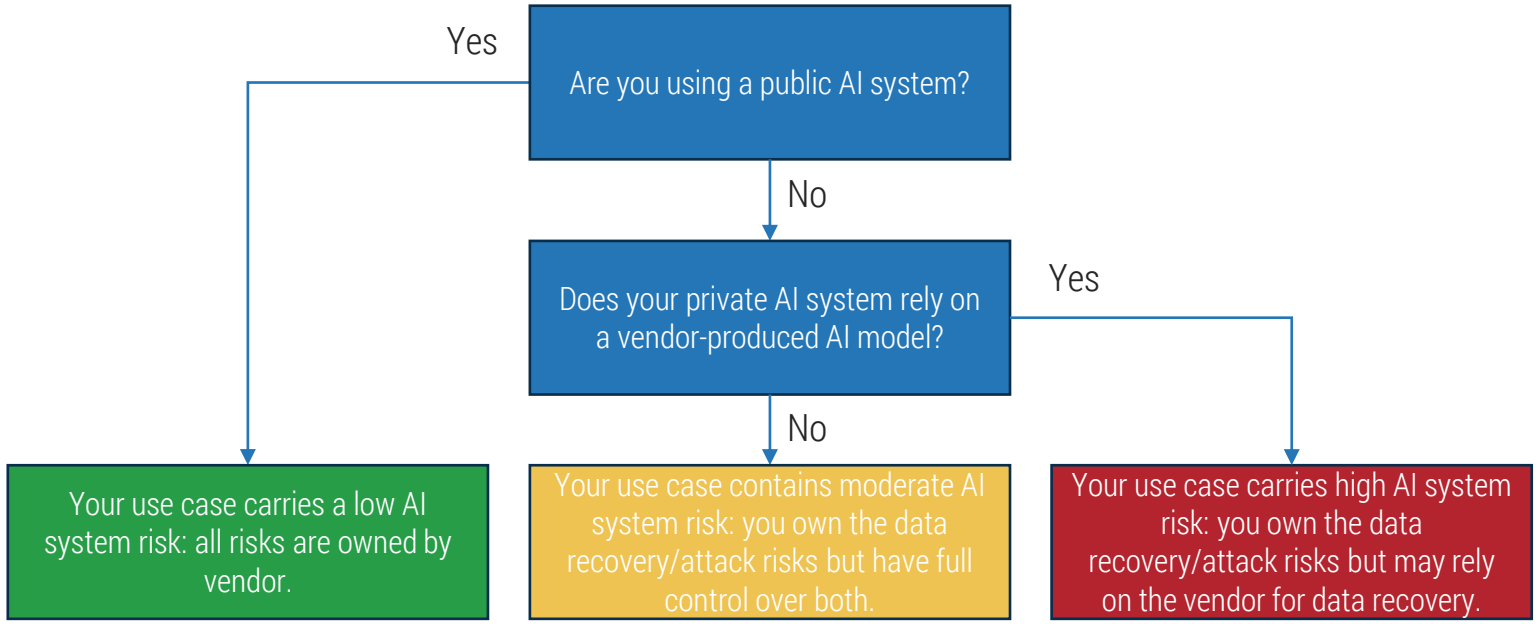


AI system risks

Risks depend on the use case

The major risks to an AI system are related to rebuilding the model in the event of a compromise (i.e. data recovery) and the system itself being attacked.

In a public AI system these risks are assumed by the vendor. In a private system, data risks may be jointly held (e.g. if a model is purchased from a vendor) or fully owned by the organization that built the model, though attack risks are owned by the organization.



Info-Tech Insight

Vendor-owned models can complicate rebuilds, as you may rely on the vendor to provide the data, which may be outdated.

Attacks on Gen AI

The threat landscape is evolving

Info-Tech Insight

Build a strong perimeter around your AI system and data. AI security is still in its early stages and best practices are still being determined. Until more specific controls and techniques are developed, the best course of action is to use a robust data security program to make your sensitive data as difficult to access as possible, and to monitor for intrusions.

