FAIR INSTITUTE

FAIR-AIR Approach Playbook

USING A FAIR-BASED RISK APPROACH TO EXPEDITE AI ADOPTION AT YOUR ORGANIZATION

Your Why

Right now, the disconnect between the security organization and the business is becoming clear once again. With the adoption of AI, the business wants to move faster and does not necessarily want to involve the security team, as they are seen in many cases as an impediment to the process. The best way to ensure AI adoption is handled in a risk-based, secure way is by using the same language as the business, based on analysis with the proven FAIR™ methodology for cyber risk quantification, **With the FAIR-AIR Approach, you will be able to speak the same language as the business and work as a partner in AI adoption rather than an impediment.**

Your Squad

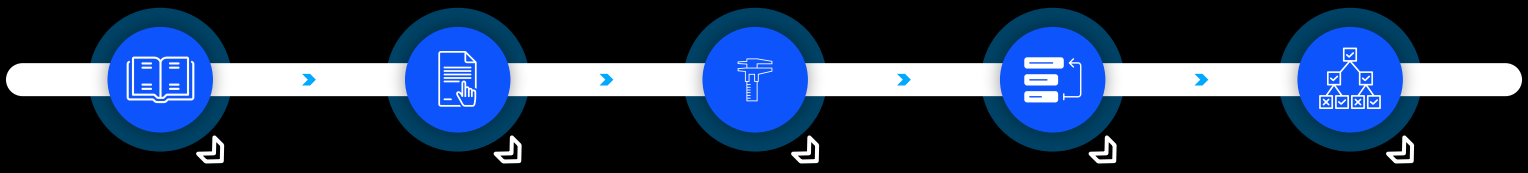
- Data Privacy Officer and team
- Privacy Officer and team
- Chief Information Security Officer and team
- Technology leadership and business leadership

These teams are important to work with to ensure that all AI processes are aligned and working towards the same goal of enhancing the business through its use and doing so with a responsible AI approach in mind – including cybersecurity.

Your Approach

The FAIR-AIR Approach for Generative AI Risk quantifies in financial terms the risk associated with AI so that you can speak the same language as the business and help your stakeholders properly prioritize AI security around other investment decisions.

Follow these steps to achieve your ultimate goal of making risk-based decisions around the secure deployment of AI.



CONTEXTUALIZE

- Understand what you're quantifying
- Identify vectors of AI risk
- Decide what risks you're trying to mitigate

SCOPE

- Visualize risk scenarios within chosen vector
- Identify the attack surface, threat actor, method of attack, and impact of threat on your asset
- Create a risk statement

QUANTIFY

- Analyze data and quantify your risk
- Review outcomes within 'loss event frequency' and 'loss Magnitude

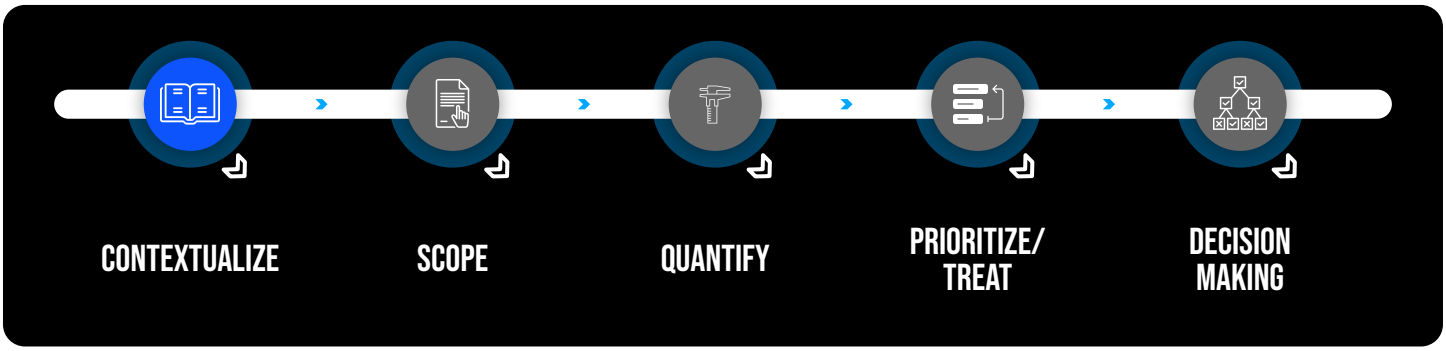
PRIORITIZE/ TREAT

- Identify results from quantification scenarios
- Pinpoint mitigation options with the largest impact
- Understand how to treat

DECISION MAKING

- Gather all quantified data, and treatment options
- Decide plan for execution and tools based on threat impact

Contextualize



In this first phase you will want to understand the why behind the risk analysis you are doing. Who are you doing an assessment for, what decisions are they trying to make? Why are they trying to make these decisions?

