RCM
RiskChainModel

Risk Assessment & Control Coordination for AI services :
Case01 Recruitment AI

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

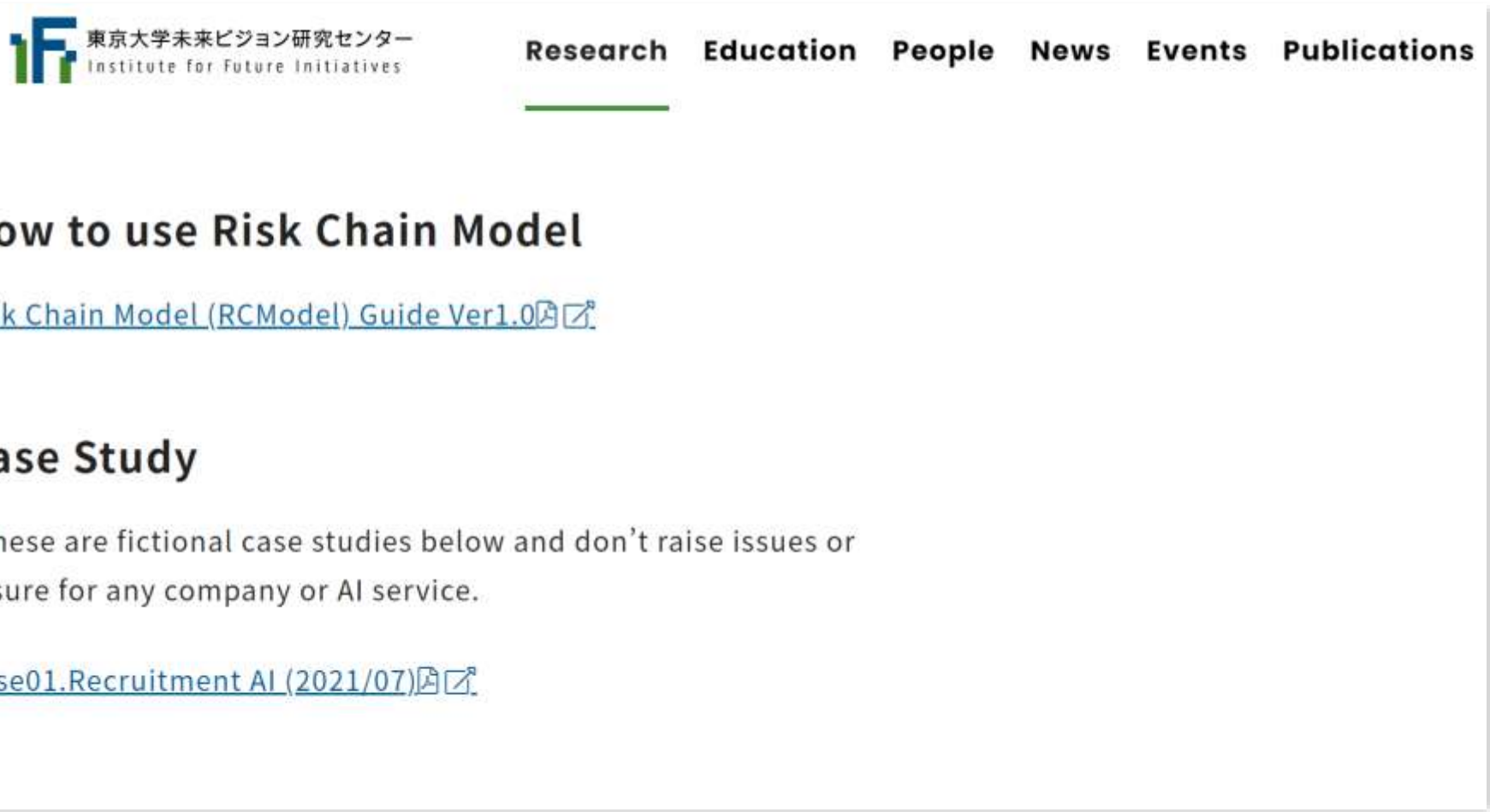
How to operate the RCModel
- Risk Assessment & Control Coordination -

Risk Assessment	Step1	Define the “values and objectives to be achieved” by AI services
	Step2	Examination of significant risk scenarios
Control Coordination	Step3	Examine the risk chain (relation of risk factors) for each important risk scenario
	Step4	Consider risk control according to the risk chain
	Step5	Organize the roles of each stakeholder based on the examination of each risk chain

Guide book and Case Studies of Risk Chain Model

AI Service and Risk Coordination Study Group

https://ifi.u-tokyo.ac.jp/en/projects/ai-service-and-risk-coordination/



RCM
RiskChainModel

Case Study

Case01 : Recruitment AI

- Define the “values and objectives to be achieved” by AI services -

Step1

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.

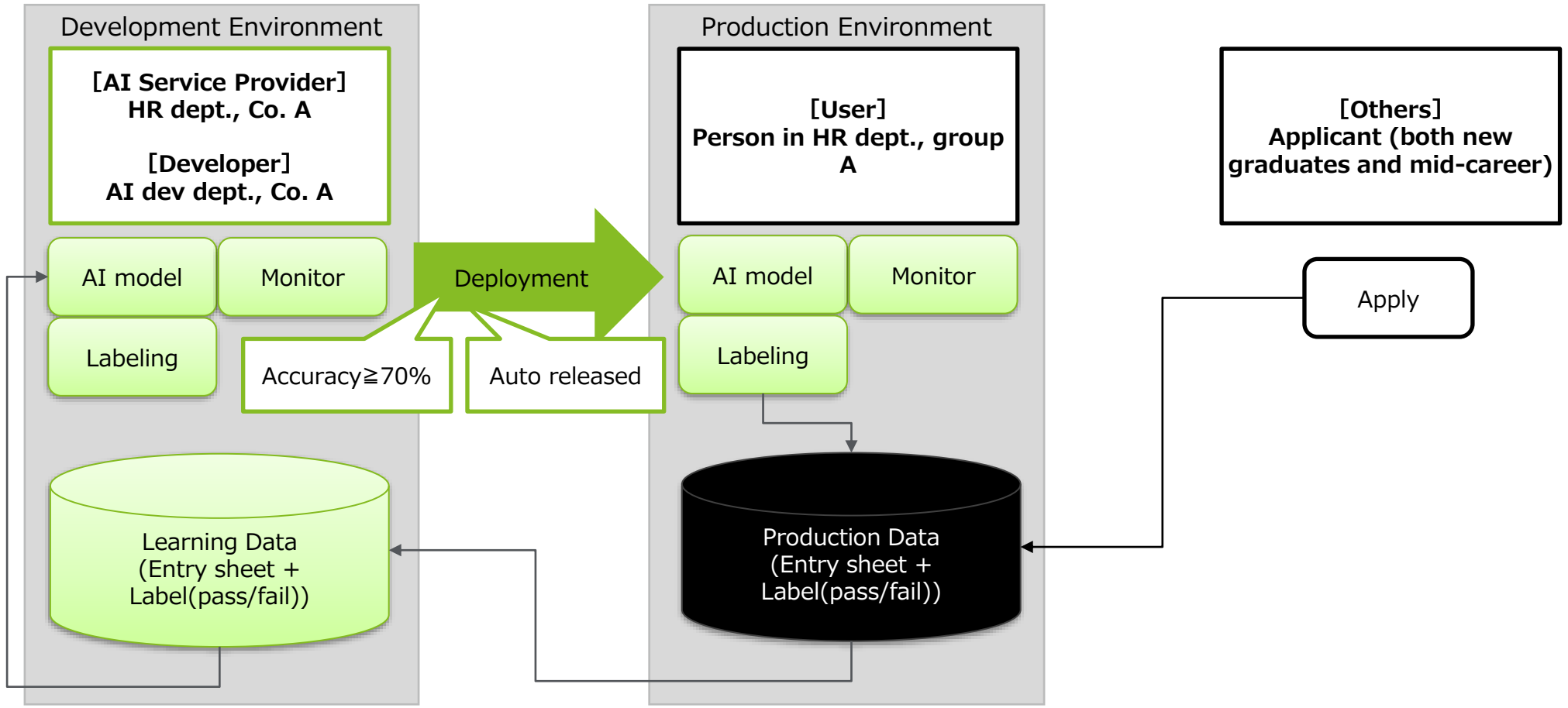
The HR dept. of Company A reads the entry sheet (through an electronic file) of the applicant (both new graduates and mid-career graduates) into the learning model and can confirm the judgment result (pass/fail) of the AI on its own personal computer (using a browser). It should be noted that not only 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.