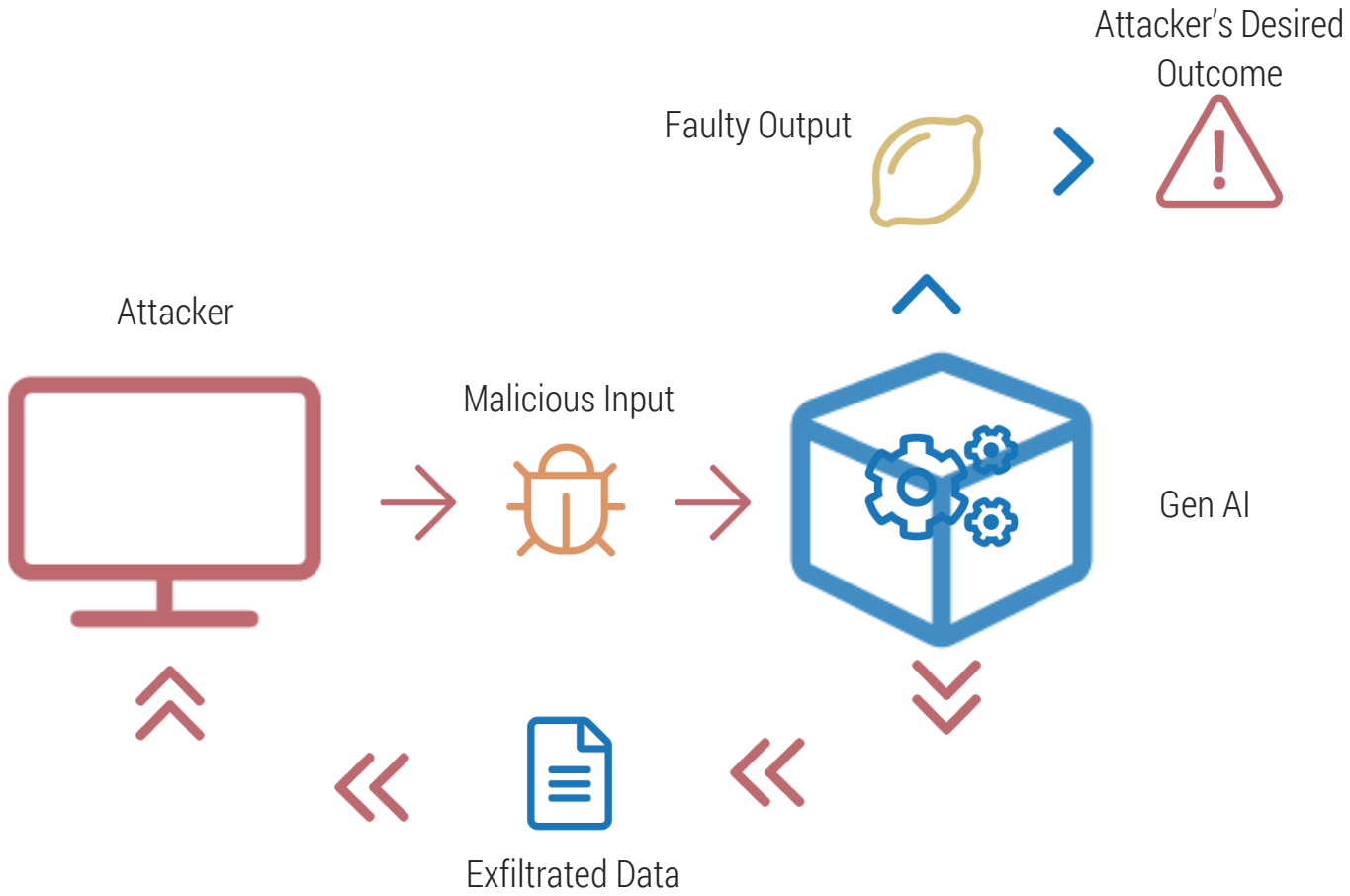
"[F]or AI attacks, a robust IT department and 90-letter passwords won't save the day. The algorithms themselves have the inherent limitations that allow for attack. Even if an AI model is trained to exacting standards using data and algorithms that have never been compromised, it can still be attacked. This bears repeating: among the state-of-the-art methods, there is currently no concept of an 'unattackable' AI system."

– Marcus Comiter, Capability Delivery Directorate at DoD Joint Artificial Intelligence Center

Input and exfiltration attacks

Attacks on Gen AI

- Input attacks are any adversarial action taken against an AI system by compromising data the system ingests and responds to.
- This type of attack appears in various forms, though not all apply to Gen AI systems.
- Input attacks are often preceded by an exfiltration attack. However, exfiltration may occur independently of a desire to harm the AI system itself.



