RCM
RiskChainModel

Risk Assessment & Control Coordination for AI services :

Case06 Verification of Recidivism Possibility AI

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

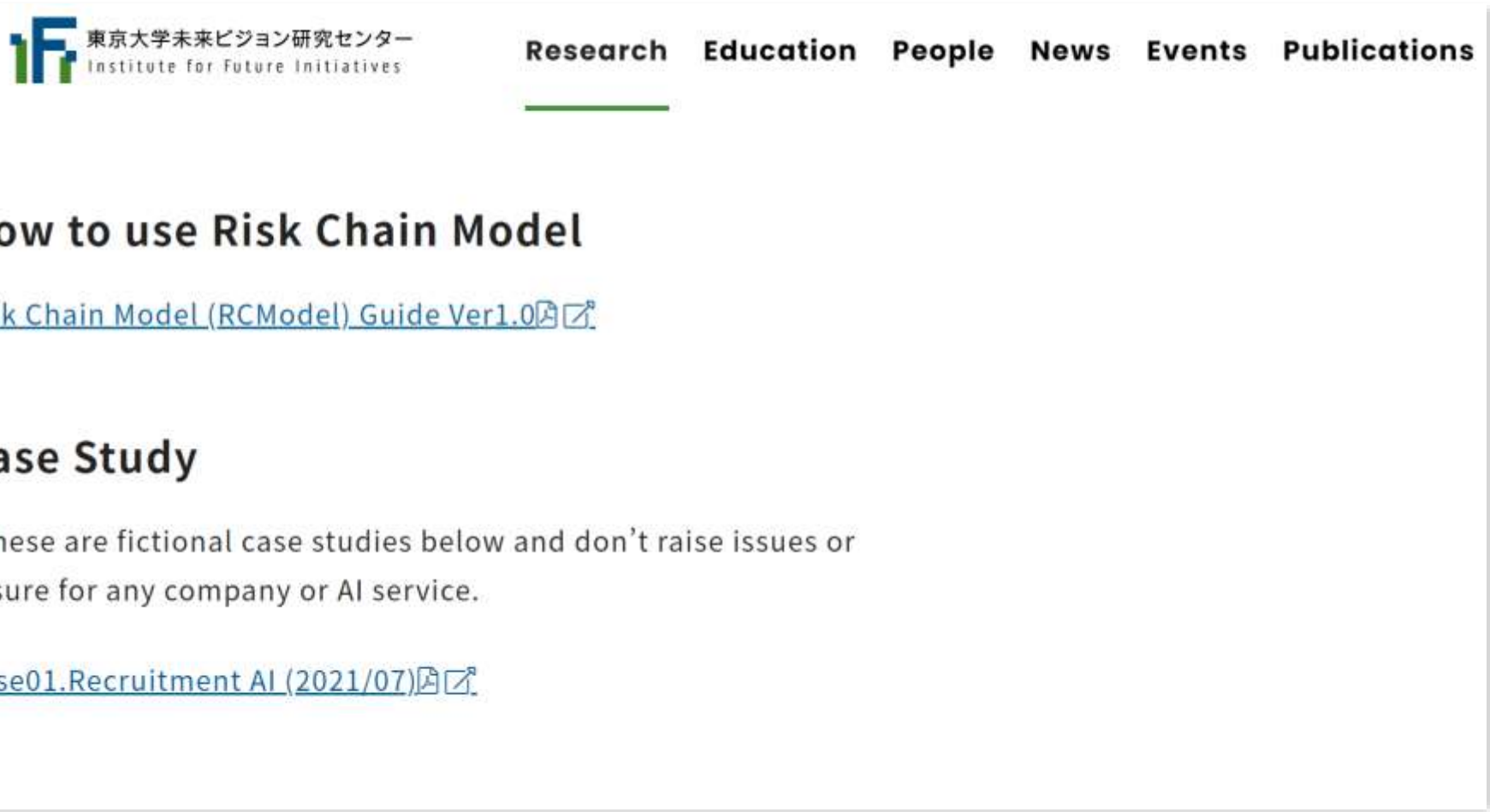
How to operate the RCModel
- Risk Assessment & Control Coordination -

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

Case06 : Verification of Recidivism Possibility AI

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

Step1

This machine-learning AI model predicts the possibility that a defendant will commit recidivism after release.

Based on questions asked by defendants and their criminal records, the AI is used to predict the possibility of recidivism (on a scale of 10) and is used by prosecutors and lay judges to determine the contents of judgments and parole decisions. It is also used for criminal investigations by the police and for the observation of parolees by probation officers.

[Values & Objectives]

- **Decrease in recidivism**
- **Proper use by each user**
- **Social responsibility**

[Flow of Actual Operations using AI Services]

① In each court, the defendant asks the questions set in the system. Questions inquire of crime, bail history, age, employment status, livelihood, education level, community connection, drug use, beliefs, family criminal history, drug use history, etc.

② By inputting the defendant’s past criminal history data and the results of ① into the AI, the possibility of the defendant's recidivism is determined in a score of 10 levels. Information on the basis of judgment (e.g., areas of interest in questionnaire responses, and examples of similar past judgments) are output along the possibility scale of recidivism.

③ Lay judges in district courts use the results of the possibility of recidivism offenses by AI as a reference opinion to determine the term of imprisonment and other matters for defendants.

✓ It can be viewed by law enforcement investigators and probation officers.

