RCM
RiskChainModel

June 2021

Risk Chain Model (RCModel) Guide Ver 1.0

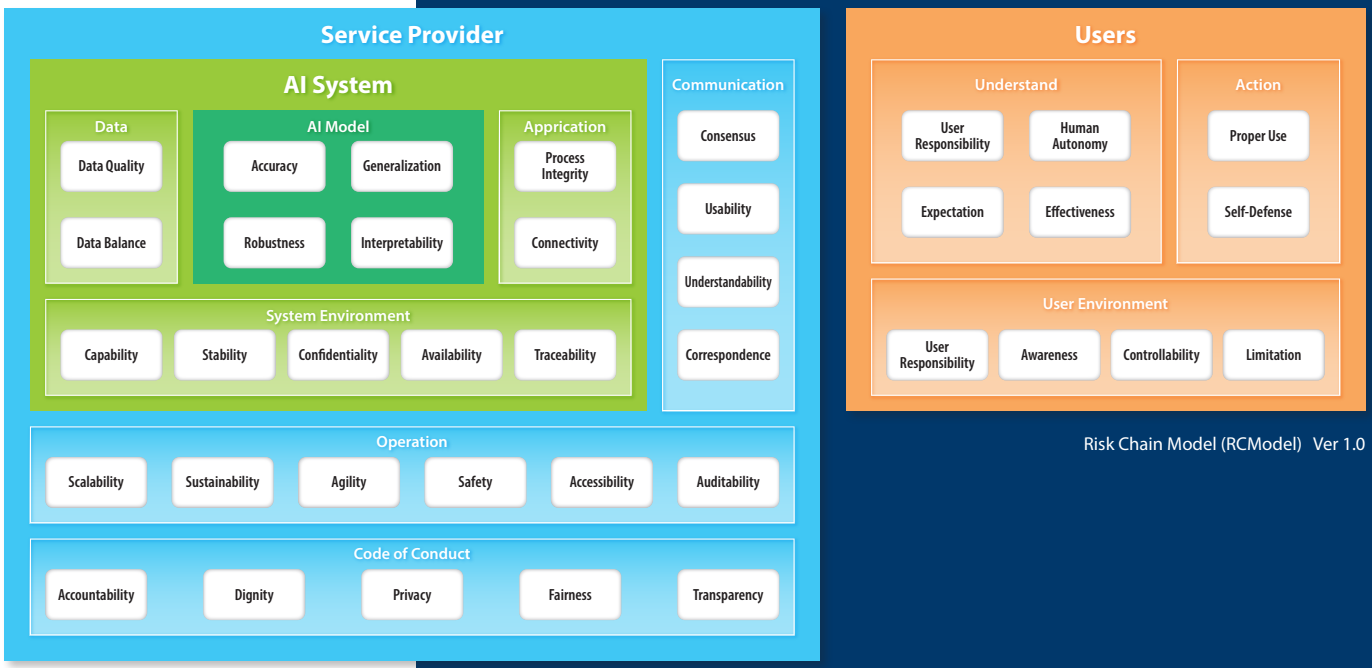
Institute for Future Initiatives, The University of Tokyo
Technology Governance Research Unit
AI Governance Project

Risk Chain Model Overview

For AI-based systems and services, it is crucial to ensure reliability and transparency; however, how exactly should this be done?

Various principles, guidelines, and checklists have been created around the world. This presentation will introduce the risk chain model (RCModel) developed by a research group at the University of Tokyo to incorporate these principles into practice.

RCM
RiskChainModel



Risk Chain Model (RCModel) Ver 1.0

Objectives of the RCModel (1)

What are the "significant risks" of AI services?

What are the significant risks in developing and using AI services? This depends on the AI service itself. For AI that evaluates individuals, such as "recruitment AI" or "loan screening AI," "fairness" is important, and for AI that handles unstructured data such as images, "infrastructure appearance inspection AI," "robustness" and "adapting to the external changes" are required.

First, the RCModel identifies significant risk scenarios for the AI service. Rather than employing a huge checklist to evaluate all AI services in a unified manner, the RCModel defines the "value and purpose of the AI service" and examines the "risk scenarios" that hinder the realization of the "value and purpose."

Objectives of the RCModel (2)

Can the AI model solely and adequately address risks?