Case01 : Recruitment AI

- System Overview -

AI System	AI dev dept., Co. A	Developing AI System (includes AI model)
AI Service Provider	HR dept., Co. A	Recommendation pass/fail for user by AI Service
User	Person in HR dept., group A	Final decision of recruitment



Case01 : Recruitment AI

- Input & Output -

[Input Data]

Data	Purpose	Collection Method	Data Manager	Including Privacy Data
Past entry sheet data	Learning	Entry sheet data submitted by the applicant to HR dept., group A and pass/fail label	Head of HR dept., Co. A/group A. (Private cloud environment)	Yes (including sensitive personal information)
Newest Entry sheet data	Production	Entry sheet data submitted by the applicant to *Company A*'s group personnel department	Head of HR dept., Co. A/group A (private cloud environment)	Yes (including sensitive personal information)

[Output]

Users	Person in charge of personnel department at *Company A*
Output	Pass/Fail
Output Method	When the entry sheet of the applicant is input into the terminal of the person in charge of the personnel department of *Company A*, th e judgment of the document selection is output
Expected Accuracy	Precision: 70% *The percentage of those who actually received job offers among those who passed the screening process
User judgment	Yes
Output of evidence information	Keywords that have had a strong impact on the decision are hi ghlighted in the entry sheet
Safety Risk	No
Connection with external system	No
Users	Person in charge of personnel department at *Company A*

RCM
RiskChainModel

Risk Assessment

Risk Assessment

- Examination of significant risk scenarios -

Step2

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

Risk Assessment&Control Summary

- Organize the roles of each stakeholder based on the examination of each risk chain -

Step5

Values & Objectives		Risk No.	Risk Scenario	Uncertainly	Environmental change	Caused by user	RC	Control Summary		
								AI System	AI service provider	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

- Organize the roles of each stakeholder based on the examination of each risk chain -

Step5

A Co) Top Management

- Values and objectives
- Approve risk controls
- Fairness consideration

Applicant (Data Provider)

A Co) Legal dept.

- Internal reporting desk
- Consult with lawyers
- Education on ethics

A Co) Internal Audit dept.

- Internal audit

Recruitment agent

- AI Service Provider - A Co) HR dept.

- Target of models
- Development system
- Ease of understanding decision basis
- Disclose AI performance
- Disclose negative decision
- Re-learning
- Consensus with groups on responsibility
- Alternative operation
- Monitoring models to verify data shift
- Verify basis for decision
- Data modification
- Monitoring for abuse
- Alerting user dept.
- Cost management

A Co) AI dev dept.

- Predictive performance
- Output of decision basis
- Model generalization
- Recognize data shift
- Validation of training data
- Learn adversarial cases
- Model robustness
- Model developers

A Co) IT dept.

- Relearning environment
- Environment for models
- Recording of usage and training logs
- Data protection

- User - Group A) HR dept.

- Responsibility for abuse
- Recognition of HR trends
- New job requirement
- Recognition of predictive performance and risk
- Linkage with HR system
- Rotation of responsibility

- User - Group A) Person in HR

- Final decision by human
- Final fair decision
- Accurate feedback
- Examination of the basis for decision
- Alternative operations
- Penalty for abuse

RCM
RiskChainModel

Control Coordination

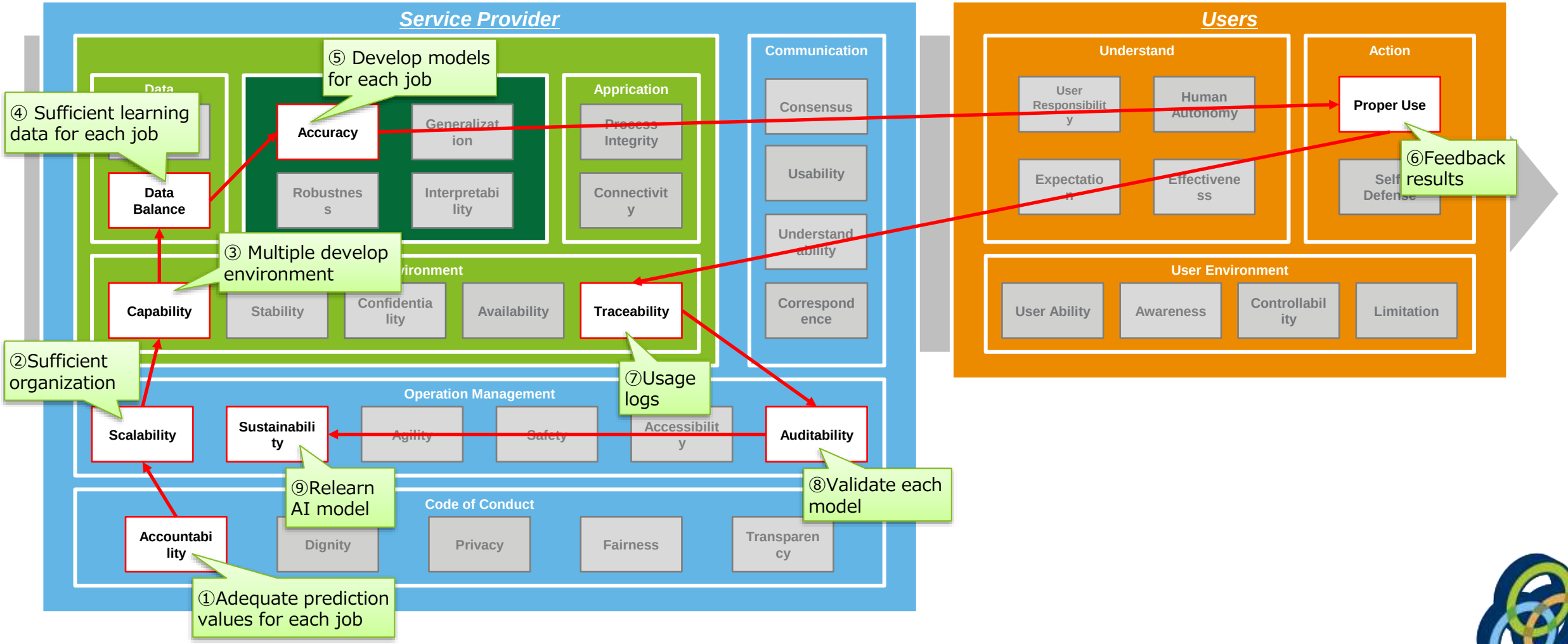
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R001