Input and exfiltration attacks

Attacks on Gen AI

Prompt Injection

Using cleverly written prompts to make an AI system comply with prohibited requests.

- Common attack against Gen AI using a large language model
- Acceptable use policy and user monitoring recommended

Evasion

Using an input that the AI system misinterprets, causing it to malfunction (i.e. go against training).

- Often used against AI systems designed for image or pattern recognition
- Rarely a significant risk for Gen AI, unless the system combines language and image processing

Data Exfiltration

Stealing data to better understand how an AI system works (e.g. training data).

- Often precedes an input attack so that it can be better executed
- A risk for all types of AI but can be mitigated using standard data protection techniques

Inversion Attacks

A process of using inputs and measuring outputs to determine if they contain sensitive information about the model or training data.

- Goal is to rebuild the AI model or the data it contains via outputs
- Can affect any AI system that includes sensitive data, especially if it is included in outputs

AI Model Theft

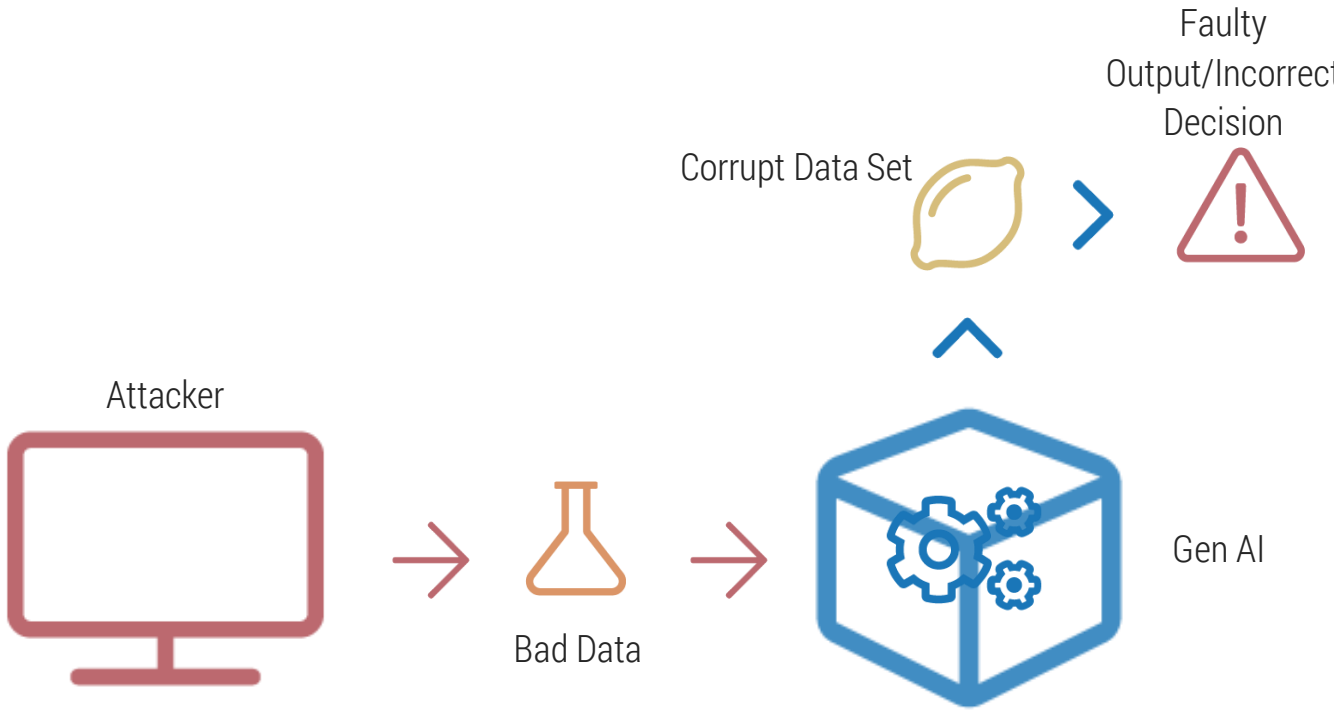
Stealing the file containing the AI model.