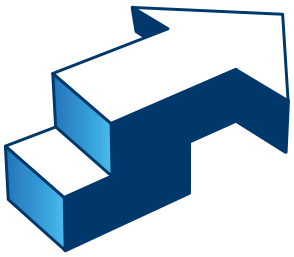
Can the AI model solely and adequately address all "significant risk scenarios?" In practice, this is very difficult owing to the uncertainty of the AI model's performance. Even if an AI model exhibits an excellent prediction performance at the time of development, its prediction performance may not be sustained towing to changes in the environment. In addition, the prediction performance of AI models may deteriorate because of misuse or abuse by users.

The RCModel considers risk mitigation (risk control) for the identified "significant risk scenarios," not only for AI models, but also for related technologies (data, system infrastructure, etc.), service providers, and users. Accordingly, risk control plans can be developed for the entire AI service.

Structure of the RCModel

Three-layer structure of AI system/service provider/user

The RCModel is structured into three layers: (1) AI models (accuracy, robustness, etc.), and AI systems (data, system infrastructure, other automated control functions), (2) service providers (code of conduct, service operation, communication with users), and (3) users (understanding, utilization, user environment).



How to operate the RCModel

This section explains the operation of the RC Model. We review each step with reference to the sample case "Recruitment AI"

Step 1 Define the " values and objectives to be achieved" by AI services.

The first step is to define the " values and objectives to be achieved" based on the following features of AI services. These do not only include business objectives such as "realization of unmanned operations," "attractive proposals," and "more advanced detection than humans," but also incorporate social responsibilities such as "prevention of inappropriate use" and "assurance of safety." In this process, a trade-off may exist between "values" In such instances, it is important to consider the "values and objectives to be achieved," which should be prioritized to reduce risks.

Features of AI services	• Purpose of use of the AI service • Conceptual diagram of the system • Algorithms and data to be used • Division of roles between developers and users • Development methods and learning frequency

■ Sample case "Recruitment AI", p. 2.

Case01 : Recruitment AI

This is an AI service used as reference information when judging the selection of documents for the entry sheet for the human resource recruitment department (HR dept.) in a global company, i.e., Company A.

The AI development department of Company A receives the past entry sheet data and results of pass/fail judgment from the HR dept. of Company A (including overseas group companies), which is a business user. The department created a learning model for judging the rates of pass/fail through machine learning (classification model).

[Values & Objectives]

- Maintaining and improving the hiring level
- Reducing costs of recruitment
- Providing service for groups
- Corporate Social Responsibility (Fair Recruitment Activities)

It is evaluated based on the precision (the percentage of successful applicants who have passed an interview and have been offered a job), and 70% is set as the expected value. The precision is used as an evaluation index because the recall (the percentage of people who did not get a job offer because they did not pass the screening process by AI) is insufficient data for the examination.

The HR dept. of Company A reads the entry sheet (through an electronic file) of the applicant (both new graduates and mid-career graduates) into the learning model and can confirm the judgment result (pass/fail) of the AI on its own personal computer (using a browser). It should be noted that not only the pass/fail judgment but also the keyword affecting the judgment is highlighted and displayed in the entry sheet on the output screen. The person in charge of the HR dept. sets up a pass/fail judgment using the judgment of the AI as reference information, obtains the approval of the head of the personnel department, and notifies the applicant of the pass/fail judgment.

The real data stored are appropriately added as learning data at the time when the determination result of the pass/fail judgment is input, and the learning model is updated daily and reflected upon in the real environment. However, the AI model in the production environment is not automatically updated when the correct answer rate is less than 70% as a result of cross-validation testing during the learning process. The AI model stores versions from the previous year.

2

Step 2 Examination of significant risk scenarios

After defining the "values and objectives to be achieved," we examine the risk scenarios that may hinder its achievement. For example, in the case of "unmanned operation," consider risk scenarios that may hinder the maintenance of services, such as "degradation of accuracy owing to environmental changes" and "abnormal operation." Furthermore, in the case of "attractive proposals," risk scenarios such as "risk of proposing strategic products to anyone" that may have a negative impact on business strategies are considered. Based on the impact on the values and objectives to be achieved, we will discuss with various stakeholders and consider the priority of each risk scenario.