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



Risk Control

- Consider risk control according to the risk chain -

Step4

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)

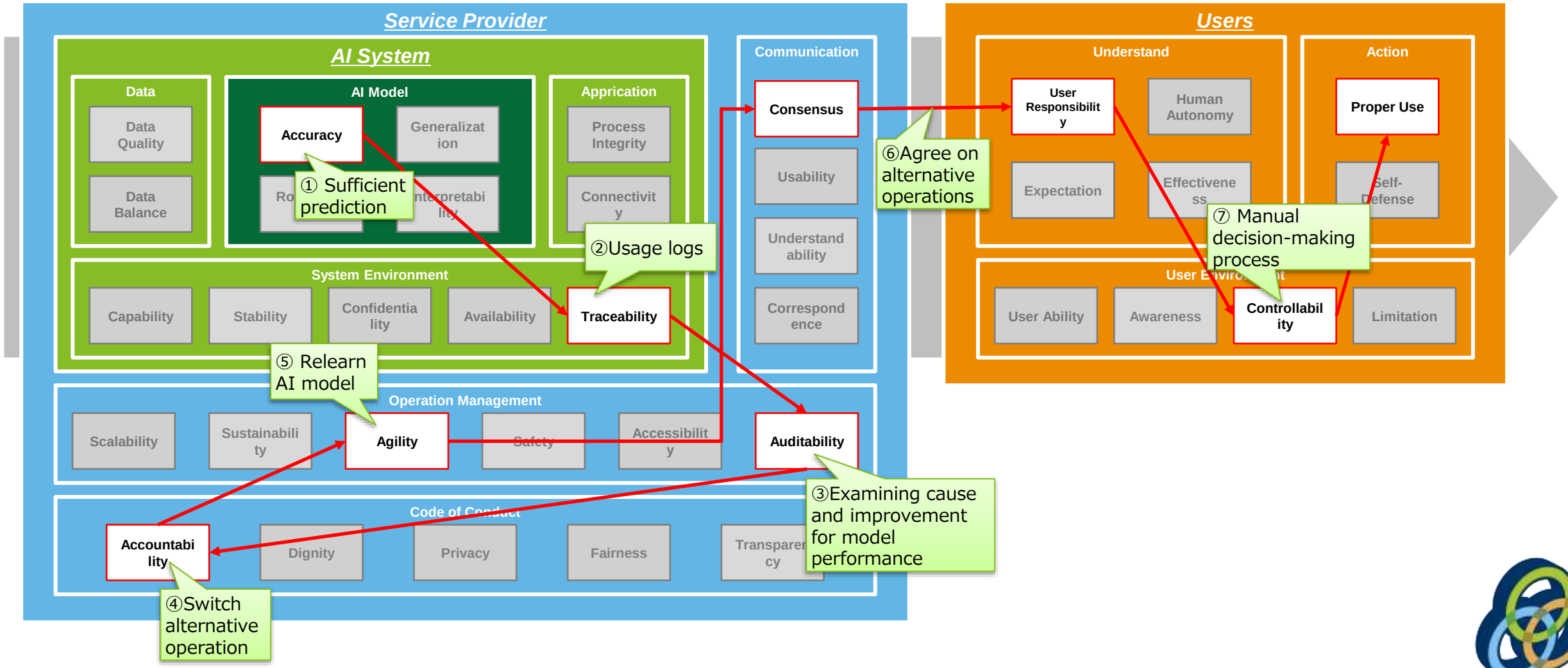
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R002

Unstable performance
Recruiting level decreases owing to deterioration of AI prediction performance



Risk Control

- Consider risk control according to the risk chain -

Step4

R002	**Unstable performance** Recruiting level decreases owing to deterioration of AI prediction performance

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
①[Accuracy] Develop models with sufficient prediction (AI dev dept., Co. A) ②[Traceability] Store usage logs (IT dept., Co. A)	③[Auditability] Examining cause and improvement of model performance when the prediction accuracy of the model deteriorates (HR dept., Co. A) ④[Accountability] Switch alternative operation for sufficient service execution (HR dept., Co. A) ⑤[Agility] Relearning AI model (HR dept., Co. A) ⑥[Consensus] Agreement on alternative manual operations if relearning is not possible in time (HR dept., Co. A)	⑥[User Responsibility] Agree on alternative manual operations if relearning is not possible in time (Person in HR dept., group A) ⑦[Controllability/Proper Use] Alternative manual decision-making (person in HR dept., group A)

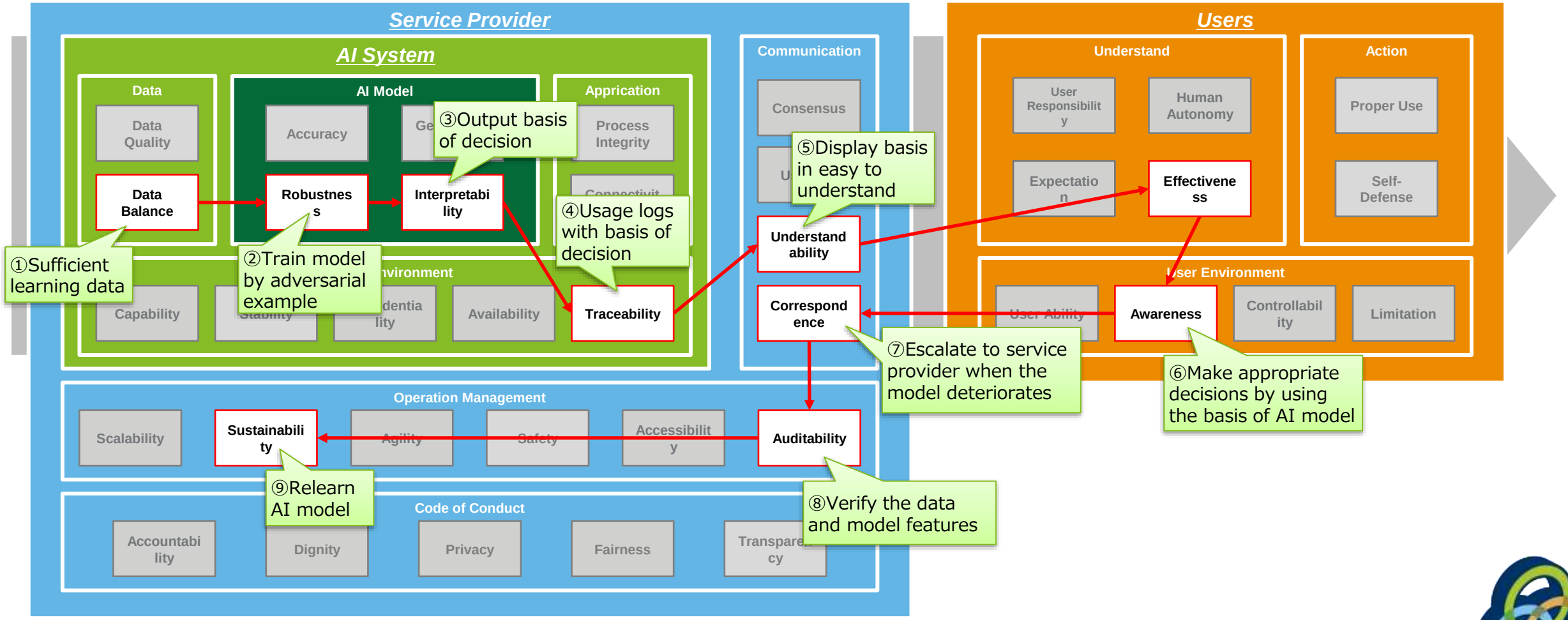
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R003