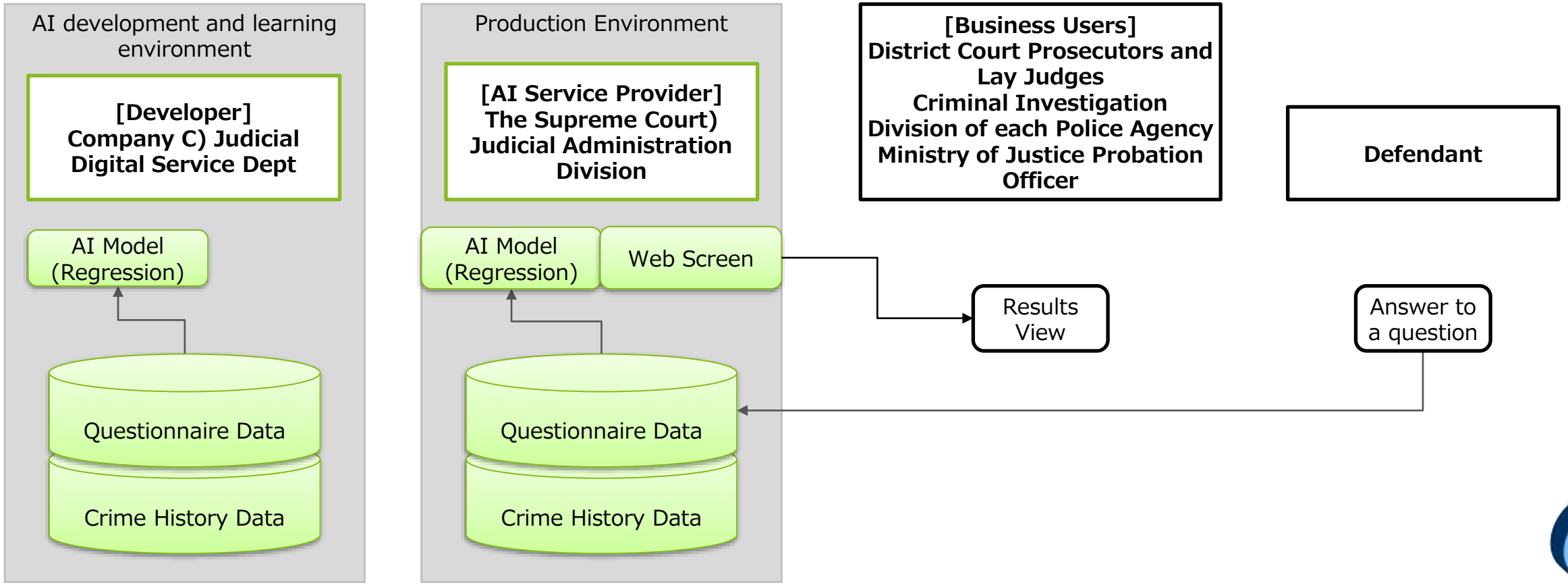
✓ Company C, which undertakes development, will use the Ministry of Justice's criminal and civil criminal history database to create an AI model common to district courts. AI model training is carried out in the development environment of Co. C, and the crime history database is added as learning data periodically to relearn the AI model.

✓ The AI service provider is the judicial administration division of the Supreme Court, which reviews the AI model updaters and provides instructions to Co. C as necessary. The actual AI models can be accessed and viewed on the web by individual users (e.g., prosecutors, courts, police, and probation officers).

Case06 : Verification of Recidivism Possibility AI

- System Overview -

AI System	Company C) Judicial Digital Service Dept	Develop and learn AI models
AI Service Provider	The Supreme Court) Judicial Administration Division	Deliver AI models from the web to the right audience
User	District Court) Prosecutor Each Police Agency) Criminal Investigation Dept Ministry of Justice) Probation Officer	Obtaining information by viewing the results of AI model decisions from the web ※The main purpose of use is the review of judgments by each court



Case06 : Verification of Recidivism Possibility AI

- Input & Output -

[Input Data]

Data	Purpose	Collection Method	Data Manager	Including Privacy Data
Past entry sheet data	Learning	Question to the defendant (Not required)	Company C server	Yes (including Special Care-required Personal Information)
Newest entry sheet data	Learning	Historical crime data	Ministry of Justice Database	Yes (including Special Care-required Personal Information)

[Output]

Users	Prosecutors, courts, police and probation officers
Output	Prediction of repeat offenses based on 10 grade evaluation
Output Method	By entering information about the accused or the subject of a criminal investigation on the Web, you can view the probability of recidivism score
Expected Accuracy	75% *Percentage of those considered to be at high risk who actually repeated offenses
User judgment	Yes (final judgment is the responsibility of each user)
Output of evidence information	Areas of interest in questionnaire responses, and examples of similar past judgments
Safety Risk	Yes (Although the final decision is the responsibility of each user, the risk of an increase in repeat offenses may arise due to erroneous judgment)
Connection with external system	No
Users	Prosecutors, courts, police and probation officers

RCM
RiskChainModel

Risk Assessment

Risk Assessment

- Examination of significant risk scenarios -

Step2

Values & Objectives		Service Requirement			Risk No.	Risk Scenario	
1	Decrease in recidivism	1-1	Ensuring predictive performance	■ Accuracy	R001	Misjudgment	Learning degrades the predictive performance of AI, leading to incorrect decisions
		1-2	Continuous guidance	■ Consistency	R002	Instable judgment	Inconsistent response due to major changes in AI decisions as they learn
		1-2	Appropriate judgment	■ Questionnaire item ■ Information management	R003	Explainability	The grounds for the final decision cannot be explained correctly
2	Proper use by each user	2-1	Clarification of the scope of use	■ Defining the scope of use ■ Access Control	R004	Unintended use	Penalize specific individuals by using AI models for unexpected purposes
		2-2	Prevention of misjudgments by users	■ Explainability ■ Fairness	R005	Excessive AI dependence	Rely on AI to make the wrong final decision
					R006	Misjudgment due to improper input	Inducing defendants to intentionally enter false information into questionnaires in order to reduce the possibility of recidivism
					R007	Loss of fairness	Produce clearly unfair predictions for specific races/sexes
3	Improvement of operational efficiency and reduction of burden	3-1	Moderate level of prediction	■ Output control	R008	Excessive judgment	AI outputs results that are too negative for everyone, overburdening users
		3-2	Less expensive data collection	■ Efficient data collection	R009	Increased burden of data collection	Large amounts of learning data are collected manually, overloading users
4	Social responsibility	4-1	Accountability	■ Explainability ■ Verifiability	R010	External explanation	Inability to adequately explain the decision process when requested to do so from outside
		4-2	Information management	■ Security assurance ■ Data management	R011	Hacking AI systems	Incorrect updates to learning data and predictive models, resulting in incorrect decisions for specific people or groups
					R012	Occurrence of harmful rumors	Inappropriate interpretations of AI model judgments are published and defames specific people or groups
					R013	Privacy protection	Violation of the Personal Information Protection Act by mishandling personal information