We have identified 5 vectors of GenAI risk to help guide this contextualization conversation. Choose the vector most relevant to you.

Shadow GenAI

- You are using Generative AI, and you just don't know it.

Foundational LLM

- You are building LLM(s) for use cases.

Hosting on LLMs

- You are hosting an LLM and using it to develop use cases.

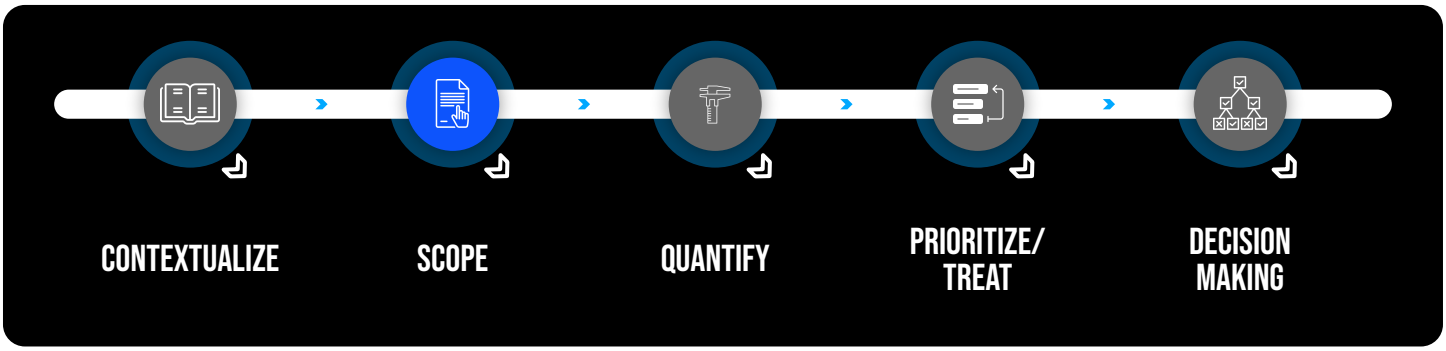
Managed LLMs

- You are using a third party LLM to develop use cases

Active Cyber Attack

- Your adversaries are using LLMs to attack you

Scope



Identify risk scenarios to quantify

Focus on your most relevant vector and think through which risk scenarios to quantify, based on the priorities of the business. In this phase, you will work to identify your assets, the relevant threats, and effects. After you understand the key assets within your chosen AI vector you can properly scope the scenarios used to quantify your risk.
Here are some sample scenarios for each of the 5 vectors.

Shadow GenAI, employees are:

- Leaking company-sensitive information via an open-source LLM Models such as ChatGPT.
- Using LLMs to mine data to enhance a job duty specific to a use case – think Meta Pixel.
- Using LLMs to complete a use case-specific job duty that is mission critical, and the LLM provides input to the process that is inaccurate or inappropriate.
- Using LLMs to complete a use case-specific job duty and the LLM goes offline, causing an outage to a mission-critical process – think no business continuity plan (BCP).

Foundational LLMs, employees are

- Not using safety controls leading to unexpected output for the LLM.
- Using training data that they have not gotten the appropriate permissions to use to train the model.
- Training the model without using the appropriate safeguards to ensure the integrity of the outcomes is not corrupted by bias with data inputs.

Hosting on LLMs, employees have not

- Defined success criteria leading to integrity issues with model output.
- Tuned the model to the use case, leading to unexpected outputs.