Impact by noise
AI decisions will change significantly by tiny different (e.g., punctuation marks)



Risk Control

- Consider risk control according to the risk chain -

Step4

R003	**Impact by noise** AI decisions will change significantly by tiny different (e.g., punctuation marks)

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
①[Data Balance] Sufficient learning data including adversarial examples (AI dev dept., Co. A) ②[Robustness] Training the model robustness using adversarial examples (AI dev dept., Co. A) ③[Interpretability] Output basis for the model decision (AI dev dept., Co. A) ④[Traceability] Usage logs with basis of decision (IT dept., Co. A)	⑤[Understandability] Display the basis for model decisions in an easy-to-understand manner (HR dept., Co. A) ⑦[Correspondence] Escalate to service provider when the model deteriorates (person in HR dept., group A/HR dept., Co. A) ⑧[Auditability] Verify the data and features that significantly affected the decision results (HR dept., Co. A/HR dept., group A.) ⑨[Sustainability] Relearning AI model (HR dept., Co. A).	⑥[Effectiveness/Awareness] Make appropriate decisions by using the basis for model decisions for evidence (person in HR dept., group A)

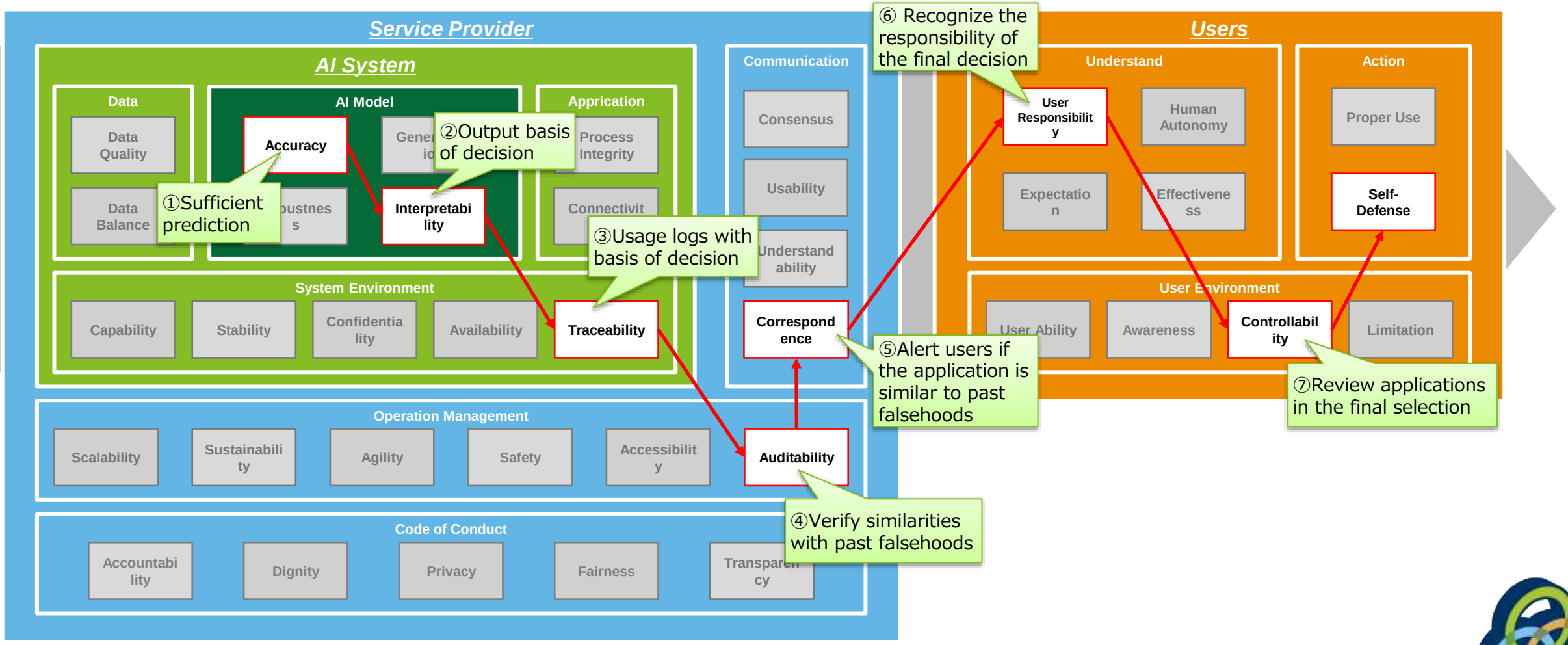
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R004

Falsehood
The application with falsehood is passed



Risk Control

- Consider risk control according to the risk chain -

Step4

R004	**Falsehood** The application with falsehood is passed

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
①[Accuracy] Develop models with sufficient prediction (AI dev dept., Co. A) ②[Interpretability] Output the basis for the model decision (AI dev dept., Co. A) ③[Traceability] Usage logs with basis of decision (IT dept., Co. A)	④[Auditability] Verify similarities with past falsehood cases (HR dept., Co. A) ⑤[Correspondence] Alert users if the application is similar to past falsehood applications (person in HR dept., group A/HR dept., Co. A)	⑥[User Responsibility] Recognize the responsibility of the final decision (person in HR dept., group A) ⑦[Controllability/Self-Defense] Review applications for falsity in the final selection process (person in HR dept., group A)