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	Decrease in recidivism	R001	Misjudgment	○			●	Prediction performance Recording execution results	Explanation of expected accuracy Relearning	Understanding of expected accuracy Ensuring necessary Knowledge Final judgment
		R002	Instable judgment	○	○		●	Examine data distribution Basis of decision	Understanding of changes in judgment grounds Relearning	Understanding of changes in Judgment and confession
		R003	Explainability	○		○	●	Basis of decision	Clarity of the grounds for judgment	Ensuring necessary knowledge Final judgment
2	Proper use by each user	R004	Unintended use			○		Data protection	Access control Proper use	Compliance
		R005	Excessive AI dependence	○		○	●	*Same as R003	*Same as R003	*Same as R003
		R006	Misjudgment due to improper input	○		○	●	Automatic cooperation of external data Validate learning data	Easy-to-use UI Verification of judgment accuracy and abnormal values Relearning	Feedback results
		R007	Loss of fairness	○			●	Data balance Impartiality of feature quantities	Arrangement of points to be considered for fairness Visualization of the rationale	Understanding of the fairness risk Review of the observation plan
3	Improvement of operational efficiency and reduction of burden	R008	Excessive judgment	○		○	●	Basis of decision Recording execution results	Consideration of detection levels Relearning	Final judgment
		R009	Increased burden of data collection			○		Simplifying data collection methods	Cost effectiveness study	Burden feedback
4	Social responsibility	R010	External explanation	○			●	Recording data understanding Record of model performance Recording execution results	Organize information to be disclosed Access control Audit response	
		R011	Hacking AI systems					Security management	Investigation and improvement of causes	Protecting user's environment
		R012	Occurrence of harmful rumors					Data protection	Compliance	Compliance
		R013	Privacy protection					Data protection	Compliance	Compliance

Organization

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

Step5

Chief Justice of the Supreme Court

- Consideration of values and objectives to be realized
- Approval of risk control methods
- External explanation

- AI Service Provider -
The Supreme Court)
Judicial Administration Division

- Explanation of expected accuracy
- Understanding of changes in judgment grounds
- Arrangement of points to be considered for fairness
- Clarity of the grounds for judgment
- Consideration of detection levels
- Relearning
- Cost effectiveness study
- Proper use
- Organize information to be disclosed
- Performance monitoring
- Access control
- Audit response
- Investigation and improvement of causes
- Compliance

C Co) Judicial Digital Service Dept.

- Prediction performance
- Basis of decision
- Impartiality of feature quantities
- Examine data distribution
- Data balance
- Recording data understanding
- Simplifying data collection methods
- Easy-to-use UI

C Co) IT Service Dept.

- Recording execution results
- Record of model performance
- Data protection
- Security management
- performance maintenance
- Protecting user’s environment

- User -
District Court) Prosecutor
Each Police Agency) Criminal Investigation Dept
Ministry of Justice) Probation Officer

- Final judgment
- Understanding changes in forecast results
- Review observation plan for overdetection
- Understanding of expected accuracy
- Understanding of the fairness risk
- Ensuring the necessary literacy
- Feedback results
- Compliance
- Alternative operation

Defendant/Subject of Observation, etc.

RCM
RiskChainModel

Control Coordination

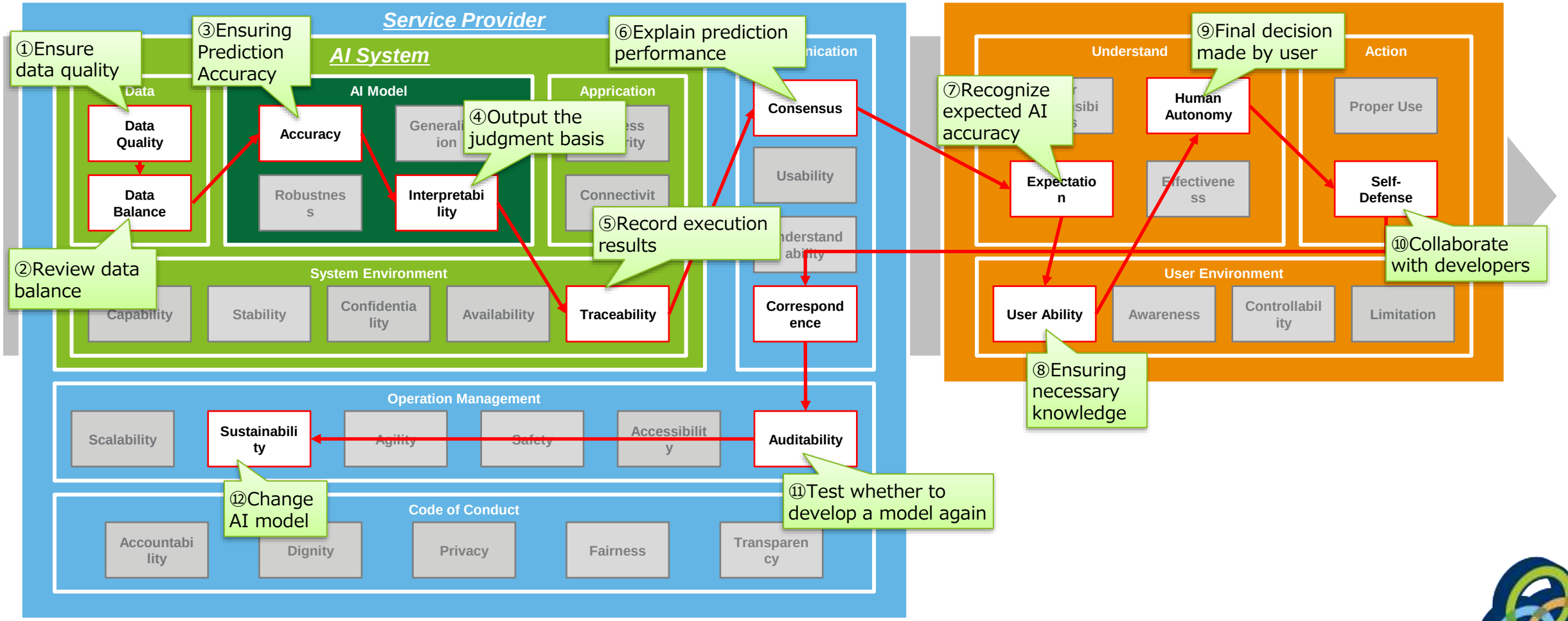
Control Coordination

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

Step3

R001

Misjudgment
Learning degrades the predictive performance of AI, leading to incorrect decisions