- May accompany an input attack or may be motivated by other factors, as in other forms of data exfiltration
- Strong data protection controls should be used to create a perimeter around AI system

Data poisoning

Attacks on Gen AI

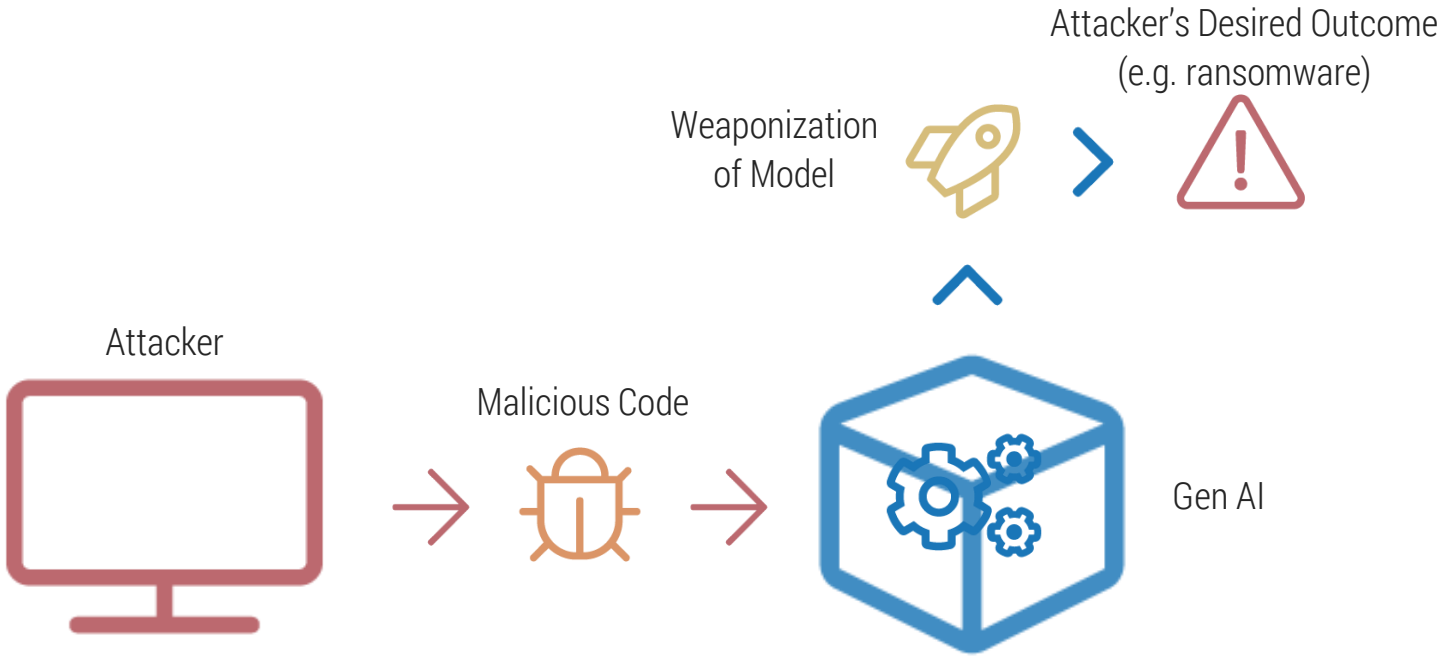
- Data poisoning occurs when an attacker gains access to AI model training data and alters or adds data with the intention of corrupting it, making its performance unreliable, potentially causing significant and even dangerous errors in its outputs.
- The greatest risk is that the attack goes unnoticed, resulting in an operational model with unknown flaws.
- The best defense is a strong perimeter around the AI model, complete with encryption and intrusion detection systems where possible.
- Rebuilding the model is the only sure recovery method, meaning reliable backups should be available.



