Enhance AI Safety in LY Corporation

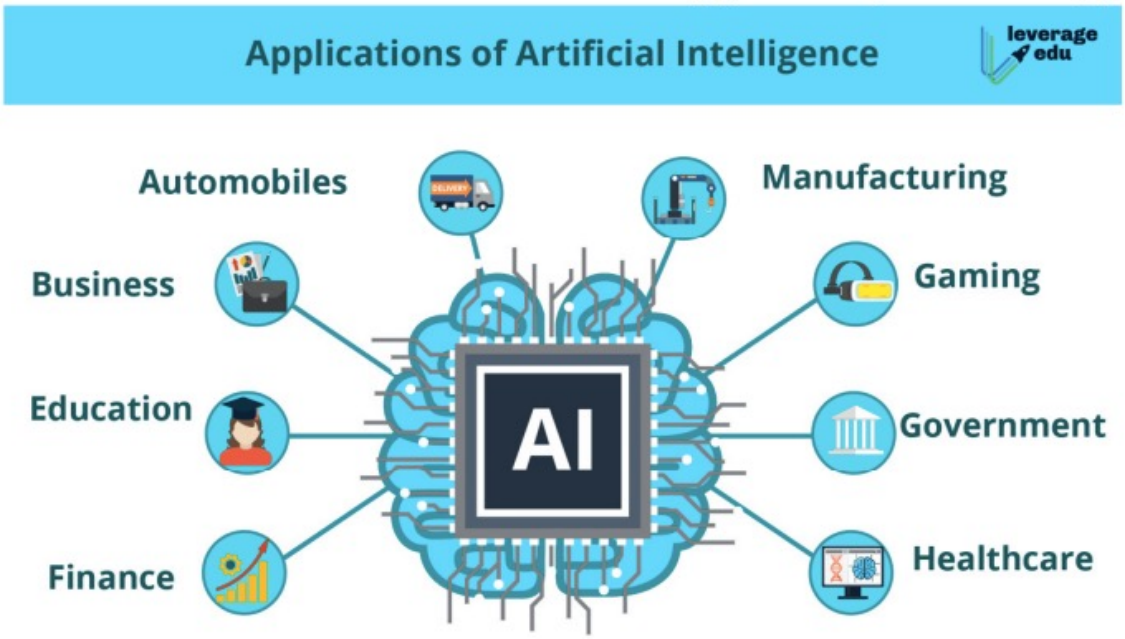
21093962D JIANG Guanlin
21080003D IU LAM AH
21096819D QIUSiyi
21100038d YUAN Yunchen
21105368dD FANG Yuji