Risk Control

- Consider risk control according to the risk chain -

Step4

R001	**Misjudgment** Learning degrades the predictive performance of AI, leading to incorrect decisions

Risk Control		
AI system (Judicial Digital Service Dept, Co. C)	**Service Provider (Judicial Administration Dept of the Supreme Court)**	**User (Public Prosecutor/Police/Probation Officer)**
① [Data Quality] Corrects data collected by mistake and ensures the quality (Judicial Digital Service Dept, Co. C) ② [Data Balance] Correct data bias (Judicial Digital Service Dept, Co. C) ③ [Accuracy] Ensure necessary prediction performance (Judicial Digital Service Dept, Co. C) ④ [Interpretability] Outputs Information on the grounds of judgment (Judicial Digital Service Dept, Co. C) ⑤ [Traceability] Record execution results (Judicial Digital Service Dept, Co. C)	⑥ [Consensus] Exploring AI's Expected Accuracy to Users (Judicial Administration Dept) ⑩ [Correspondence] Collaborate with developers when prediction accuracy is in doubt (Judicial Administration Dept) ⑪ [Auditability] Whether models should be developed based on new hypotheses (Judicial Administration Dept) ⑫ [Sustainability] Changing the AI model (Judicial Administration Dept + Judicial Digital Service Dept, Co. C)	⑦ [Expectation] Recognizes AI's expected accuracy (User) ⑧ [User Ability] Hold study sessions and share case studies to ensure the knowledge and literacy necessary for final decisions (User) ⑨ [Human Autonomy] The final decision is made by the user (User) ⑩ [Self-Defense] Engage developers when prediction accuracy is in doubt (User)

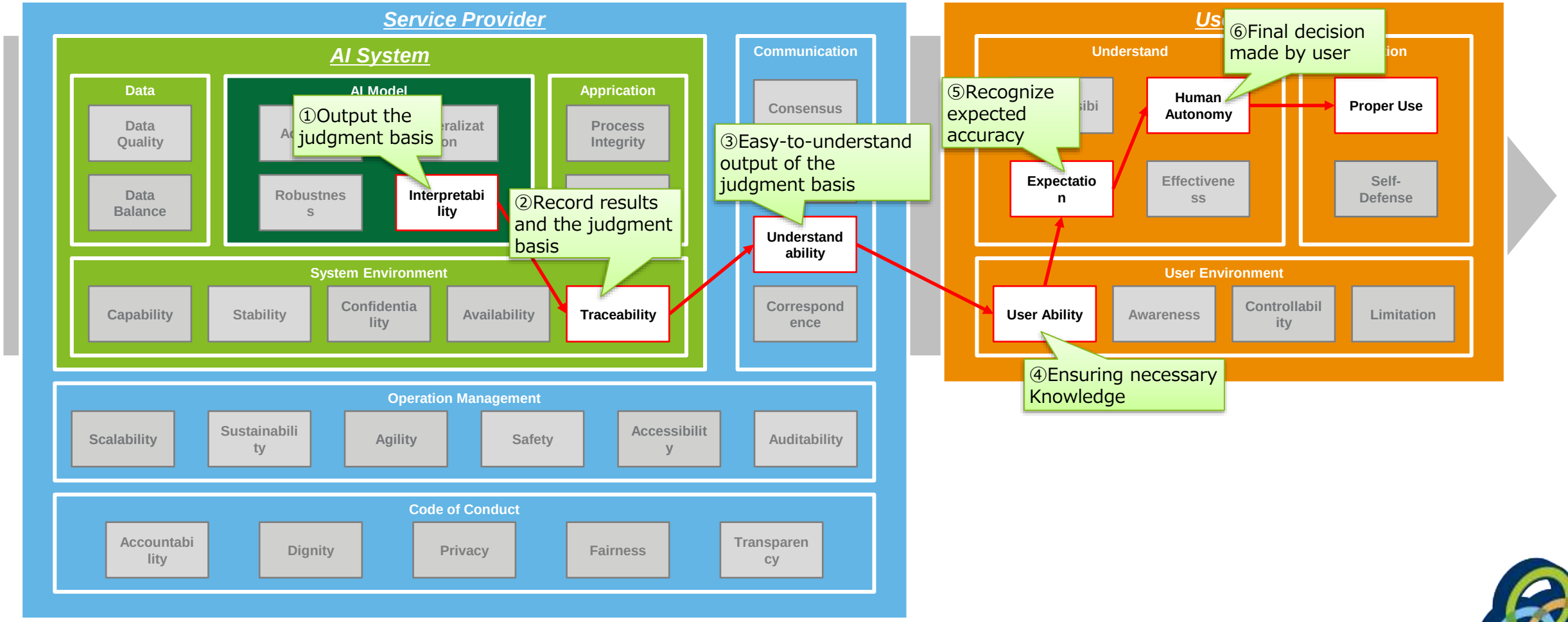
Control Coordination

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

Step3

R002

Instable judgment
Inconsistent response due to major changes in AI decisions as they learn



Risk Control

- Consider risk control according to the risk chain -

Step4

R002	**Instable judgment** Inconsistent response due to major changes in AI decisions as they learn