Weaponization of AI model

Attacks on Gen AI

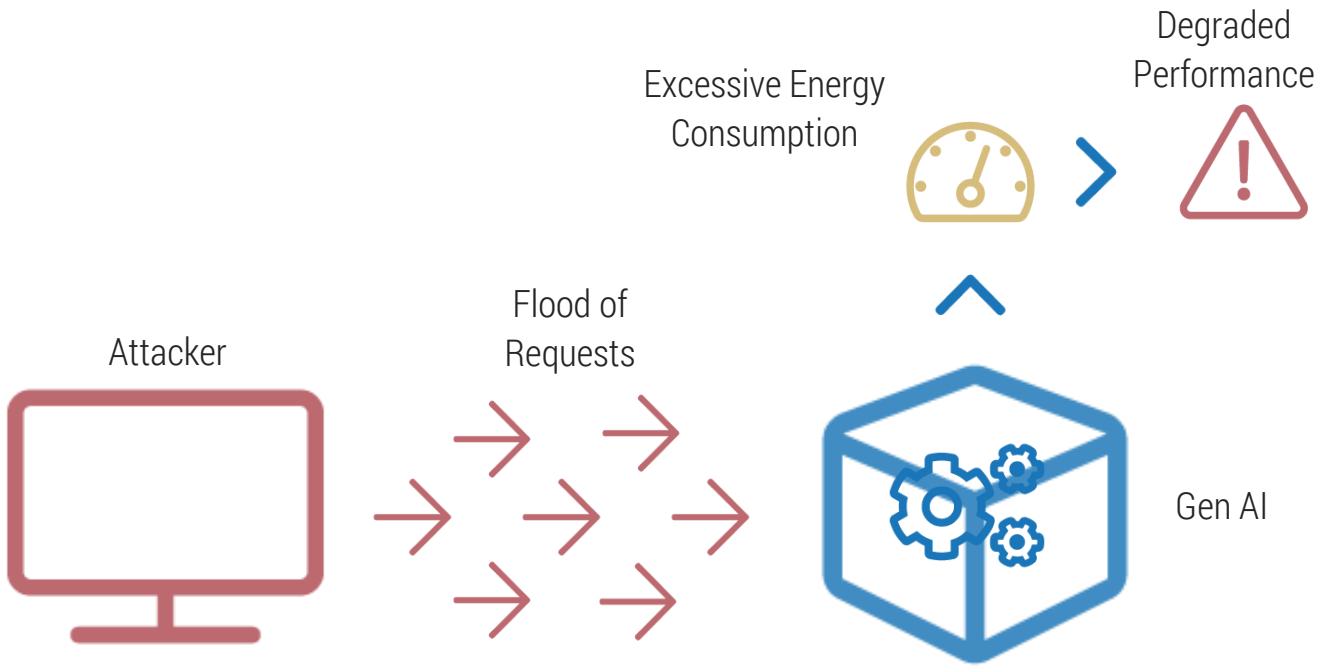
- Weaponization refers to an attack where a bad actor embeds malicious code in the AI model, using it as the vector to launch another attack, such as ransomware.
- Knowledge of AI model, system, and training data should be accessed on a need-to-know basis.
- Unverified code should not be incorporated into the AI system.



Sponge attacks

Attacks on Gen AI

- Sponge attacks are like DoS attacks. They involve flooding the AI system with more requests than it can handle to degrade performance and cause excessive energy consumption.
- The best defense is to configure the AI system to reset if a given energy consumption threshold is reached.



AI-assisted cyberattacks

The barrier to entry just got lower

- When a bad actor uses Gen AI to launch a cyberattack, the greatest risks occur when they write malware and craft highly realistic phishing materials (e.g. natural-sounding writing, audio-visual deepfakes).
- Public AI systems are typically trained not comply with requests that are illegal or unethical. However, this safeguard can be overcome by changing the way the AI system's help is requested (e.g. prompt injection – a type of input attack).

Info-Tech Insight

Risk likelihood just went up. The use of Gen AI to facilitate cyberattacks doesn't fundamentally change the nature of the risk. But because Gen AI makes the process easier, we should account for this in our risk assessments.

"It's irrelevant whether an attack was developed using generative AI or not. An exploit is an exploit, and an attack is an attack, regardless of how it was created."

– Tony Bradley, Editor-in-Chief, TechSpective, in "Defending," Forbes, 2023

Activity

- Using the information contained in slides 20-30, determine which risks apply to your Gen AI use case (i.e. data confidentiality, data integrity, and attacks on the AI system).
- Review this risk map on tab 3 to better understand how those risks are realized and to determine what mitigating tactics and policy statements are required.
- Note any key mitigating tactics or policy statements that are not currently well represented in your security program.

Leverage Info-Tech's *Generative AI Risk Map*

Generative AI Risk Map

Use the following matrix to help you determine mitigating tactics and policy statements that apply to the risks associated with your generative-AI use case. For any policy statements that don't apply, select "No" using the dropdown menu in Column I. This will cause those statements to be crossed out.