CONTENTS


Current State
PART
1

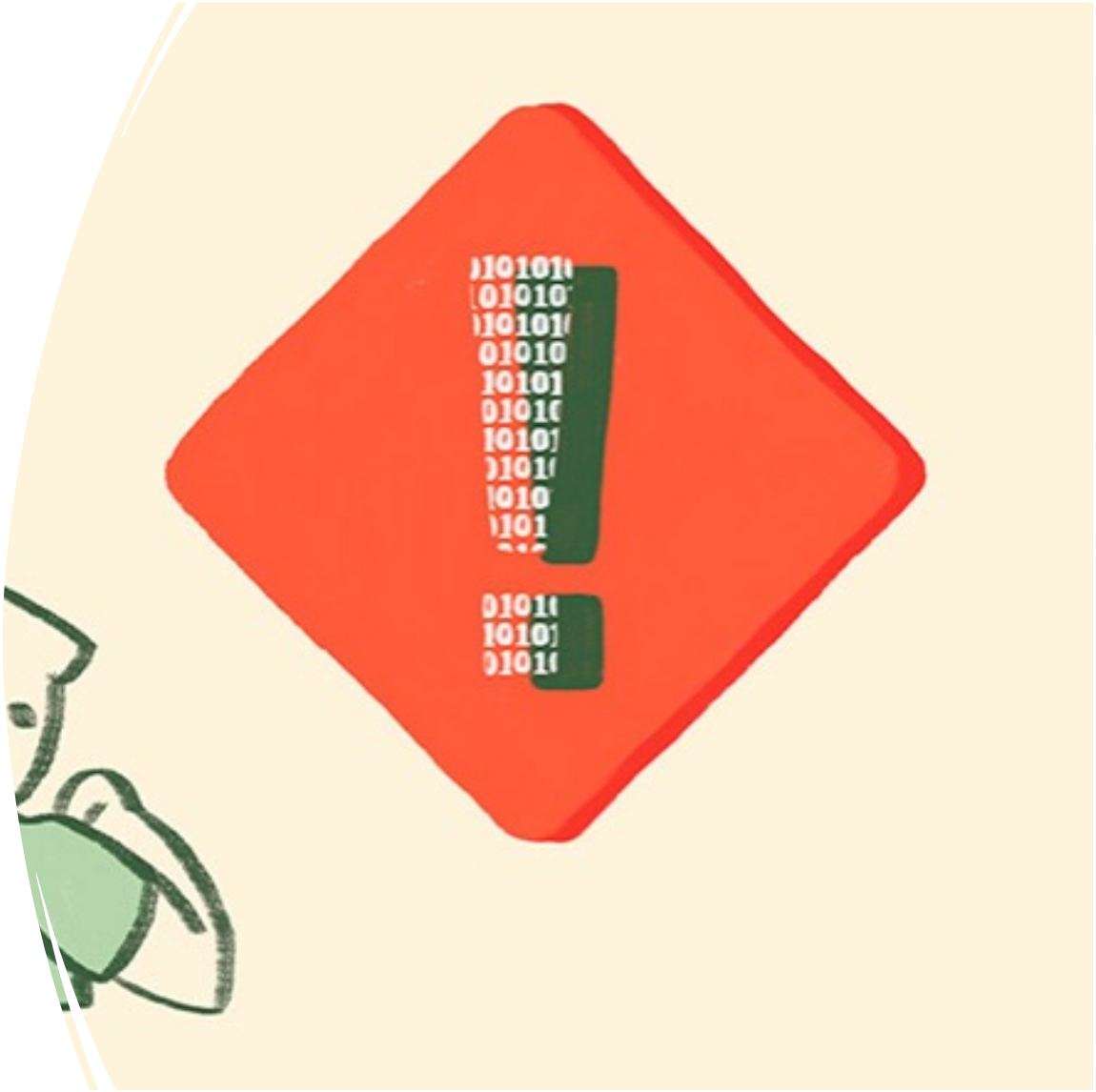
AI & AI Safety

- Recent Years, AI is using in many areas
- AI Technology cannot be prefect
- AI Safety is important



Current AI Safety Problem: Unreliability

- Decision-making processes being opaque
- hard to understand or explain
- prone to hallucination

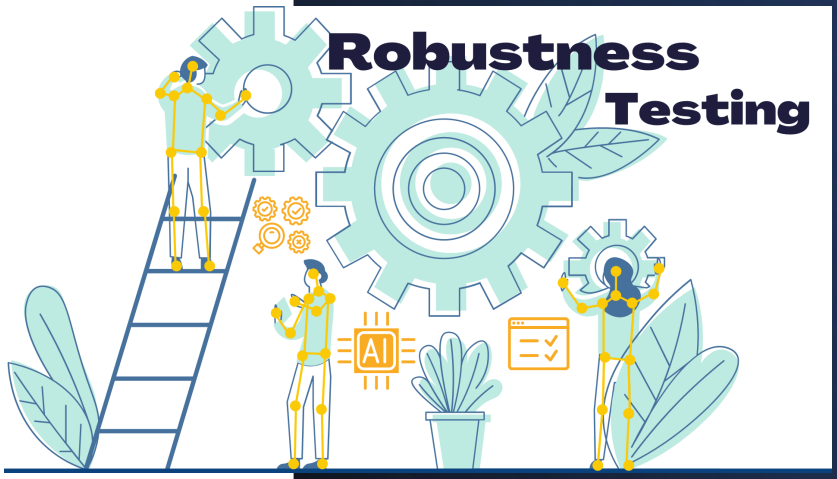


Current AI Safety Problem: Robustness

- The Challenge of Adversarial Attacks
- *(Wrong Inputs &Wrong Outputs)*

- Transfer learning and generalization ability issues

(Different environments lead to different performances)

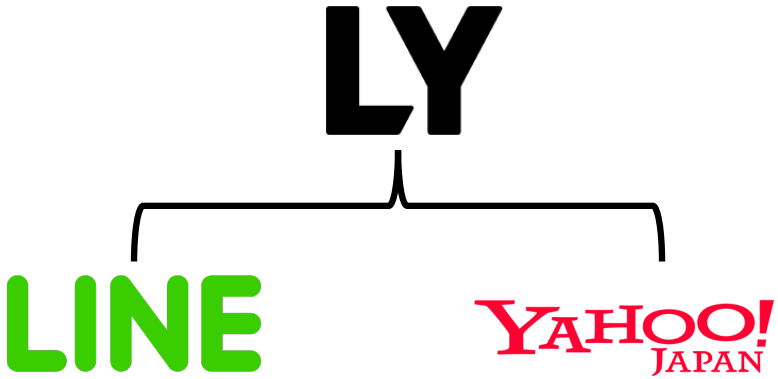


About LY Corporation

LY Corporation - Line & Yahoo Japan Corporation (Before: Z Holdings)

- An Internet Company in Japan
- **Products**
    - Yahoo Japan Search
    - Line Products
    - Yahoo Mobile Network Operators

- **History**
    - Before 2023 October – Z Holdings (Line and Yahoo JP are separate)
    - After 2023 October – LY Corporation



Current Actions for Enhance AI Safety in LY Corporation

- Introduce AI policies inside the company
- Control systems that control and manage inputs and outputs
- Employee AI safety briefing

Source: https://linecorp.com/tw/pr/news/en/2023/4636

Japan Government AI Regulation

- **AI Regulation Standards**
  - JIS X22989 (Same as International Standard - ISO/IEC 22989)
  - JISQ 38507
  - Machine Learning Quality Management Guideline
- **Focus on**
  - Data Privacy
  - The ethical use of AI

Source: https://www.linkedin.com/pulse/overview-regulation-ai-japan-ray-proper-cissp-rlzrc/

Current Solution

- Strict internal regulations and control measures
- Partnership and Supplier Management
- Data privacy protection measures
- Continuous monitoring and updating

2，Proposed Initiative

Core objective: Security

Reliability and Robustness :

→ Enterprise risk management system maintains a high degree of functional stability even under Deliberate attacks

Prevent

- false
- Harmful
- Biased

Data Privacy

Ensure data privacy is adequately protected

Edge goal: business operations

Cost control:

reduce costs caused by risk management measures

Efficiency improvement:

Optimize the time to build a security system

Reduce the delay of service because of risk measures

Standards & Manage:

Management cost and difficulty

Compliance with policy and standards

Future state

security :

The integrity and privacy of data and services

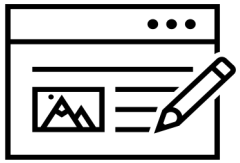
The security of system

Enterprise operations :