Risk Control		
AI system (Judicial Digital Service Dept, Co. C)	**Service Provider (Judicial Administration Dept of the Supreme Court)**	**User (Public Prosecutor/Police/Probation Officer)**
① [Data Balance] Confirmation of changes in data distribution (Office of Judicial Digital Services, Co. C) ② [Accuracy] Ensure necessary prediction accuracy (Judicial Digital Service Dept, Co. C) ③ [Interpretability] Outputs judgment basis information (Judicial Digital Service Dept, Co. C) ④ [Traceability] Record execution results and Information on the grounds of judgment (Judicial Digital Service Dept, Co. C)	⑤ [Understanding] Information on the grounds of judgment is output so that changes from the previous one can be seen (Judicial Administration Dept) ⑧ [Correspondence] Collaborate with developers when Information on the grounds of judgment is in doubt (Judicial Administration Dept) ⑨ [Auditability] Whether models should be developed based on new hypotheses (Judicial Administration Dept) ⑩ [Sustainability] Changing the AI model (Judicial Administration Dept + Judicial Digital Service Dept, Co. C)	⑥ [Effectiveness] Consider the adverse effects of inconsistent judgments (User) ⑦ [Human Autonomy] Make final decisions based on changes to the forecast target (User) ⑧ [Self-Defense] Engage with developers when Information on the grounds of judgment is in doubt (User)

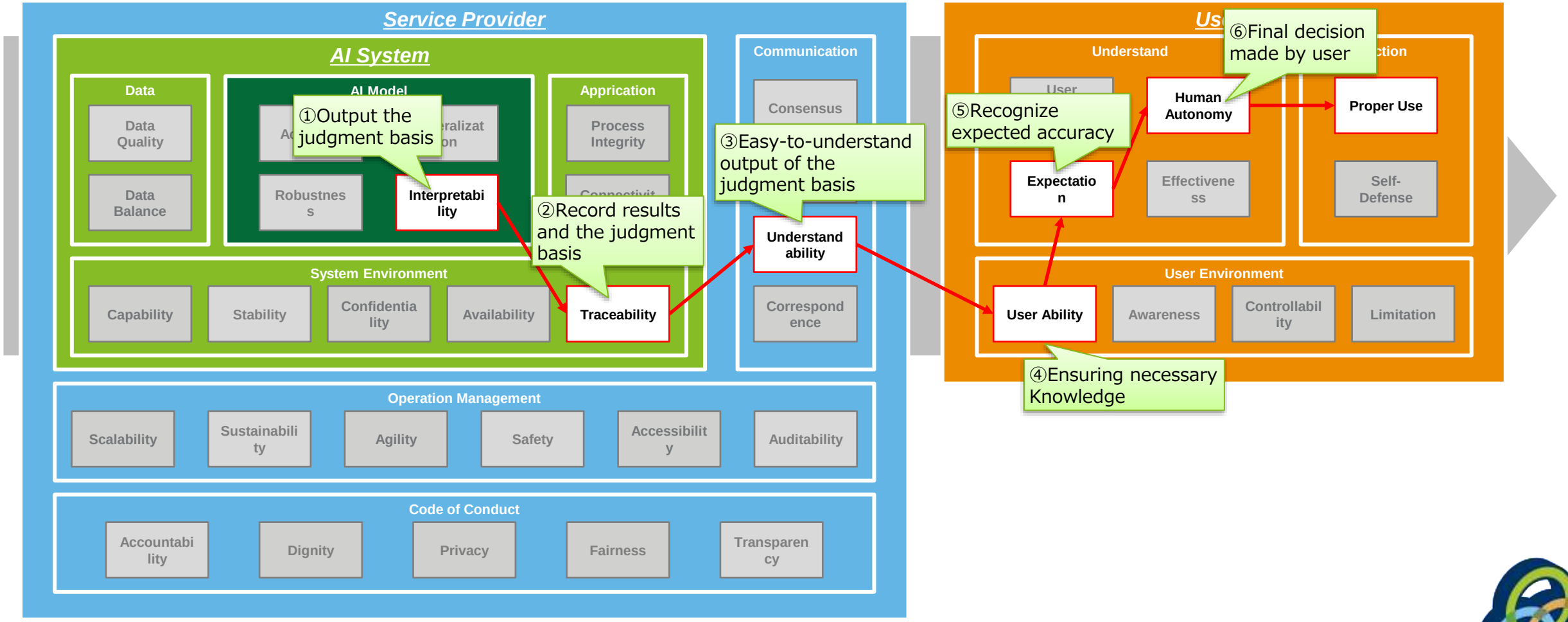
Control Coordination

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

Step3

R003

Explainability
The grounds for the final decision cannot be explained correctly.



Risk Control

- Consider risk control according to the risk chain -

Step4

R003	**Explainability** The grounds for the final decision cannot be explained correctly.

Risk Control		
AI system (Judicial Digital Service Dept, Co. C)	**Service Provider (Judicial Administration Dept of the Supreme Court)**	**User (Public Prosecutor/Police/Probation Officer)**
① [Interpretability] Outputs Information on the grounds of judgment (Judicial Digital Service Dept, Co. C) ② [Traceability] Record execution results and Information on the grounds of judgment (Judicial Digital Service Dept, Co. C)	③ [Understanding] To output Information on the grounds of judgment in an easy-to-understand manner (Judicial Administration Dept)	④ [User Ability] Hold study sessions and share case studies to ensure the knowledge and literacy necessary for final decisions (User) ⑤ [Expectation] Recognizes AI's expected accuracy (User) ⑥ [Human Autonomy] The final decision is made by the user (User) ⑦ [Proper Use] Make appropriate final decisions (User)