Once complete, review the policy statements that remain and use them to update the AI Security Policy Template. Policy statements containing square brackets indicate Info-Tech's recommendation but should be updated to match your organizational standards and terms.
Note: just because a given policy statement does not apply in one context does not mean it won't be important in another. Be sure to evaluate each instance carefully before choosing whether or not to include the policy statement. For this reason, we recommend completing the risk map before updating the policy template.

Be sure to watch for any gaps between what the risk map recommends and your current security posture. Make a note of these gaps, as you will use them later to plan a data-security improvement roadmap.

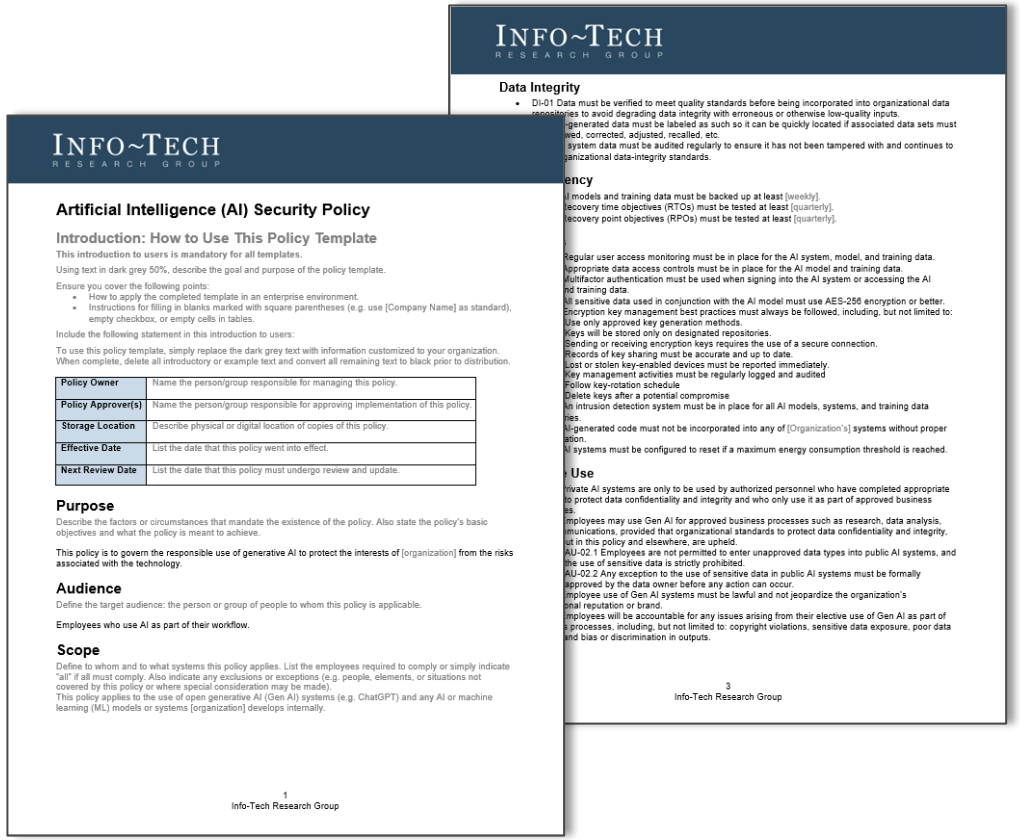
Risk Category	Risk Description	Summary	Mitigating Tactics	Policy Section	Code	Policy Statements	Include?
Data Confidentiality Compromise	Policy non-compliance leading to exposure of sensitive data	Prohibited data type is entered into AI Model	Training and awareness materials for end users should be up to date with the latest guidance for using generative AI.	Acceptable Use	AU-01	Private AI systems are to be used only by authorized personnel who have completed appropriate training to protect data confidentiality and integrity and are only to use it as part of approved business processes.	Yes
			Acceptable use policy should be in place before authorizing enterprise use of generative AI.		AU-02	Employees may use generative AI for approved business processes, such as research, data analysis, communications, provided that organizational standards to protect data confidentiality and integrity, as laid out in this policy and elsewhere, are upheld.	Yes
			Privacy policy should be in place to help end users determine whether or not using a given data type with create privacy risks and to clarify points in the acceptable use policy.		AU-02.1	Employees are not permitted to enter unapproved data types into public AI systems and the use of sensitive data is strictly prohibited.	Yes
					AU-02.2	Any exception to the use of sensitive data in public AI systems must be formally approved by the data owner before any action can occur.	Yes
					AU-03	Employee use of generative AI systems must be lawful and not jeopardize the organization's professional reputation or brand.	Yes
					AU-04	Employees will be accountable for any issues arising from their elective use of generative AI as part of business processes, including, but not limited to, copyright violations, sensitive data exposure, poor data quality, bias, or discrimination in outputs.	Yes
					AU-05	Prior to use of generative AI, employees must complete training related to data protection, privacy, data quality, data integrity, and responsible AI use.	Yes
					AU-06	Employees must not violate any privacy or data protection regulations when using generative AI systems.	Yes
			Access monitoring should be applied for all devices that interact with the AI system, model, or related data to control against unauthorized usage.	IT Controls	ITC-01	Regular user-access monitoring must be in place for the AI system, model, and training data.	Yes
			Access controls should be applied to AI system and training data to control against non-approved use of AI system and related data.		ITC-02	Appropriate data access controls must be in place for the AI model and training data.	Yes
					ITC-03	Multifactor authentication must be used during sign-in to AI system or accessing the AI model and training data.	Yes
			Data classification labels should be applied to all records used in the AI system to control against prohibited data type being used.	Data Confidentiality	DC-01	All existing data-confidentiality controls and best practices must be in-place and observed when using AI as part of a business process.	Yes
					DC-02	Data used to train the AI model shall be classified as [restricted] and must be encrypted while at rest to secure against data exfiltration by a bad actor.	Yes
					DC-03	Privacy regulations and organizational processes designed to comply with them must be followed with entering data into the AI system, especially in cases involving a public AI system (e.g. ChatGPT).	Yes
					DC-04	All suspected or confirmed cases of compromised data confidentiality must be reported to [Cybersecurity] using the established channels as soon as possible.	Yes