Easy to monitor, update and maintain

The costs are effectively controlled

Design



- **Core Issue:** Lack of systematic approach

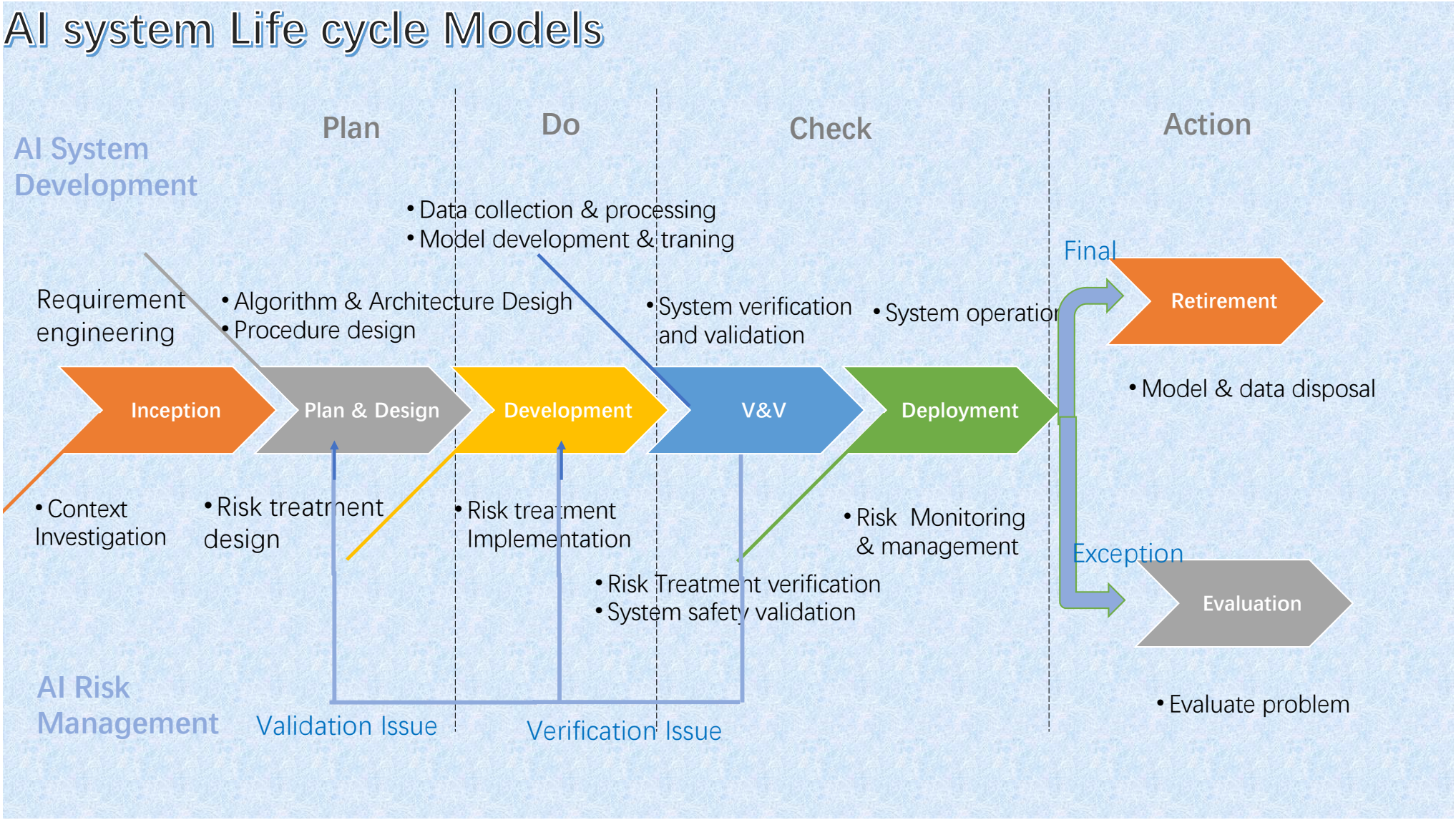
- **Proposed Solution:** AI system security + Life cycle management framework

→ integrates the development and deployment process of AI systems with AI risk management measures, introducing a model that combines functionality and security with a complete lifecycle.