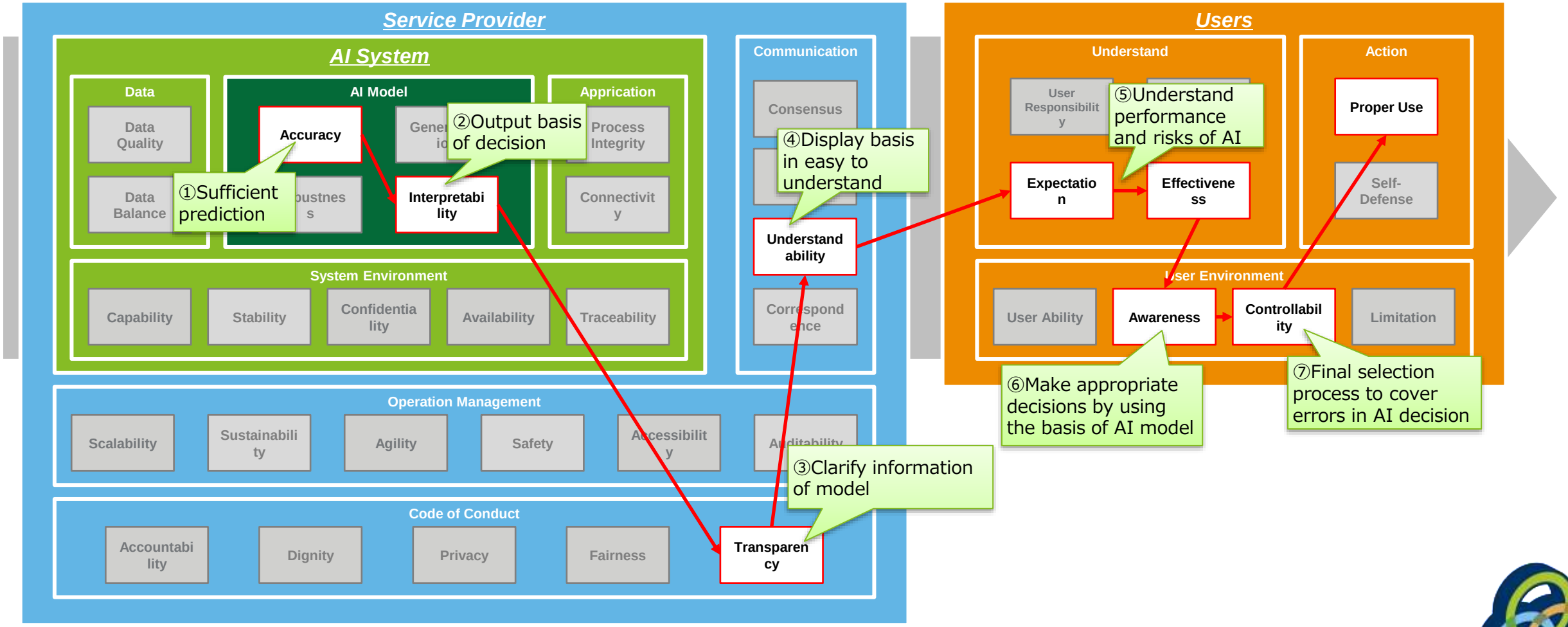
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R005

Excessive AI dependence
Person in charge of HR dept. relies excessively on AI decisions



Risk Control

- Consider risk control according to the risk chain -

Step4

R005	**Excessive AI dependence** Person in charge of HR dept. relies excessively on AI decisions

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
①[Accuracy] Develop models with sufficient prediction (AI dev dept., Co. A) ②[Interpretability] Output basis for the model decision (AI dev dept., Co. A)	③[Transparency] Clarify information of model performance and basis of decision (AI dev dept., Co. A) ④[Understandability] Display basis for model decisions in an easy-to-understand manner (HR dept., Co. A)	⑤[Expectation/Effectiveness] Understand performance and risks of AI (person in HR dept., group A) ⑥[Awareness] Make appropriate decisions by using the basis for model decisions for evidence (person in HR dept., group A) ⑦[Controllability/Proper Use] Prepare a final selection process to cover errors in AI decision and make the selection (person in HR dept., group A)

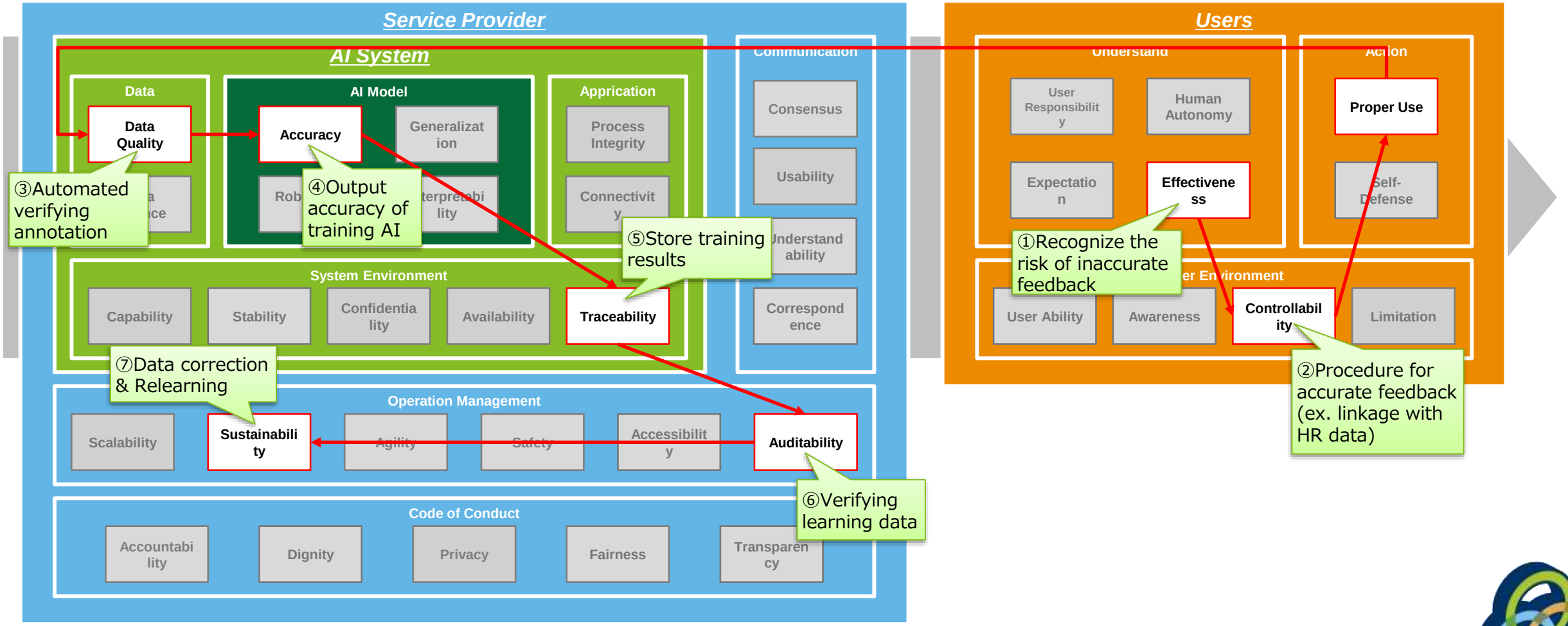
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R006

Inaccurate feedback
Inaccurate feedback (pass/fail labeling) to AI by HR dept. degrades AI performance



Risk Control

- Consider risk control according to the risk chain -

Step4

R006	**Inaccurate feedback** Inaccurate feedback (pass/fail labeling) to AI by HR dept. degrades AI performance

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
③[Data Quality] (Group company, if possible) Automated verification annotations for accuracy on a regular basis with external systems (IT dept., Co. A) ④[Accuracy] Output accuracy of the model during training AI (AI dev dept., Co. A) ⑤[Traceability] Store the prediction results and anomalies at each training stage (IT dept., Co. A)	⑥[Auditability] Verify changes in prediction performance and anomalies in learning data (data that seem to be mislabeled) (HR dept., Co. A/HR dept., group A.) ⑦[Sustainability] Data correction and relearning (HR dept., Co. A)	①[Effectiveness] Recognize that inaccurate feedback deteriorates the performance of AI model (person in HR dept., group A) ②[Controllability/Proper Use] Procedure to ensure accurate feedback (e.g., linkage with HR system) (person in HR dept., group A)