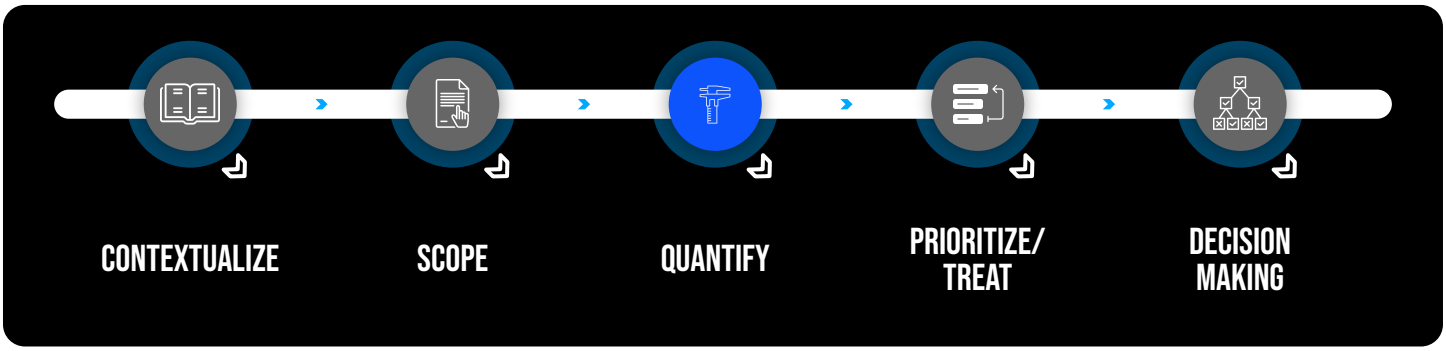
Managed LLMs

- Employees have not properly defined success criteria, leading to integrity issues with model output.
- Third Party LLM provider has not employed proper security controls and your sensitive data is leaked via prompt injection.

Active Cyber Attack, adversaries are

- Leveraging LLMs to enhance phishing attacks, leading to a breach of sensitive data.
- Using LLM's in order to discover zero-day vulnerabilities that can be exploited via ransomware or malware.

Quantify



Apply Factor Analysis of Information Risk (FAIR)

Analyze the scoped risks with FAIR and FAIR MAM to quantify how much risk (or loss exposure) exists. To do this you will need to collect internal data and industry data in order to complete this analysis. Below are some examples of how quantified scenarios would look:

Shadow GenAI

- There is a 5% probability in the next year that Employees will leak company-sensitive information via an open-source LLM Model (like chat GPT), which will lead to $5 million dollars of losses.

Foundational LLMs

- There is a 2% probability that employees will not use safety controls in the coming year, leading to unexpected output for the foundational LLM and $10 million in losses.

Hosting on LLMs

- There is a .01% probability that employees have not properly defined success criteria, leading to integrity issues with model output that result in $50 million dollars in losses.

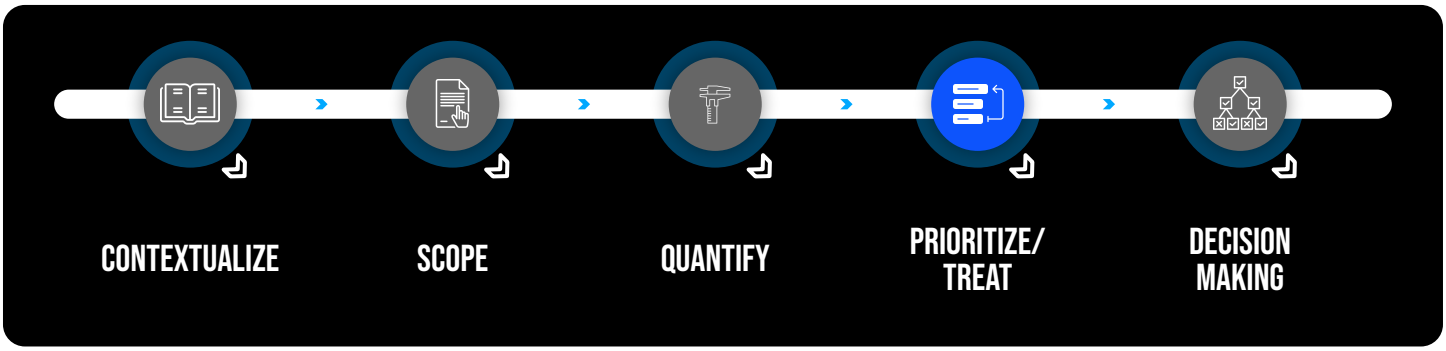
Managed LLM

- Over the next year, there is a 5% probability that a third-party LLM provider will not employ proper security controls. Your sensitive data will be leaked via prompt injection, leading to $200 million in losses.