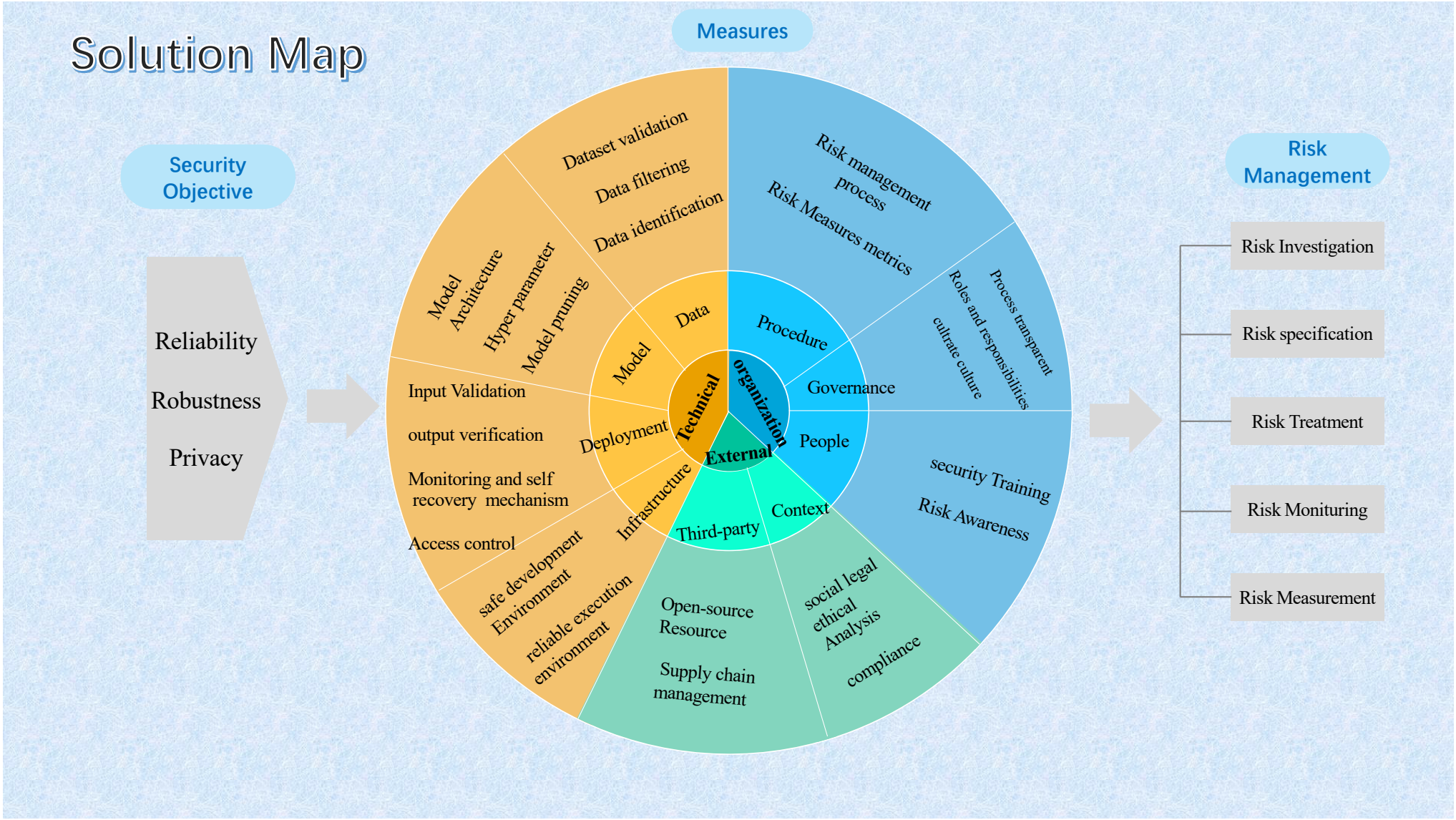
- **Basis of Solutions:**

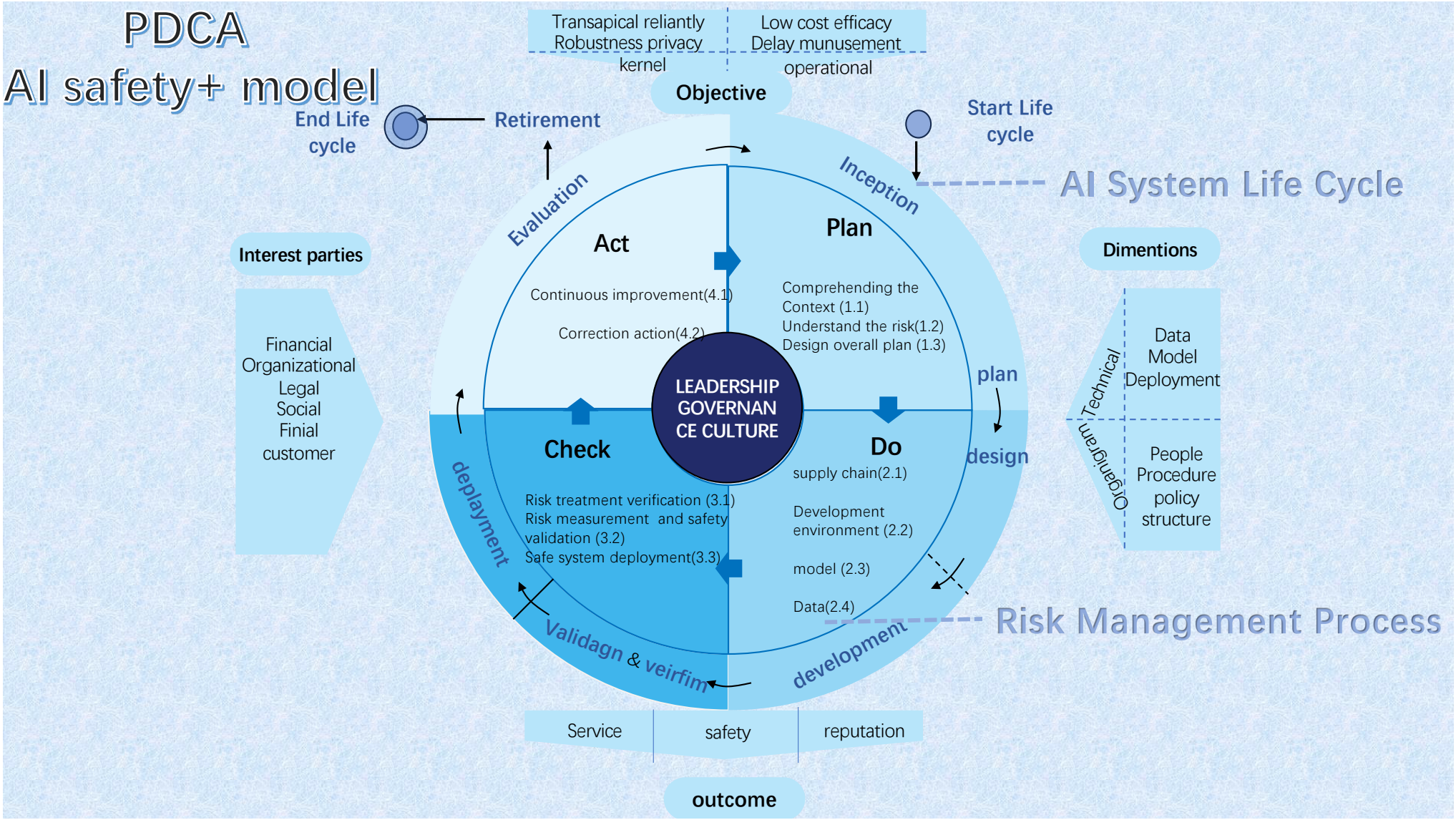
AI safety standards from National Institute of Standards and Technology (NIST) & National Cybersecurity Center (NCSC)