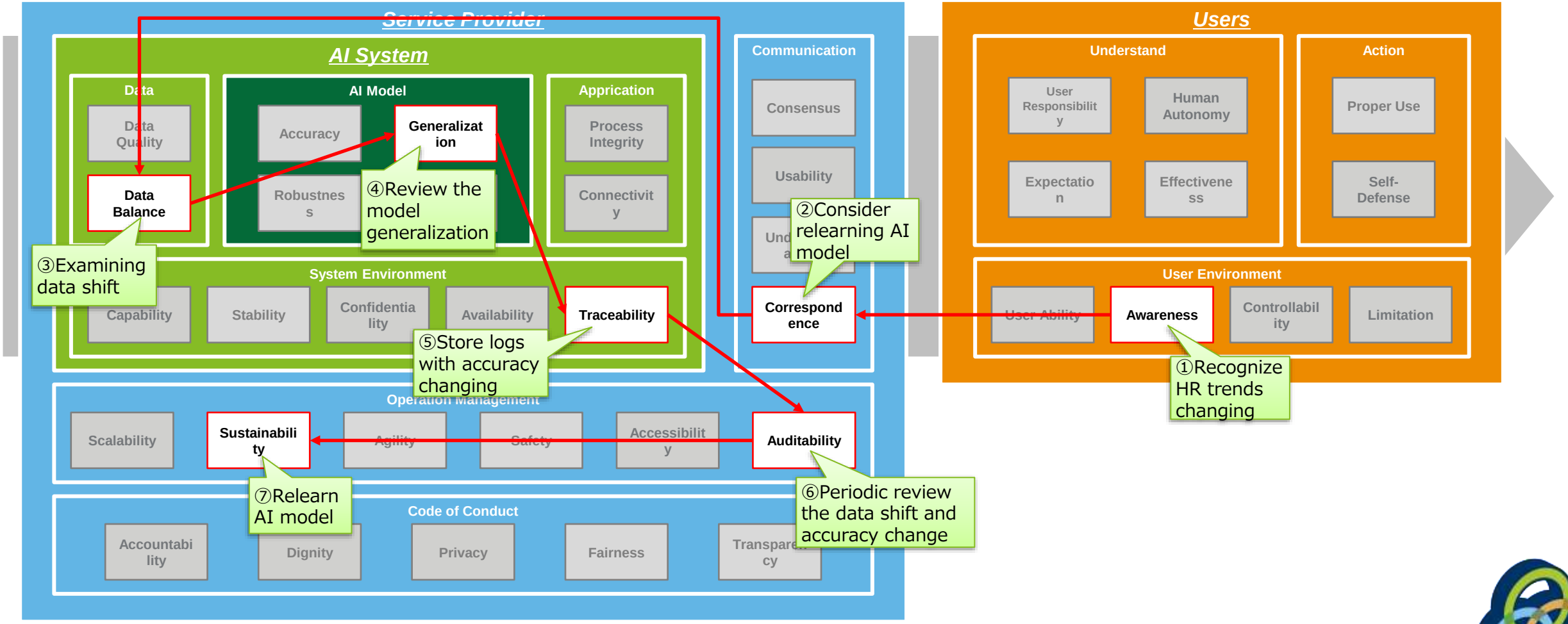
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R007

Change in HR trends
AI model cannot adapt to changes in HR trends of talent required



Risk Control

- Consider risk control according to the risk chain -

Step4

R007	**Change in HR trends** AI model cannot adapt to changes in HR trends of talent required

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
③[Data Balance] Examining data shift (AI dev dept., Co. A) ④[Generalization] Review the model generalization (AI dev dept., Co. A) ⑤[Traceability] Store logs with accuracy changing (IT dept., Co. A)	②[Correspondence] Consider relearning AI model when HR trends change significantly (HR dept., Co. A/ HR dept., group A.) ⑥ [Auditability] Periodic review of the data shift and accuracy change to determine relearning (HR dept., Co. A/HR dept., group A.) ⑦ [Sustainability] Request relearning the AI model to ensure continuous prediction accuracy and generalization performance (HR dept., Co. A)	①[Expectation] Recognize changing HR trends (person in HR dept., group A)

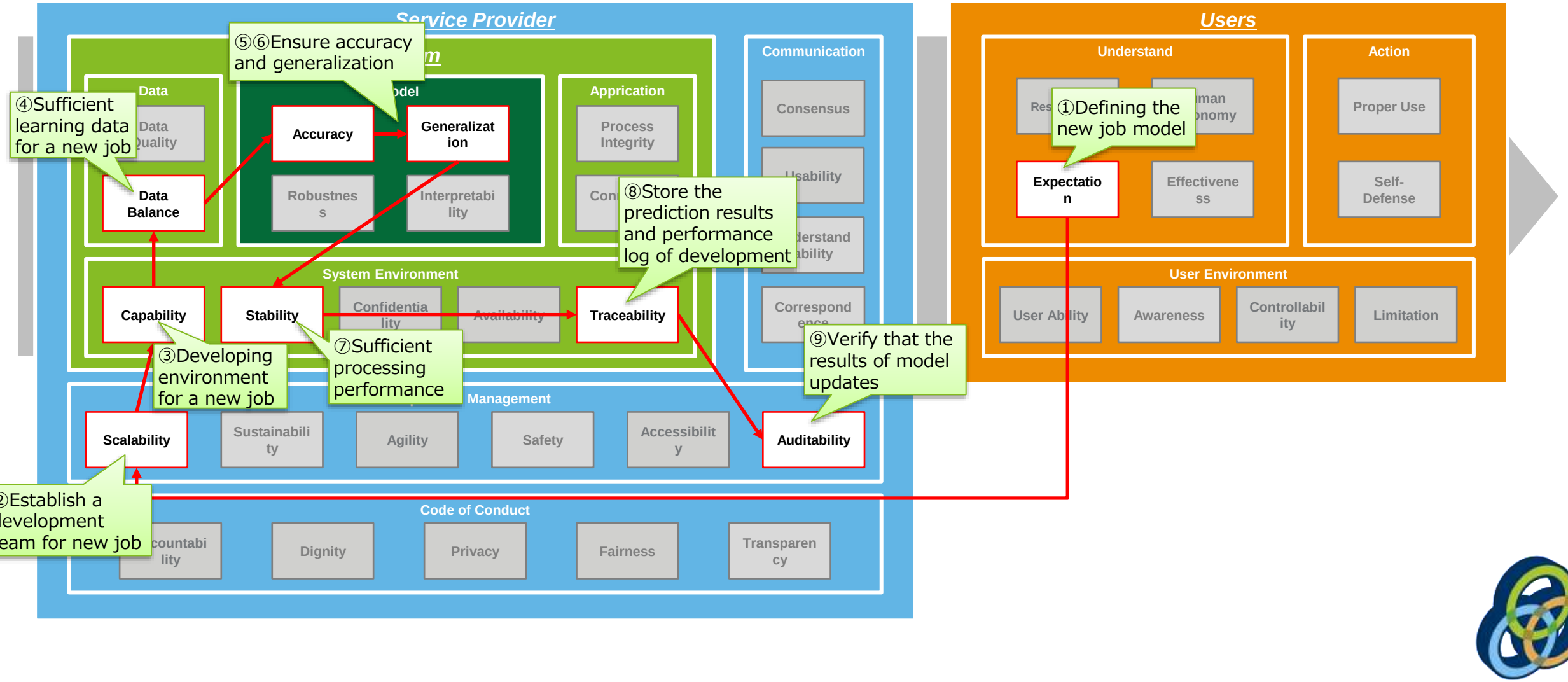
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R008