Download *Generative AI Risk Map*

Activity

- After determining the applicable policy statements, update this policy template by deleting the ones that don't apply.
- Each policy statement is cross-referenced using the code provided in the risk map.

Update the *AI Security Policy Template*



INFO-TECH RESEARCH GROUP

Artificial Intelligence (AI) Security Policy

Introduction: How to Use This Policy Template

This introduction to users is mandatory for all templates.

Using text in dark grey 50%, describe the goal and purpose of the policy template.

Ensure you cover the following points:

- How to apply the completed template in an enterprise environment.
- Instructions for filling in blanks marked with square parentheses (e.g. use [Company Name] as standard), empty checkbox, or empty cells in tables.

Include the following statement in this introduction to users:

To use this policy template, simply replace the dark grey text with information customized to your organization. When complete, delete all introductory or example text and convert all remaining text to black prior to distribution.

Policy Owner	Name the person/group responsible for managing this policy.
Policy Approver(s)	Name the person/group responsible for approving implementation of this policy.
Storage Location	Describe physical or digital location of copies of this policy.
Effective Date	List the date that this policy went into effect.
Next Review Date	List the date that this policy must undergo review and update.

Purpose

Describe the factors or circumstances that mandate the existence of the policy. Also state the policy's basic objectives and what the policy is meant to achieve.

This policy is to govern the responsible use of generative AI to protect the interests of [organization] from the risks associated with the technology.

Audience

Define the target audience: the person or group of people to whom this policy is applicable.

Employees who use AI as part of their workflow.

Scope

Define to whom and to what systems this policy applies. List the employees required to comply or simply indicate "all" if all must comply. Also indicate any exclusions or exceptions (e.g. people, elements, or situations not covered by this policy or where special consideration may be made).
This policy applies to the use of open generative AI (Gen AI) systems (e.g. ChatGPT) and any AI or machine learning (ML) models or systems [organization] develops internally.

1
Info-Tech Research Group

INFO-TECH RESEARCH GROUP

Data Integrity

- DI-01 Data must be verified to meet quality standards before being incorporated into organizational data

3
Info-Tech Research Group

Download *AI Security Policy Template*

Activity

- As a final step, use a whiteboard to brainstorm a list of proposed program improvements based on the gaps noted while completing the risk map exercise.
- Be sure to prioritize improvements using an effort-to-benefit evaluation, targeting the improvements that will provide the greatest boost to your security program maturity.