■ Sample case " Recruitment AI", p.6

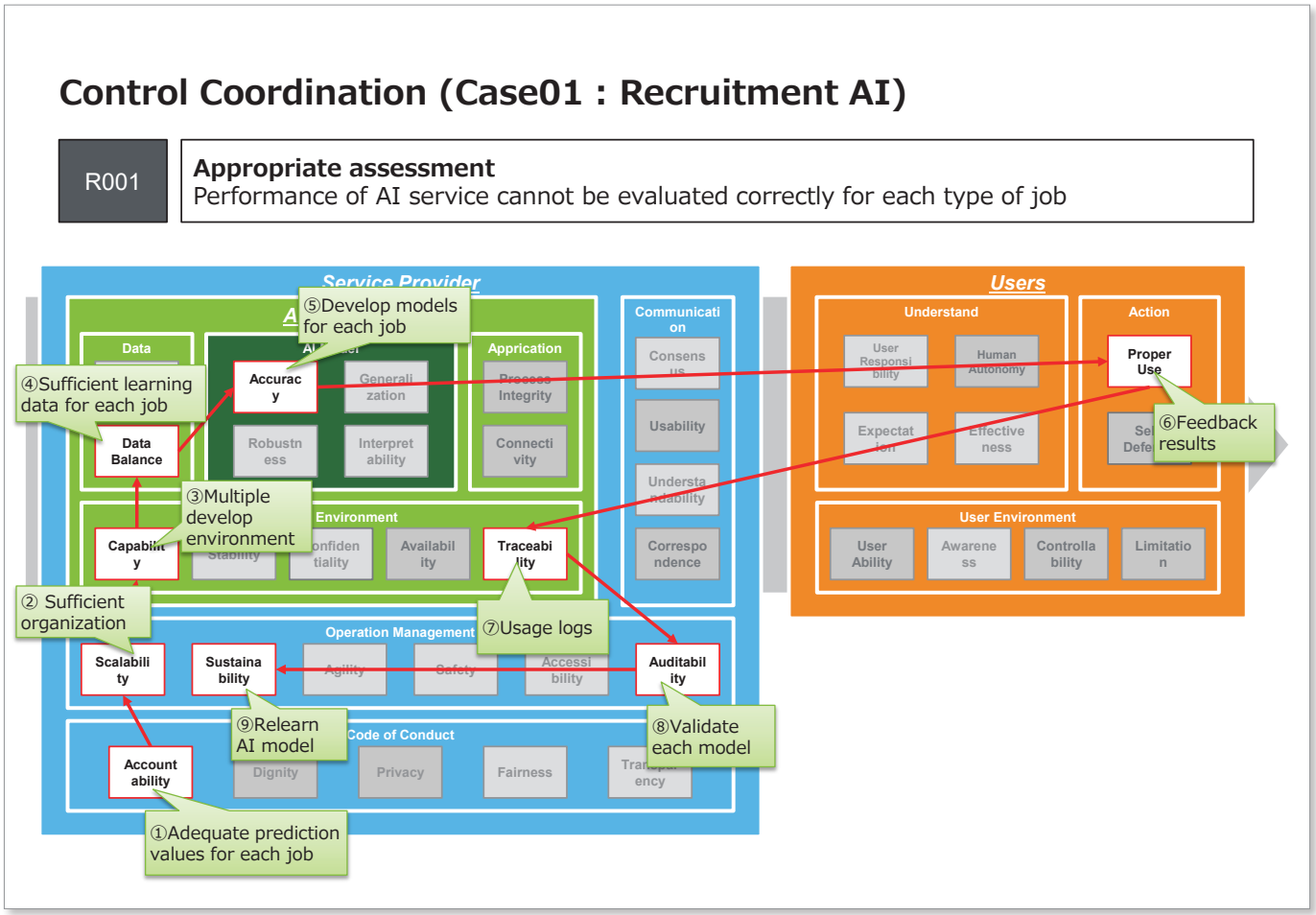
Risk Assessment (Case01 : Recruitment AI)