New occupation
AI model cannot achieve sufficient prediction when seeking new occupations



Risk Control

- Consider risk control according to the risk chain -

Step4

R008	**New occupation** AI model cannot achieve sufficient prediction when seeking new occupations

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] Environment to develop models for new job model (IT dept., Co. A) ④[Data Balance] Sufficient learning data for new job model (AI dev dept., Co. A) ⑤⑥[Accuracy/Generalization] Ensure accuracy and generalization performance by including data of new jobs (AI dev dept., Co. A) ⑦[Stability] Sufficient processing performance of the system environment when the models are added or updated (IT dept., Co. A) ⑧[Traceability] Store the prediction results and performance log at training stage (AI dev dept., Co. A)	②[Scalability] Establish a development team for new job model (HR dept., Co. A) ⑨[Auditability] Verify that the results of model updates are acceptable for service delivery (HR dept., Co. A)	①[Expectation] Defining the requirement for a new job model (person in HR dept., group A)

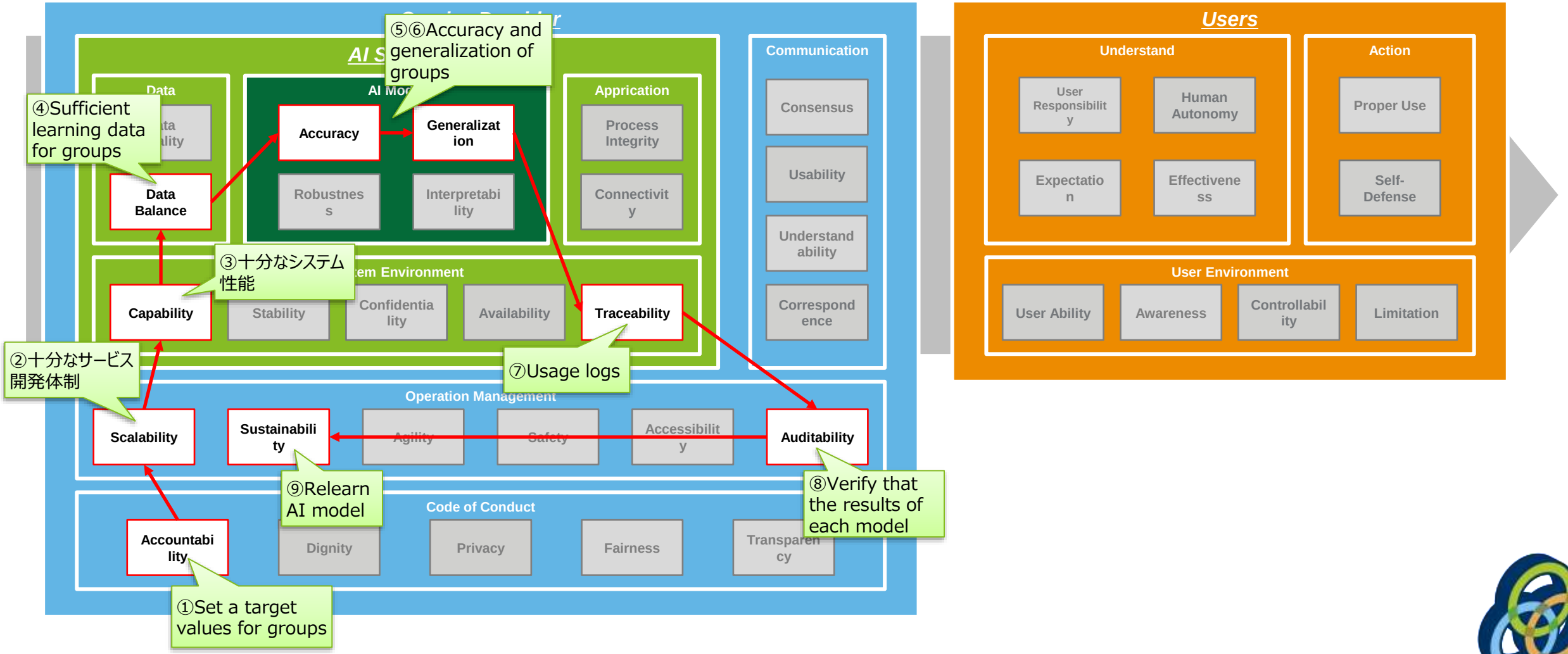
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R010

Differences of groups
AI does not work effectively owing to differences in the circumstances of group companies



Risk Control

- Consider risk control according to the risk chain -

Step4

R010	**Differences of groups** AI does not work effectively owing to differences in the circumstances of group companies

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] Environment to develop models for each group and region (AI dev dept., Co. A) ④[Data Balance] Sufficient learning data for each group and region (AI dev dept., Co. A) ⑤⑥[Accuracy/Generalization] Ensure accuracy and generalization performance by including data of groups (AI dev dept., Co. A) ⑦[Traceability] Store usage logs (AI dev dept., Co. A)	①[Accountability] Set appropriate target values for each group and region (HR dept., Co. A/HR dept., group A.) ②[Scalability] Establish a development team for each group and region (HR dept., Co. A) ⑧[Auditability] Verify that the results of each model (HR dept., Co. A/HR dept., group A.) ⑨[Sustainability] Request relearning the AI mode (HR dept., Co. A)	