Plan for future improvements

Initiative Planning

Quarter 1

1. Update privacy policy to include AI use
2. Expand use of intrusion detection system (IDS) to include AI system and training data

Quarters 2-3

1. Implement AI-produced content verification process
2. Design training and awareness materials to address AI risks

Quarter 4

1. Create AI backup plan

Research Contributors and Experts

Birye Abebe
Chief Information Officer
Armed Forces Benefit Association

Charles Beierle
Chief Information Officer
Randolph Brooks Federal Credit Union

Todd Heinz
Practice Manager – Governance, Risk, and Compliance
Heartland Business Systems

Trenton Schuttler
Co-Owner
TBI IT

Related Info-Tech Research

[Secure Your High-Risk Data]

[Build a Data Privacy Program]

[Develop Responsible AI Guiding Principles]

Bibliography

Bradley, Tony. "Defending Against Generative AI Cyber Threats." *Forbes*, 27 Feb. 2023, https://www.forbes.com/sites/tonybradley/2023/02/27/defending-against-generative-ai-cyber-threats/?sh=154fd0310884Defending Against Generative AI Cyber Threats (forbes.com). Accessed 26 May 2023.

Comiter, Marcus. "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It." *Belfer Center for Science and International Affairs*, Aug 2019, https://www.belfercenter.org/publication/AttackingAIAttacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It | Belfer Center for Science and International Affairs. Accessed 29 May 2023.

Daniels, Jodi. "How Generative AI Can Affect Your Business' Data Privacy." *Forbes*, 1 May 2023, https://www.forbes.com/sites/forbesbusinesscouncil/2023/05/01/how-generative-ai-can-affect-your-business-data-privacy/?sh=6bc7c63e702dHow Generative AI Can Affect Your Business' Data Privacy (forbes.com). Accessed 17 May 2023.

"IT Leaders Call Generative AI a 'Game Changer' but Seek Progress on Ethics and Trust." *Salesforce*, 6 Mar. 2023,https://www.salesforce.com/news/stories/generative-ai-research/?d=cta-body-promo-8. Accessed 15 May 2023.

Jackson, Terrance. "Exploring the Security Risks of Generative AI." *Forbes*, 19 Apr. 2023, https://www.forbes.com/sites/forbestechcouncil/2023/04/19/exploring-the-security-risks-of-generative-ai/?sh=a73609735942Exploring The Security Risks Of Generative AI (forbes.com). Accessed 15 May 2023.

"KPMG U.S. Survey: Executives Expect Generative AI to Have Enormous Impact on Business, But Unprepared for Immediate Adoption." *KPMG*, Mar. 2023, https://info.kpmg.us/news-perspectives/technology-innovation/kpmg-generative-ai-2023.htmlKPMG Generative AI Survey. Accessed 15 May 2023.

Lawton, George. "Multimodal AI," *TechTarget*, 22 May 2023, https://www.techtarget.com/searchenterpriseai/definition/multimodal-AI. Accessed 13 Jul. 2023.

Lemos, Robert. "Adversarial AI Attacks Highlight Fundamental Security Issues." *Dark Reading*, 22 Nov. 2022, https://www.darkreading.com/vulnerabilities-threats/adversarial-ai-attacks-highlight-fundamental-security-issues. Accessed July 4, 2023.

NIST. *NIST AI 100-01: Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. *NIST*, 26 Jan. 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. Accessed 13 Jul. 2023

Schmitt, Eric. "Understanding The Risks of Generative AI for Better Business Outcomes," *Venture Beat*, 29 Apr. 2023, https://venturebeat.com/ai/understand-risks-generative-ai-better-business-outcomes/. Accessed 2 May 2023.

Bibliography

Seifried, Kurt et al. "Security Implications of ChatGPT." *CSA*, 23 Apr. 2023, https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt/. Accessed 17 May 2023.

"You'll Probably Need a ChatGPT Company Policy." *Legal.io*, 6 Apr. 2023, https://www.legal.io/articles/5429675/You-ll-Probably-Need-a-ChatGPT-Company-PolicyYou'll Probably Need a ChatGPT Company Policy. Accessed 25 May 2023.


INFO~TECH
RESEARCH GROUP