Research paper on International Conference on Software Engineering (ICSE)


AI system Life cycle Models
Plan
Do
Check
Action
AI System Development
• Data collection & processing
• Model development & traning
Final
Requirement engineering
• Algorithm & Architecture Desigh
• Procedure design
• System verification and validation
• System operation
Retirement
Inception
Plan & Design
Development
V&V
Deployment
• Model & data disposal
• Context Investigation
• Risk treatment design
• Risk treatment Implementation
• Risk Monitoring & management
Exception
• Risk Treatment verification
• System safety validation
Evaluation
AI Risk Management
Validation Issue
Verification Issue
• Evaluate problem

Solution Map



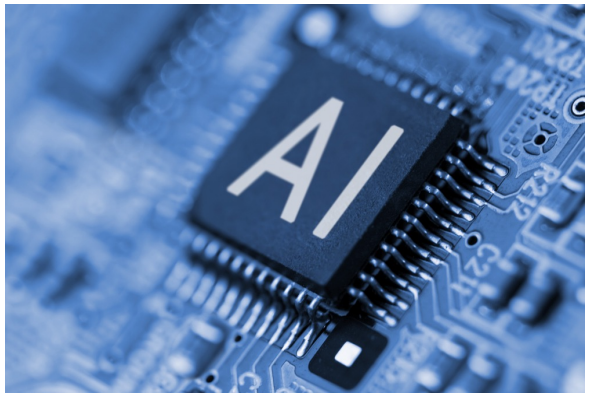

PDCA
AI safety+ model
Transapical reliantly
Robustness privacy
kernel
Low cost efficacy
Delay munusement
operational
Objective
End Life
cycle
Retirement
Start Life
cycle
AI System Life Cycle
Evaluation
Inception
Act
Plan
Interest parties
Dimentions
Continuous improvement(4.1)
Correction action(4.2)
Comprehending the
Context (1.1)
Understand the risk(1.2)
Design overall plan (1.3)
Financial
Organizational
Legal
Social
Finial
customer
LEADERSHIP
GOVERNAN
CE CULTURE
plan
design
Technical
Organigram
Data
Model
Deployment
People
Procedure
policy
structure
Check
Do
deplayment
Risk treatment verification (3.1)
Risk measurement and safety
validation (3.2)
Safe system deployment(3.3)
supply chain(2.1)
Development
environment (2.2)
model (2.3)
Data(2.4)
Validagn & veirfim
development
Risk Management Process
Service
safety
reputation
outcome

Impacts and differentiation analysis

Impacts:

enhanced system security
And service security guarantees

user trust and improved
corporate reputation



Innovation and technological leadership position

Impacts and differentiation analysis

Differentiation:

Customized risk management solutions And reduce security risks from multiple dimensions.

PDCA modeling

→ integrate AI risk management measures into the AI system lifecycle

Compliance with the actual business environment local policy

3, Plan of action and criteria for success

PART 3

Criteria of success

Quantitative Metrics

Reliability :

Under high loads and extreme conditions
(Stress and load testing)

- Mean Time Between Failures (MTBF)
- MTTR (Mean Time To Recover)

adversarial attack benchmarks

Robustness testing:

attacks are simulated to test the system's defense capabilities

- Minimize the success rate
- adversarial attack benchmarks

Enterprise Operations :

changes in operational costs, system deployment times → system's additional costs
standard compatibility reports → Compliance with standards
personnel training time. → management complexity

1. Plan

1.1: Comprehending the Context

Impact Factors:
social, legal, and ethical implications related to system risk

Business Environment:
Business goals, customer needs and market trends
→ restrictions of risk management.

System Requirements:
Define AI systems and model the user requirements.

1.2: Understanding Risks

Identifying Risks and specification:

Analyzing the system type and execution context
Systematically identify and document potential risks associated with system

Analyze Risks :

Considering past incidents/feedback from similar deployments and Evaluate the potential impact

Risk management Objective:

Set clear safety goals for AI Risk management
→ Meet organizational standards and user expectations

Risk Measurement:

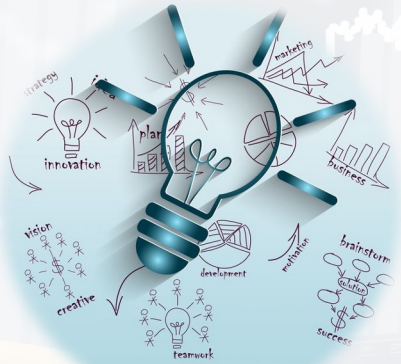
Establish clear success metrics for system security

1.3: Design Overall Plan

System Architecture:

Clearly design and document:

- Overall structure
- Functionality



Risk Management:

A documented process :