Active Cyber Attack

- There is a 40% probability that adversaries will use LLMs to enhance phishing attacks, leading to a breach of sensitive data and resulting in losses of $350 million.

Prioritize / Treat



Clarify your path to decision-making

Prioritize the scenarios based on the probability of occurrence and dollar value of probable losses, then identify the key risk drivers for the scenarios to inform your decision on risk treatment. These key risk drivers will give you an understanding of how you can treat the identified risk and prioritize control implementation.

Active Cyber Attack

- Scenario: Enhanced phishing attacks lead to breach
- 40% probability, $350 million loss
- Key risk driver: phishing click rate amongst employees with access to large amounts of sensitive data.

Managed LLMs

- Scenario: Improper security controls
- Analysis result: 5% probability, $200 million loss
- Key risk driver: Amount of data fed to the third party LLM.

Hosting on LLMs

- Scenario: Improper success criteria
- Analysis result: .01% probability, $50 million loss
- Key risk driver: Undefined success criteria and the number of users for said use case.

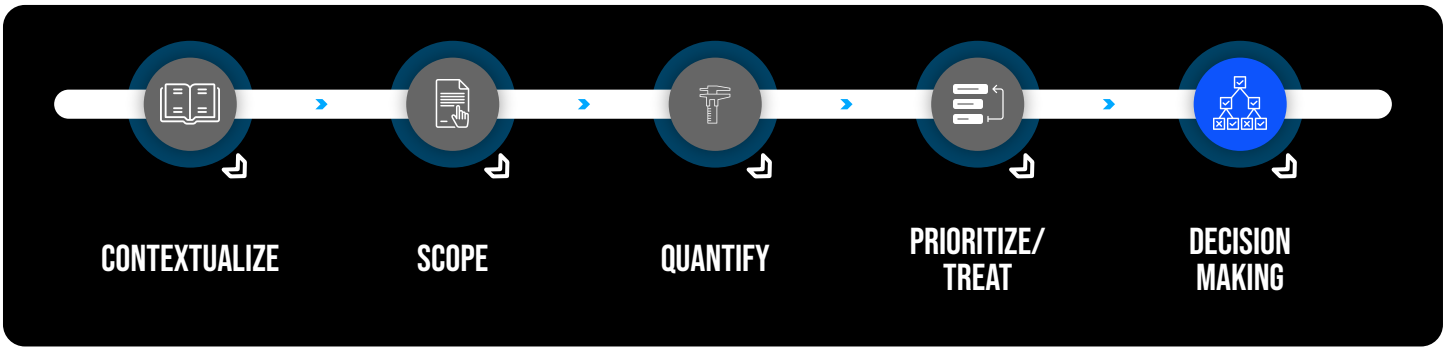
Foundational LLMs

- Scenario: Employees not using safety controls
- Analysis result: 5% probability, $5 million loss
- Key risk driver: Lack of safety controls for the model

Shadow GenAI

- Scenario: Employees leaking company sensitive information
- Analysis result: 5% probability, $5 million loss
- Key risk driver: Number of employees with exceptions to access Open AI and access to sensitive information

Decision Making



Target investments for risk reduction

In this step, you use the data points to determine what scenarios you should prioritize for treatment based on where investments will have the largest impact. You will use this data to answer the questions you collected during the contextualize phase and ensure you are able to provide the decision makers with the information needed to take a responsible AI approach.

An example from above might be: On the surface, you can remediate the most risk by employing security controls for the Active Cyber Attack scenario to reduce the phishing click rate amongst employees with access to sensitive information. However, depending on how effective those controls are, you might reduce risk by employing security controls around key risk drivers in other situations where the controls would be more effective.

Your Takeaways

1. The FAIR-AIR Approach will help you identify your AI loss exposure and make risk-based decisions on how to treat your identified loss event scenarios.
2. You will need to work across teams to ensure proper data and alignment for scenarios and use cases.
3. The purpose of this approach is to meet the business needs, not create additional obstacles to AI deployment.

The FAIR Institute

www.fairinstitute.org

info@fairinstitute.org