Values & Objectives		Service Requirement			Risk No.	Risk Scenario	
1	Maintaining and improving the hiring level	1-1	Precision Performance	■ Accuracy ■ Robustness ■ Explainability	R001	Appropriate assessment	Performance of AI service cannot be evaluated correctly for each type of job
					R002	Unstable performance	Recruiting level decreases owing to deterioration of AI prediction performance
					R003	Impact by noise	AI decisions will change significantly by tiny different (e.g., punctuation marks)
					R004	Falsehood	The application with falsehood is passed
		1-2	User interaction	■ Feedback for AI by User	R005	Excessive AI dependence	Person in charge of HR dept. relies excessively on AI decisions
					R006	Inaccurate feedback	Inaccurate feedback (pass/fail labeling) to AI by HR dept. degrades AI performance
		1-3	Adopting the changes in environment	■ Data shift	R007	Change in HR trends	AI model cannot adapt to changes in HR trends of talent required
					R008	New occupation	AI model cannot achieve sufficient prediction when seeking new occupations
2	Reducing costs of recruitment	2-1	Adequate cost	–	R009	Excess costs	Operation costs are exceeded
3	Providing service for groups	3-1	Service localization	■ Multiple models	R010	Differences in groups	AI does not work effectively owing to differences in the circumstances of group companies
		3-2	Development and support system	■ Organization	R011	Insufficient speed	No timely improvement when model accuracy deteriorates
4	Corporate Social Responsibility (Fair Recruitment Activities)	4-1	Compliance with ethics	■ Explainability ■ Fairness	R012	Internal abuse	Identify key phrases that make a pass with a high probability and leak them illegally outside the company by using AI services numerous times
					R013	Fairness	Unfair forecast results for a particular group
		4-2	Data Protection	■ Data protection	R014	Unintended use	The use of AI decision for other purposes causes a disadvantage to a specific individual.
					R015	Harmful rumors	Leakage of AI prediction damages the reputation of a specific person
					R016	Leakage of privacy data	Performance of AI service cannot be evaluated correctly for each type of job

Step 3 Examine the risk chain (relation of risk factors) for each important risk scenario.

For each risk scenario, the risk factors at each layer of the RCModel (AI system/service provider/user) is recognized, a risk chain(line) is drawn, and the order in which the risks become apparent is considered. The chain may not be in one direction or a single line, and the starting point may vary depending on the risk or problem being addressed.

■ Sample case " Recruitment AI" p.10 (Example of consideration of risk scenario "R001. Appropriate evaluation")



Step 4 Consider risk control according to the risk chain

We consider risk reduction measures (risk controls) for each element associated with the risk chain. Although risk controls may be unable to sufficiently mitigate risk in isolation, they incrementally mitigate risk by association. The RCModel approach considers the optimal risk control based on the magnitude of the risk and its cost-effectiveness, such that the risk response measures do not become excessive and unrealistic.

Consider what countermeasures (controls) should be taken at each layer among the parties involved (development project, data scientist, service provider, user, etc.), and the scope of responsibility. Because several risk scenarios draw lines at multiple layers, there is a need for not only developers, but also service providers and users to consider risk measures for AI.

■ Sample case " Recruitment AI" p.11 (Example of consideration of risk scenario "R001. Appropriate assessment")

Risk Control (Case01 : Recruitment AI)

R001	**Appropriate assessment** Performance of AI service cannot be evaluated correctly for each type of job

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
③[Capability] Multiple environments to develop models for each type of job (IT dept., Co. A) ④[Data Balance] Sufficient learning data to develop models for each type of job (AI dev dept., Co. A) ⑤[Accuracy] Develop models with sufficient prediction for each type of job (AI dev dept., Co. A) ⑦[Traceability] Store usage logs (IT dept., Co. A)	①[Accountability] Define adequate prediction values for each type of job (HR dept., Co. A/HR dept., group A) ②[Scalability] Establish service organization to develop models for each type of job (HR dept., Co. A) ⑧[Auditability] Validate each model performance (HR dept., Co. A/HR dept., group A) ⑨[Sustainability] Relearning AI model (HR dept., Co. A)	⑥[Proper Use] Feed the pass/fail results back to AI models (person in HR dept., group A)

11

Step 5 Organize the roles of each stakeholder based on the examination of each risk chain.

After considering risk control through the risk chain for each risk scenario, the roles of each stakeholder, including internal and external organizations, contractors, and users should be organized.