Oversight and checks consistent with organizational policies

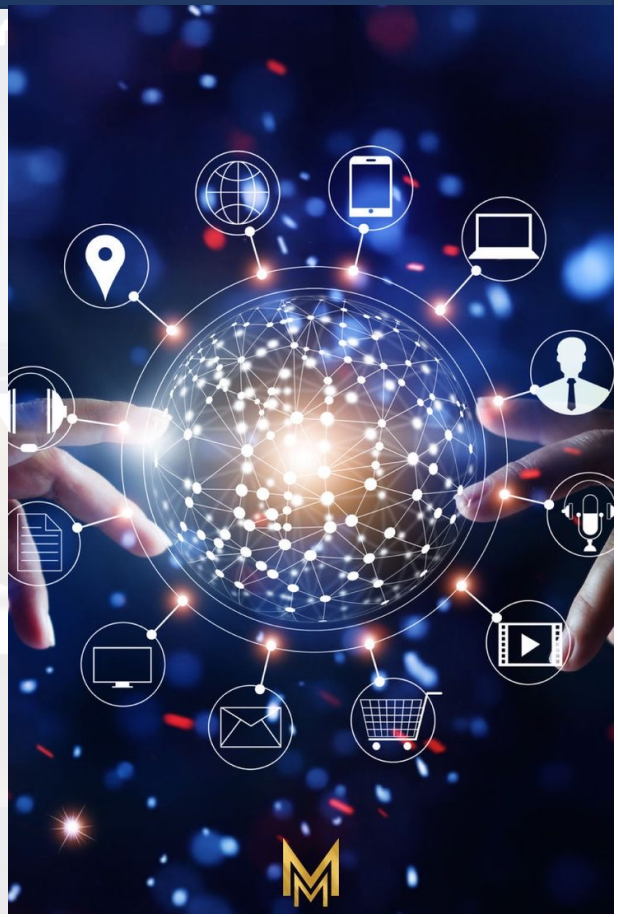
Procedure and management:

Clear, actionable timeline

→ Guides the AI deployment process

Define AI systems and model the user requirements

- The management strategies
- Human resource plans



2. Do

2.1 Make Supply chain safe

Monitor the external resource:

E.g: Dataset, API services

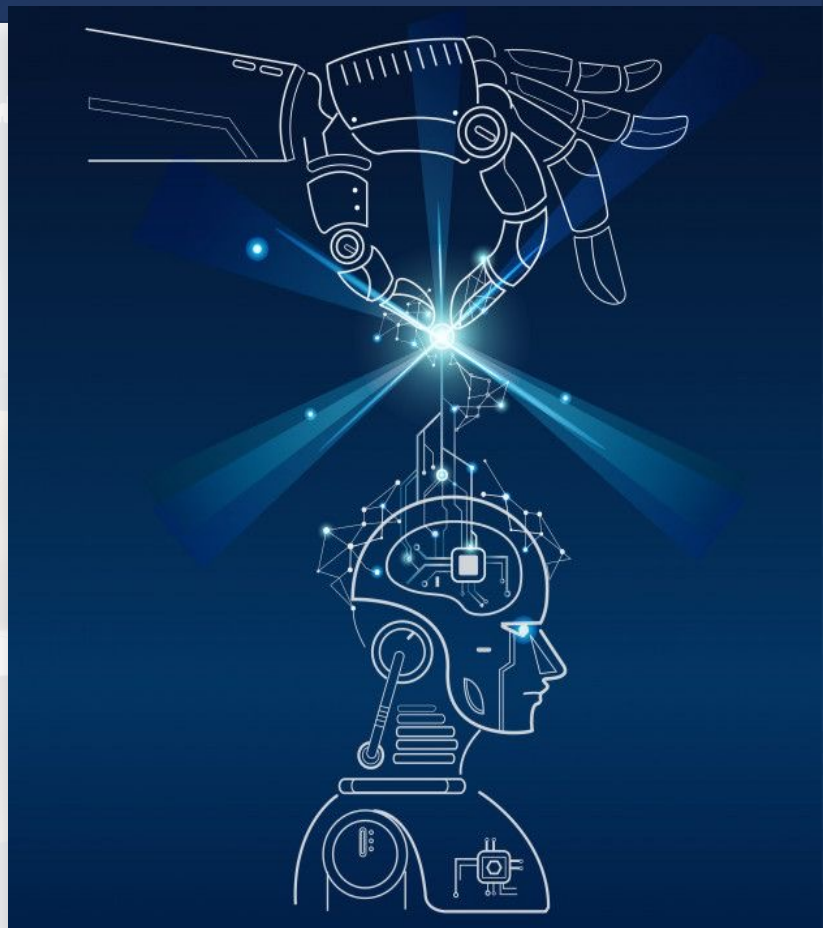
Design the risk control strategy

→ Manage the third- party resources

Monitor and investigate:

The open-source pretrained model

Weights to be used

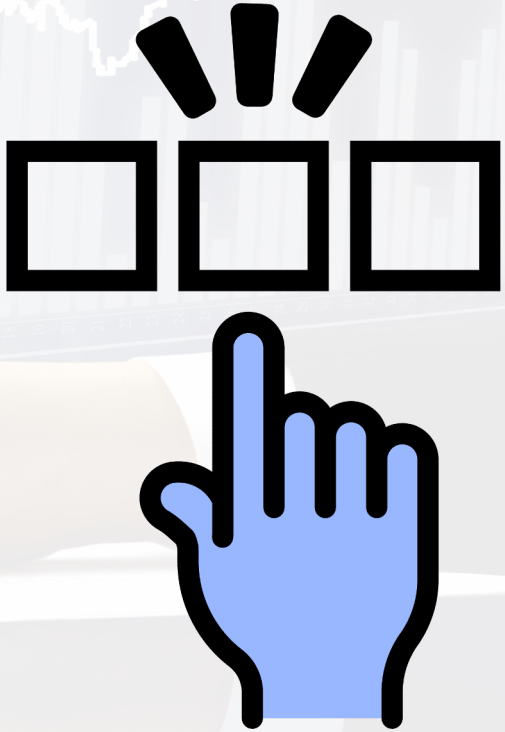


2.2 Make Development Environment safe

Ensure a safe environment setup:

Selecting suitable:

- Sourcing software
- Data warehouses
- MLOps tools
- Testing platforms



Monitor the whole development process :

Ensuring strict access control to training
Environment based on the principle of least privilege
Implement thorough access logging and monitoring

2.3 Make Model safe

Trade-off between performance and Security and explainability :

Use bottom-up architecture selection approach:

Valuate a variety of model types on data

- → Identify the most suitable ones
- → proceed to hyperparameter selection
- → Model pruning

2.4 Make data safe

Data filtering and preprocessing.

Implementing out-of-distribution detection mechanisms for inference inputs

Use privacy enhancement technologies :

Data de-identification, minimization

3. Check

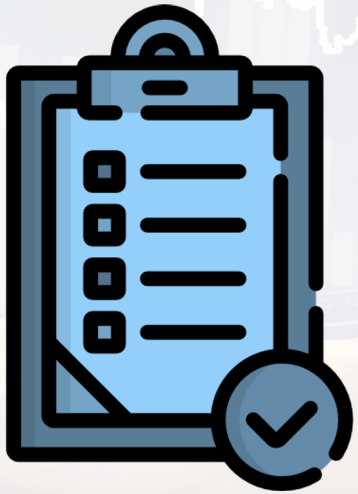
3.1 . 3.2 Verification and Validation (V&V)

Risk Treatment Verification:

Ensure that all aspects of the AI system, comply with risk management policies and Procedures

Risk Measurement and safety Validation:

The AI system undergoes tests on measurement metrics, following established model risk assessment criteria



3.3 Safe System Deployment

Design and Implementation of Safety Guardrails :

- Input validations
- Output verification mechanisms

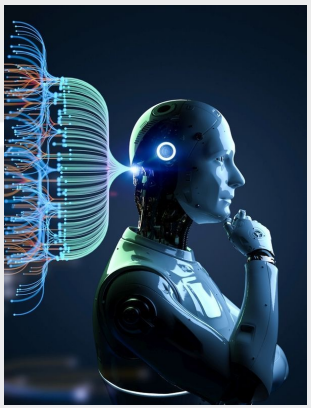
Utilizing a combination of technological tools and manual review

Establishment of Anomaly Detection and Self-healing Mechanisms :

Real-time monitoring systems with anomaly detection and self-healing capabilities

Access Control Implementation:

Leveraging cybersecurity frameworks



4. Action

4.1 Continuous Improvement & 4.2 Corrective Action

Refined AI system's risk management framework :

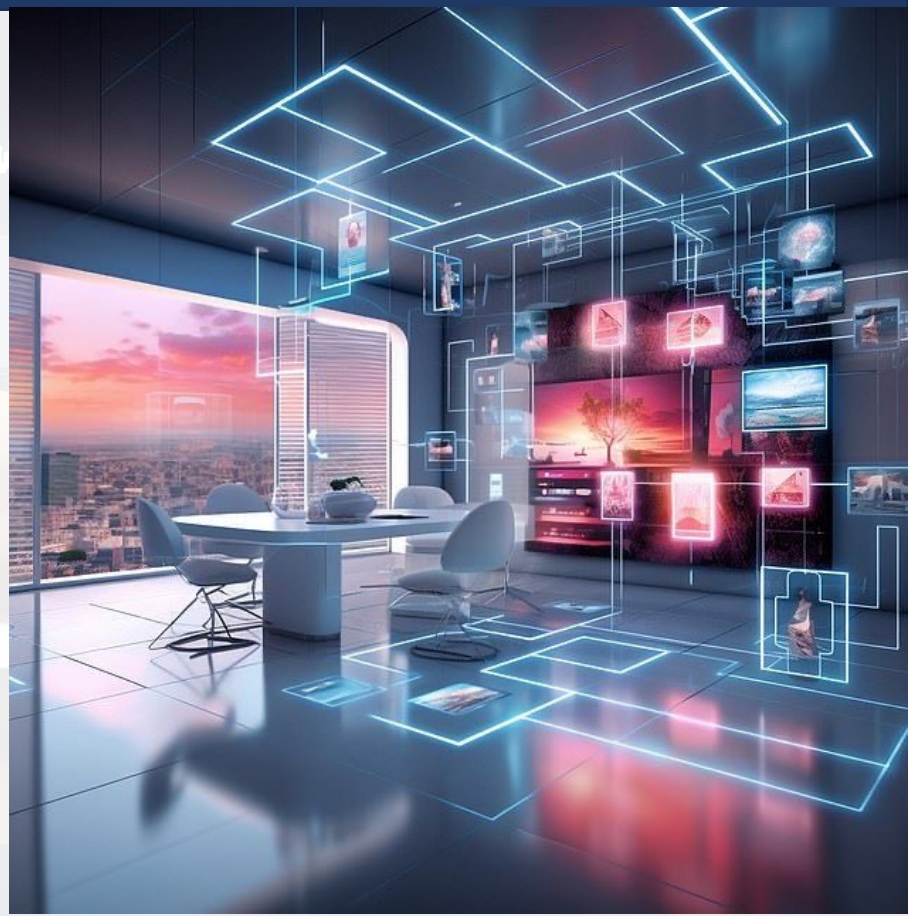
By systematic learning from operational data and feedback

Enabling proactive adaptation to emerging risks and maintaining alignment

Corrective action:

When a risk-related issue is identified, when resolve the specific problem

→ Strengthen the overall risk posture through improved controls and mitigation strategies.