Control Coordination

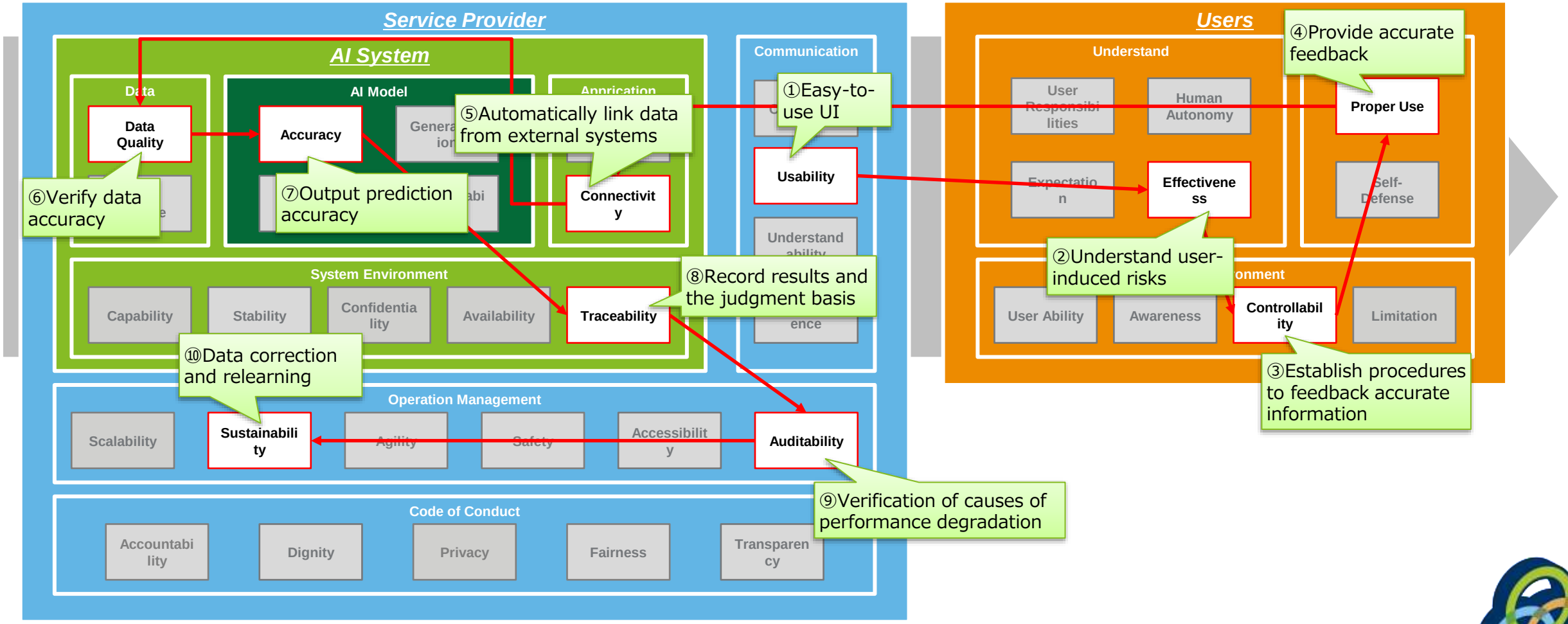
- Examine the risk chain (relation of risk factors) for each important risk scenario -

Step3

R006

Misjudgment due to improper input

Inducing defendants to intentionally enter false information into questionnaires in order to reduce the possibility of recidivism



Risk Control

- Consider risk control according to the risk chain -

Step4

R006	**Misjudgment due to improper input** Inducing defendants to intentionally enter false information into questionnaires in order to reduce the possibility of recidivism

Risk Control		
AI system (Judicial Digital Service Dept, Co. C)	**Service Provider (Judicial Administration Dept of the Supreme Court)**	**User (Public Prosecutor/Police/Probation Officer)**
⑤ [Connectivity] Automated linkage of possible data from external systems (Judicial Digital Service Dept, Co. C) ⑥ [Data Quality] Verifying the accuracy of learning data (Judicial Digital Service Dept, Co. C) ⑦ [Accuracy] Ensuring the accuracy of model judgments (Judicial Digital Service Dept, Co. C) ⑧ [Traceability] Storing the contents of prediction results and abnormal values at each learning stage (Judicial Digital Service Dept, Co. C)	① [Usability] Prepare a user-friendly UI that is hard to misoperate (Judicial Administration Dept + Judicial Digital Service Dept, Co. C) ⑨ [Auditability] Verification of changes in prediction performance and abnormal values in learning data (data that seems to be label errors) (Judicial Administration Dept + Judicial Digital Service Dept, Co. C) ⑩ [Sustainability] Correct teacher labels in data and relearn AI models as needed (Judicial Administration Dept + Judicial Digital Service Dept, Co. C)	② [Effectiveness] Recognizing that feedback errors affect the performance degradation of AI models (User) ③ [Controllability] Set up procedures to provide accurate feedback (User) ④ [Proper Use] Accurate Feedback (User)