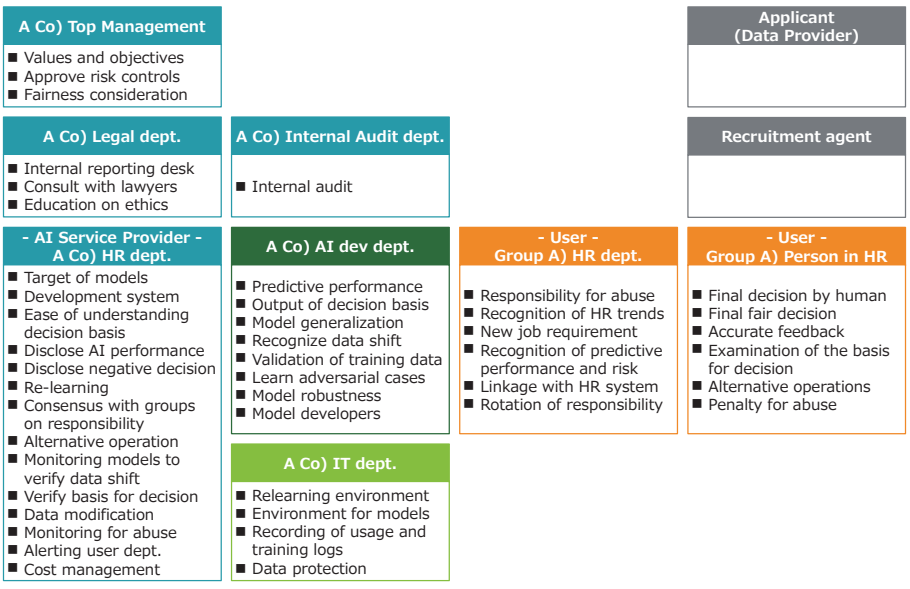
By using the RCModel to organize a series of review processes related to "values and objectives to be achieved," "significant risk scenarios," and "roles of each stakeholder," it is expected that all parties involved should have a common understanding.

■ Sample case " Recruitment AI," p. 7-8.

Risk Assessment&Control Summary (Case01 : Recruitment AI)

Values & Objectives		Risk No.	Risk Scenario	Technical difficulty	Environmental change	Caused by user	RC	Control Summary: AI System	Control Summary: AI service provider	Control Summary: User
1	Maintaining and improving the hiring level	R001	Appropriate assessment	○			●	Multiple environment Multiple models	Values of each job Relearn AI model	Feedback results
		R002	Unstable performance	○			●	Prediction performance Store usage logs	Validate model Relearn AI model	Alternative manual operation
		R003	Impact by noise	○			●	Adversarial examples Basis of decision	Easy to understand Validate the basis	Using basis of model decision
		R004	Falsehood	○			●	Prediction performance Basis of decision	Similarities with past falsehood cases Alert Users	Review of selection
		R005	Excessive AI dependence	○		○	●	Basis of decision	Information of AI Easy to understand	Understand risk Manual process
		R006	Inaccurate feedback	○		○	●	Verifying annotation Store training logs	Data correction Relearning	Accurate feedback Linkage with HR data
		R007	Change in HR trends	○	○		●	Examine data shift Review generalization	Periodic review Relearn AI model	Recognize HR trends
		R008	New occupation	○	○		●	Processing performance Sufficient learning Data	Development team Verify models	Requirement of new job
2	Reducing costs of recruitment	R009	Excess costs					Appropriate pricing	Cost control	
3	Providing service for groups	R010	Differences in groups	○	○		●	System environment Individual models	Target of individual Monitoring model Development team	
		R011	Insufficient speed					Relearning efficiently Developer	Project organization	
4	Corporate Social Responsibility (Fair Recruitment Activities)	R012	Internal abuse	○		○	●	Store usage logs	Monitoring abuse Consultation with lawyer	Internal retraction
		R013	Fairness	○		○	●	Data balance Model generalization	Fairness consideration Clarify negative tendency	Understand AI tendency Human decision
		R014	Unintended use			○		Data protection	Access control Proper use	Compliance
		R015	Harmful rumors			○		Data protection	Compliance	Compliance
		R016	Leakage of privacy data			○		Data protection	Compliance	Compliance

Organization (Case01 : Recruitment AI)



When to use the RCModel and its further development

It is advisable to consider the timely use of RC Model in the PoC or development stage, and to achieve the necessary risk control before the start of utilization.
However, the significant risk scenarios may change owing to moderations in the business environment and related legal systems. Therefore, it is desirable to periodically review and revise the necessary risk controls for utilization.
In addition, it had better to understanding the social and legal issues and findings related to AI, and to examining future issues related to AI services in advance, by holding discussions on the series of examination processes by the RCModel, including each internal and external stakeholder as well as relevant experts and specialists.