5. Leadership Governance Culture

5.1 Roles and Responsibilities & 5.2 People and awareness:

Establish a system of accountability and process transparency:

* Detailing who is responsible at every stage
* Ensuring that mechanisms are in place for human Operators

People and awareness:

* Implement comprehensive training
* To ensure understanding and effective management of AI risks
* Coupled with fostering a corporate culture that emphasizes ethical use and risk-awareness in AI development and application

5.3 Procedures and Methodology & 5.4 Policy and standard Compliance

Define risk management processes
Conduct systematic assessments

To ensure transparency

- integrating policies
- resource allocation

based on assessed levels of risk and potential impact within a broader enterprise risk management strategy

Align risk management practices:

With applicable laws, regulations, and norms

4, Project Charter

PART 4

4.1 Business case & Goal

Help LY corporation's application services build secure AI systems.

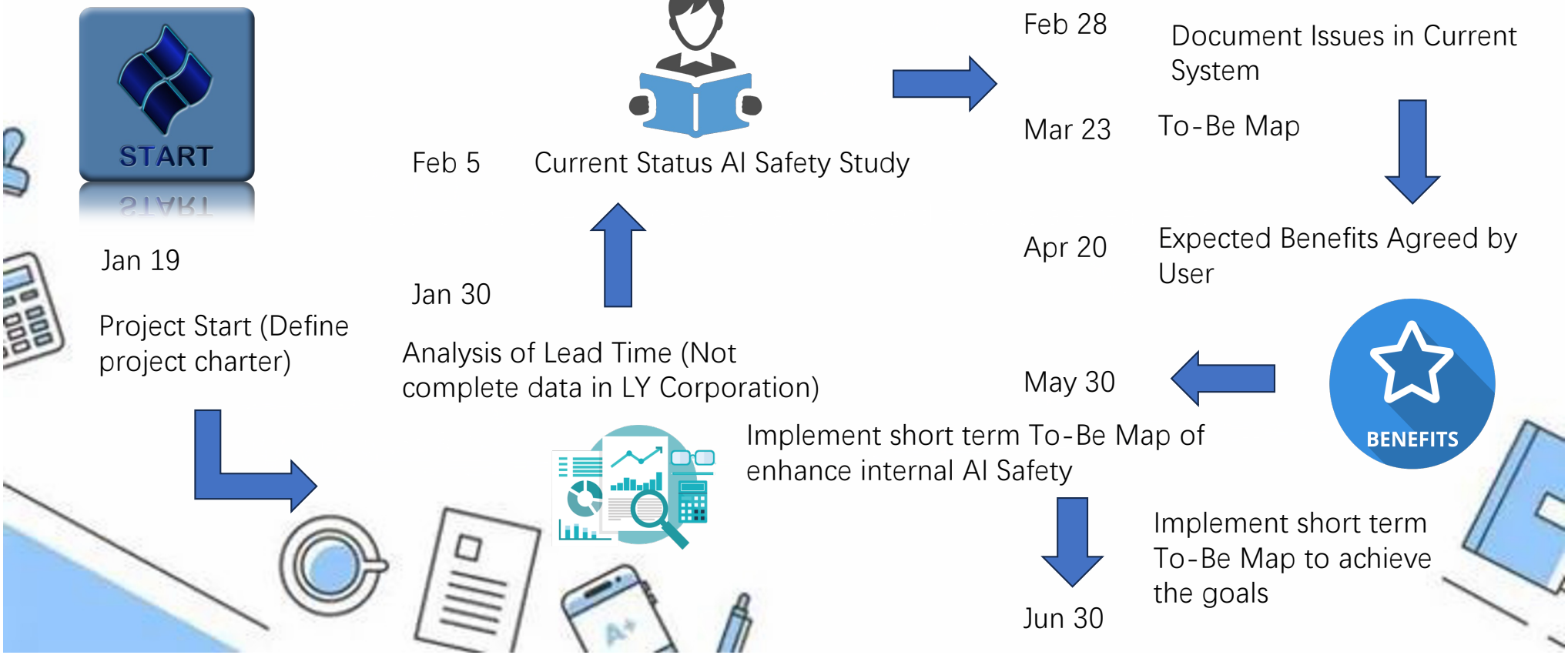
Integrate AI security risk management measures into the process of AI system development.

To enhance their internal AI system more:

- **Reliable**
- **Safety**
- **Robust**

- Strengthen employees' AI safety awareness

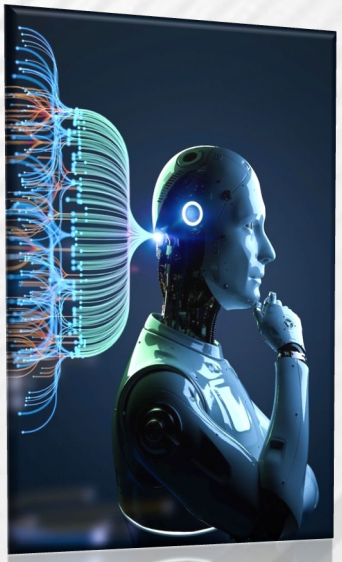
4.2 Milestone



4.3 Scope, Benefits and Cost

Scope: In LY Company

Out: Enhance LY Corporation
AI system safety



Tangible: $ 100000

Intangible:

- Employees
- Policy and legal
- Management
- Time: Service delay and System deployment


Thanks