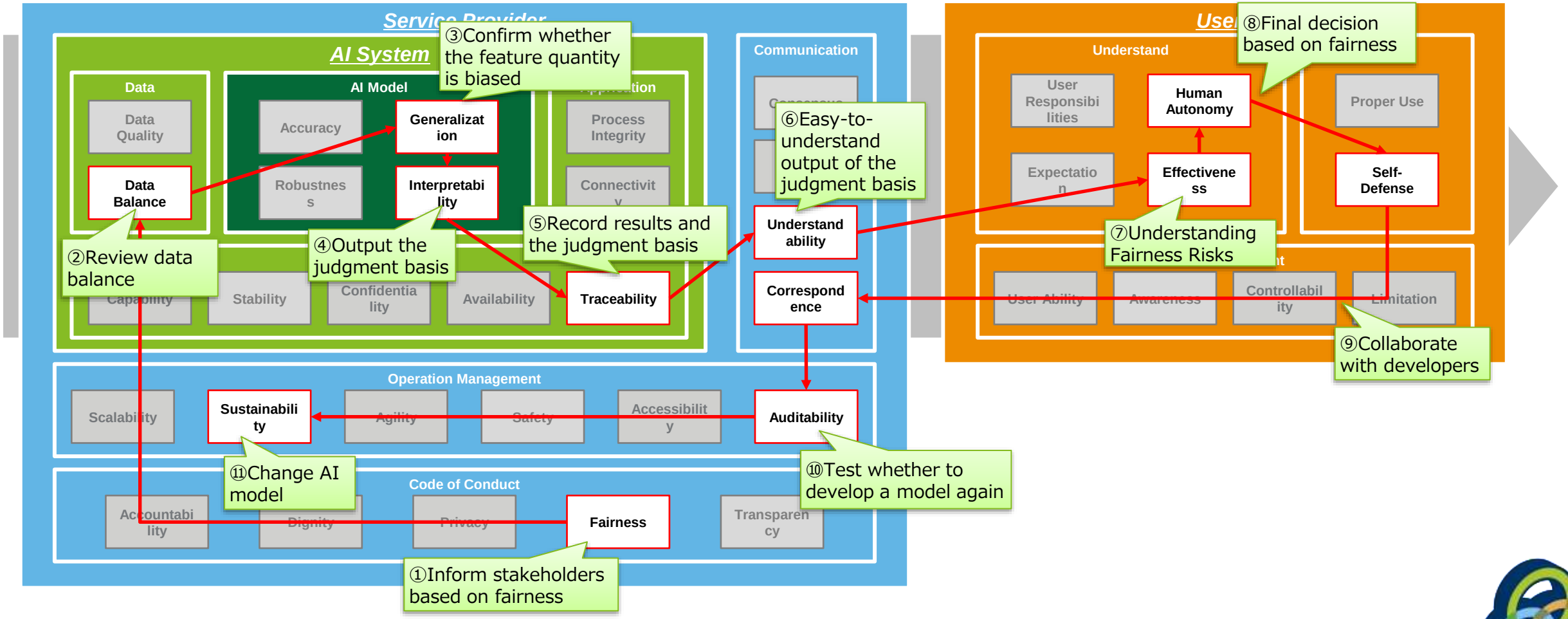
Control Coordination

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

Step3

R007

Fairness
Produce clearly unfair predictions for specific races/sexes



Risk Control

- Consider risk control according to the risk chain -

Step4

R007	**Fairness** Produce clearly unfair predictions for specific races/sexes

Risk Control		
AI system (Judicial Digital Service Dept, Co. C)	**Service Provider (Judicial Administration Dept of the Supreme Court)**	**User (Public Prosecutor/Police/Probation Officer)**
② [Data Balance] Review the balance of data (Judicial Digital Service Dept, Co. C) ③ [Generalization] Check whether the importance of the feature quantity is biased to a single item (Judicial Digital Service Dept, Co. C) ④ [Interpretability] Outputs Information on the grounds of judgment (Judicial Digital Service Dept, Co. C) ⑤ [Traceability] Record execution results and Information on the grounds of judgment (Judicial Digital Service Dept, Co. C)	① [Fairness] Based on the consideration of fairness, confirm that there is no problem in the tendency of the model's judgment (Judicial Administration Dept) ⑥ [Understanding] To output Information on the grounds of judgment in an easy-to-understand manner (Judicial Administration Dept) ⑨ [Correspondence] Collaborate with developers when fairness is questionable (Judicial Administration Dept) ⑩ [Auditability] Whether models should be developed based on new hypotheses (Judicial Administration Dept) ⑪ [Sustainability] Changing the AI model (Judicial Administration Dept + Judicial Digital Service Dept, Co. C)	⑦ [Effectiveness] Understanding Fairness Risks (User) ⑧ [Human Autonomy] Make a final decision based on the fairness of the judgment grounds (User) ⑨ [Self-Defense] Engage developers when fairness is in doubt (User)

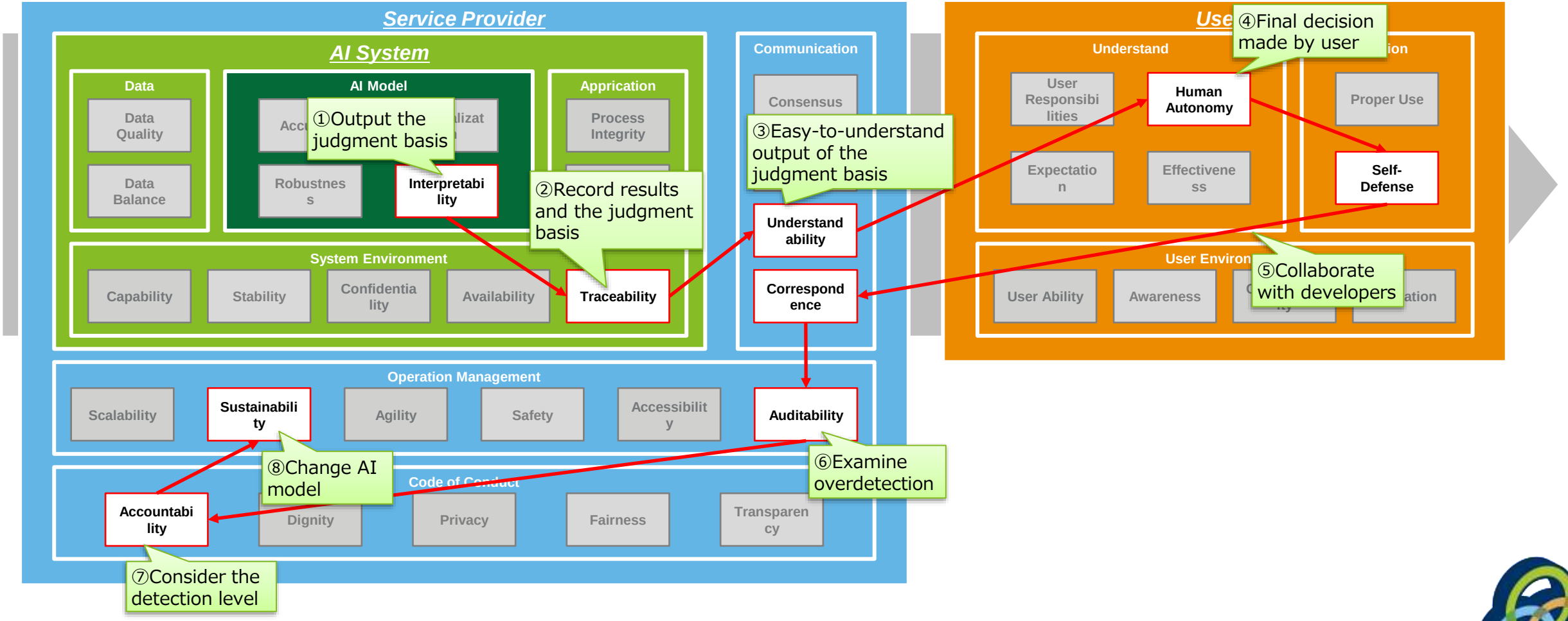
Control Coordination

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

Step3

R008

Excessive judgment
AI outputs results that are too negative for everyone, overburdening users