Main Issues

1. How will the business environment and social demands change in the fields covered in this case?
2. How can science technology respond to the above changes?
3. What will be the relationship between technology and humans?
4. What is happening overseas?
5. What are the trends in legal reform?

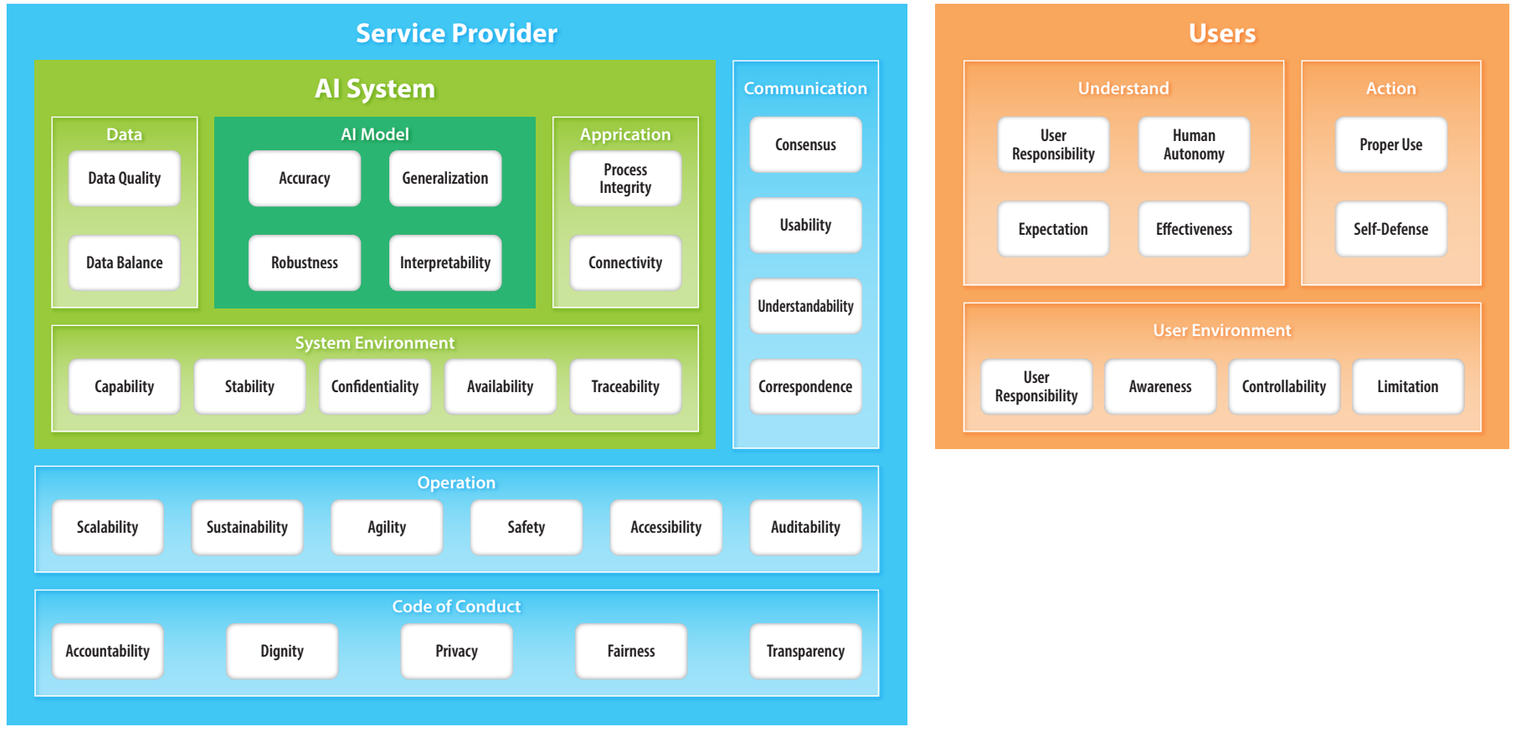
About using the RCModel

Technology Governance Research Unit, Institute for Future Initiatives, The University of Tokyo
ifi_tg@ifi.u-tokyo.ac.jp