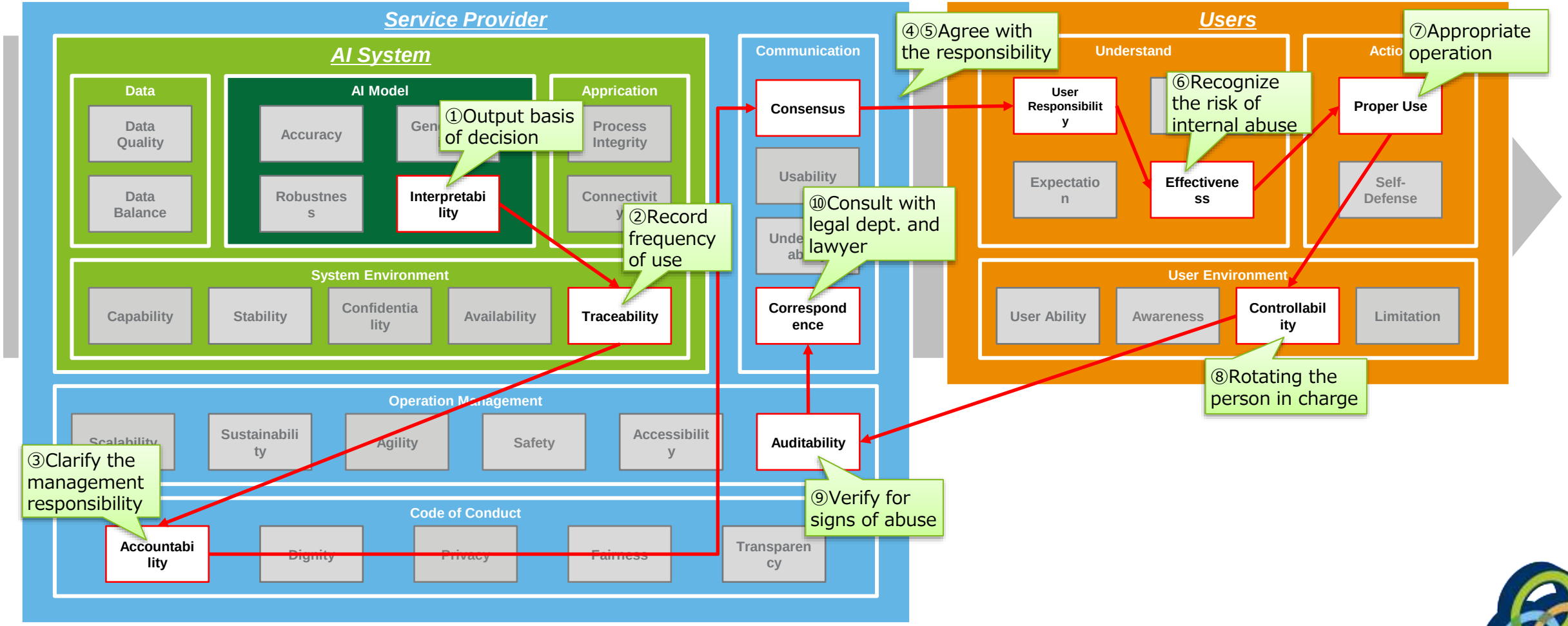
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

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



Risk Control

- Consider risk control according to the risk chain -

Step4

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

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
①[Interpretability] Output the basis for the model decision (AI dev dept., Co. A) ②[Traceability] Record frequency of use (AI dev dept., Co. A)	③[Accountability] Clarify the management responsibility for internal abuse (HR dept., Co. A) ④[Consensus] Agree with groups that they have responsibility for the management of internal abuse (HR dept., Co. A) ⑨[Auditability] Verification for signs of abuse, such as operation history outside of work or reading the same entry sheet multiple times (HR dept., Co. A) ⑩[Correspondence] Consulting with legal dept. and lawyer when internal abuse is discovered (HR dept., Co. A／Legal dept., Co. A)	⑤[User Responsibility] Agreement with groups that they have responsibility for the management of internal abuse (HR dept., group A.) ⑥[Effectiveness] Recognize the risks associated with unauthorized use and inform the department about the penalties when they occur (HR dept., group A.) ⑦[Proper Use] Conduct recruitment operations appropriately (person in HR dept., group A) ⑧[Controllability] Reduce the risk of fraud by rotating the person in charge of contacting external agents (HR dept., group A.)

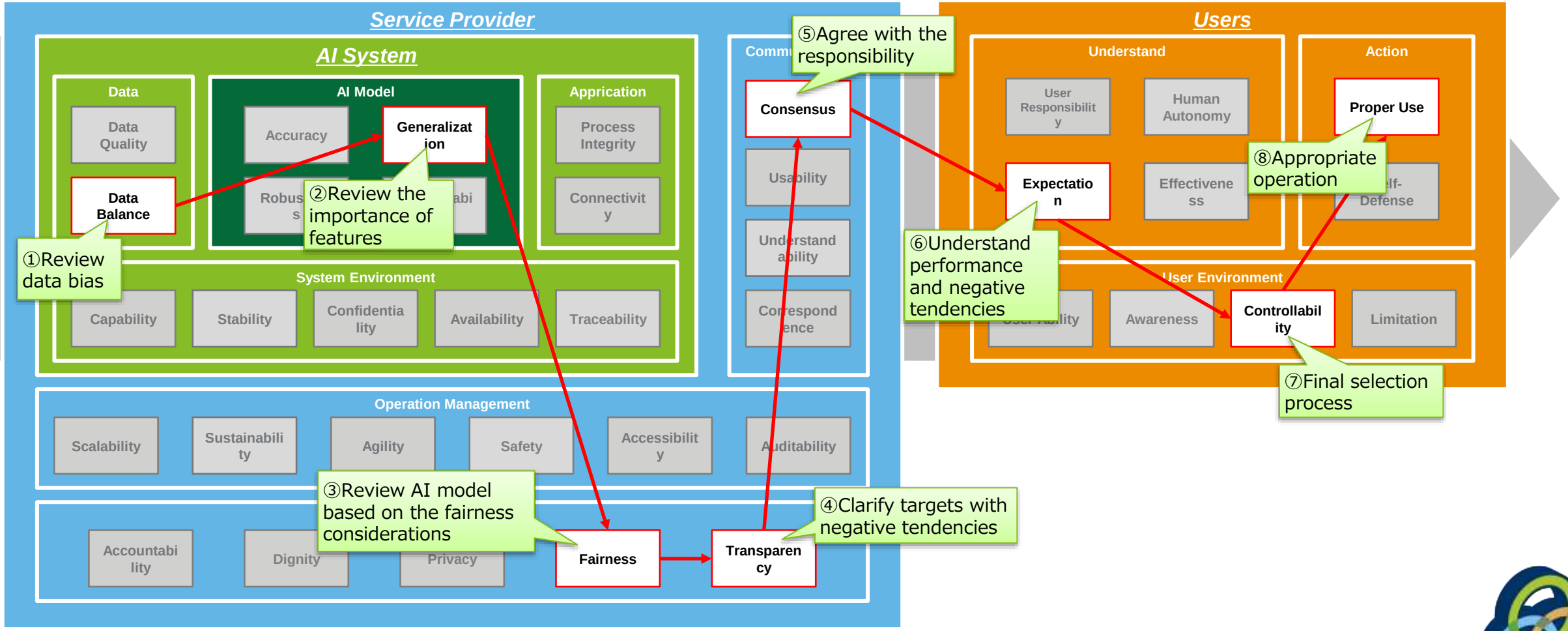
Control Coordination

- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R013

Fairness
Unfair forecast results for a particular group



Risk Control

- Consider risk control according to the risk chain -

Step4

R013	**Fairness** Unfair forecast results for a particular group

Risk Control		
AI System (AI dev dept., Co. A)	**AI Service Provider (HR dept., Co. A)**	**User (Person in HR dept., group A)**
①[Data Balance]Review data bias (AI dev dept., Co. A) ②[Generalization]Review the importance of features (AI dev dept., Co. A)	③[Fairness] Review tendency of AI model to make decisions, based on the fairness considerations (HR dept., Co. A) ④[Transparency] Clarify targets that cannot be excluded from negative decisions owing to lack of data or other reasons (HR dept., Co. A) ⑤[Consensus] Agreement with groups that they have responsibility for final decision (HR dept., Co. A)	⑥[Expectation] Understanding predictive performance and negative decision tendencies (HR dept., group A.) ⑦[Controllability] Consider that humans should make decisions (HR dept., group A.) ⑧[Proper Use] Conduct recruitment operations appropriately (Person in HR dept., group A)