Risk Control

- Consider risk control according to the risk chain -

Step4

R008	**Excessive judgment** AI outputs results that are too negative for everyone, overburdening users

Risk Control		
AI system (Judicial Digital Service Dept, Co. C)	**Service Provider (Judicial Administration Dept of the Supreme Court)**	**User (Public Prosecutor/Police/Probation Officer)**
① [Interpretability] Outputs Information on the grounds of judgment (Judicial Digital Service Dept, Co. C) ② [Traceability] Record execution results and Information on the grounds of judgment (Judicial Digital Service Dept, Co. C)	③ [Understanding] To output Information on the grounds of judgment in an easy-to-understand manner (Judicial Administration Dept) ⑥ [Correspondence] Engage with developers when they believe overly negative predictions are being made (Judicial Administration Dept) ⑦ [Auditability] Examination of the status of overdetection (Judicial Administration Dept) ⑧ [Accountability] Investigate appropriate detection levels (Judicial Administration Dept) ⑨ [Sustainability] Changing the AI model (Judicial Administration Dept + Judicial Digital Service Dept, Co. C)	④ [Human Autonomy] The final decision is made by the user (User) ⑤ [Self-Defense] Engage developers when they think they are making overly negative predictions (User)

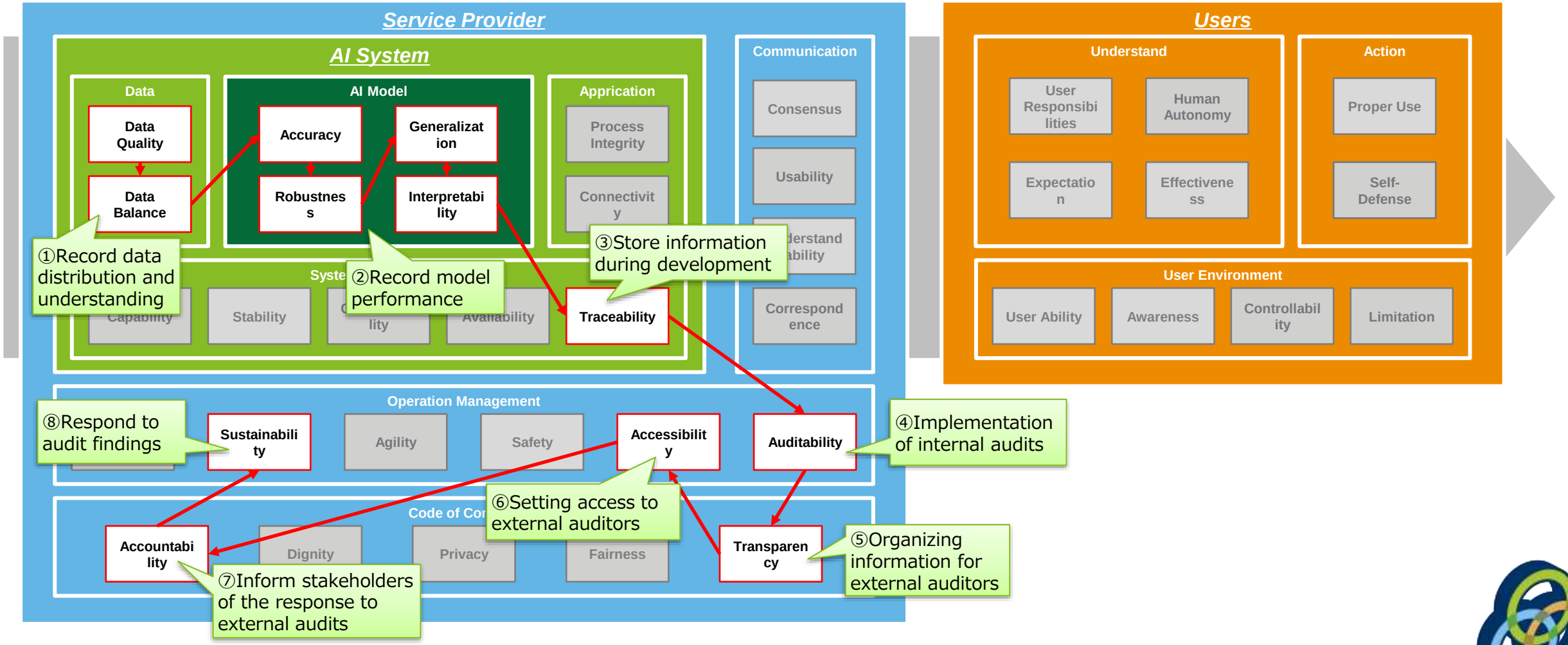
Control Coordination

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

Step3

R010

External explanation
Inability to adequately explain the decision process when requested to do so from outside



Risk Control

- Consider risk control according to the risk chain -

Step4

R010	**External explanation** Inability to adequately explain the decision process when requested to do so from outside

Risk Control		
AI system (Judicial Digital Service Dept, Co. C)	**Service Provider (Judicial Administration Dept of the Supreme Court)**	**User (Public Prosecutor/Police/Probation Officer)**
① [Data Quality] [Data Balance] Record data distribution and understanding (Judicial Digital Service Dept, Co. C) ② [Accuracy] [Robustness] [Generalization] [Interpretability] Record the performance of the model (Judicial Digital Service Dept, Co. C) ③ [Traceability] Record the results of AI decisions (Judicial Digital Service Dept, Co. C)	④ [Auditability] Conducts internal audits and responds in advance (Judicial Administration Dept) ⑤ [Transparency] Organize information to be disclosed to external auditors (Judicial Administration Dept) ⑥ [Accessibility] Establish necessary access rights for external auditors (Judicial Administration Dept) ⑦ [Accountability] Disseminate appropriate response to external audits (Judicial Administration Dept) ⑧ [Sustainability] Responding to issues discovered during the audit (Judicial Administration Dept)	