RCM
RiskChainModel

Components of RCModel

(Ver 1.0)



AI System

Components	contents	Examples of major controls
AI Model		
Accuracy	Prediction performance	Verification of prediction performance
Generalization	Decisions without bias toward specific cases	Verification of generalization performance
Robustness	Resistance to noise	Learning adversarial examples
Interpretability	Basis of decision	Output the basis of decisions (ex. importance of feature quantity)
Data		
Data Quality	Ensuring data quality	Validation of annotation, noise correction, classification of special data (strategic products, etc.)
Data Balance	Adequate data bias	Sufficient training data, verifying data bias
Applications		
Process Integrity	Automatic processing	Correction of outliers, abnormal prediction alert (ex. over detection)
Connectivity	Cooperation with external systems, etc.	Automated data linkage, protocol
System runtime environment		
Capability	A system environment that can cope with an increase in the number of usage environments, etc.	Environment for implementing multiple models
Stability	A system environment for stable operation	Ensuring adequate performance, maintaining the IoT
Confidentiality	A system environment where functions and information are protected	System protection, access Control
Availability	A readily available system environment.	Ensure system uptime, switching to alternative functions
Traceability	Properties for post-verification of AI services	Storage of execution history with basis of decision, records of learning results

Service provider

Components	contents	Examples of major controls
Code of Ethics		
Accountability	Accountability, protection of users	Clarification of responsibilities, appropriate expectations
Dignity	Consideration of priority rights on the part of users	Defining the degree of human intervention
Fairness	Fairness across services,	Equality criteria
Privacy	Appropriate management of personal information	Compliance with privacy policy
Transparency	Disclosure of necessary information related to AI services	Disclosure of information for each stakeholder
Operation		
Scalability	An organization that can respond to the increase in usage environment, etc.	An organization for service providing, support for the languages required
Sustainability	Maintaining sustainable services	Review of service requirements, re-learning and continuous learning
Agility	Timely response	Agile development
Safety	Ensuring safety throughout the service	Safety measures out of AI system, dealing with abuse by users
Accessibility	Management of appropriate system usage authority and scope	Access rights, restriction of unauthorized users
Auditability	Necessary monitoring, including of third parties	Monitoring of AI systems, verification when trouble occurs, monitoring abuse
Communication		
Consensus	Acknowledgement of with users	Agreement on User Responsibility, agreement to change AI system
Usability	Ease of usage	Easy-to-use user interface
Understandability	Ease of understanding of information output by AI services	Clear expression of the basis of decision
Correspondence	Remote support for users	Remote assistance for users, collaboration with experts

User

Components	contents	Examples of major controls
Understanding		
User Responsibility	Understanding user responsibilities	Agreement on user responsibility (final decision, etc.), agreement to change AI system
Human Autonomy	Understanding of autonomous decisions	Correcting AI decisions
Expectation	Understanding the expected accuracy of AI services	Understanding of expected performance (ex. prediction accuracy), understanding negative tendencies
Effectiveness	Understanding the impact on users	Understanding the risks associated with AI services, understanding changes in work styles
Usage Environment		
User Ability	Acquisition of knowledge and skills necessary for service use	Ensuring the necessary literacy, user education
Awareness	Awareness of AI's existence	Clarification of the existence of AI, notification to users
Controllability	Control on the user side	Maintenance of manual procedures, ability to correct AI decisions
Limitation	Restrictions on users to prevent unauthorized use, etc.	Technical restrictions on users, legal constraints
Action		
Proper Use	Correct service use	Use of services in accordance with appropriate purposes, correct setting of teacher labels
Self-Defense	Protection of the users	Claims when disadvantages occur, advanced collaboration on environmental changes

June 2021 **Risk Chain Model** (RCModel) **Guide** Ver 1.0