RCM
RiskChainModel

リスクチェーンモデル(RCModel)ケース検討事例：
Case06 再犯可能性の検証AI

東京大学未来ビジョン研究センター
技術ガバナンス研究ユニット　AIガバナンスプロジェクト

ケース検討のステップ

リスク アセスメント	Step1	**AIサービスが「実現すべき価値・目的」の定義** AIサービスの特徴を踏まえて「実現すべき価値・目的」を検討(ビジネス目的及び社会的責任)
	Step2	**重要なリスクシナリオの検討** 「実現すべき価値・目的」を阻害する「リスクシナリオ」を検討
コントロール コーディネーション	Step3	**重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討** 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化
	Step4	**リスクチェーンに従ってリスクコントロールを検討** リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討
	Step5	**各リスクチェーンでの検討を元にステークホルダー毎の役割を整理** 各リスクチェーンで検討したコントロールを集約してステークホルダー毎の役割を整理し合意形成

ケース事例（AIサービスとリスクコーディネーション研究会）

東京大学未来ビジョン研究センター　技術ガバナンス研究ユニット

AIガバナンスプロジェクト　AIサービスとリスクコーディネーション研究会

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

RCM
RiskChainModel

ケーススタディの概要

ケーススタディの概要（Case06：再犯可能性の検証AI）

- AIサービスが「実現すべき価値・目的」の定義 -

Step1

各裁判所において、被告人が釈放後に再犯を起こしてしまう可能性を予測する機械学習のAIモデルである。

被告人への質問と犯罪履歴を元に、AIで再犯可能性（10段階）で予測し、検察・裁判員が当該事案の判決内容や仮釈放の判断等に用いる。また警察での犯罪捜査や保護観察官による仮釈放人の観察にも用いられている。

【実現すべき価値・目的】

- **再犯の減少**
- **各利用者での適正利用**
- **社会的責任**

【AIサービスを用いた実運用の流れ】

① 各裁判所において、被告人に対してシステムで設定された質問を行う。（質問には、犯罪、保釈の履歴や年齢、雇用状況、暮らしぶり、教育レベル、地域とのつながり、薬物使用、信条、家族の犯罪歴、薬物使用歴等が含まれる）

② 被告人の過去の犯罪履歴データと①の結果を、AIにインプットし、対象者の再犯可能性を10段階の点数として割り出す。再犯可能性と併せて、判断根拠に係る情報（アンケート回答の注目部分、類似した過去の判断事例等）を出力する。

③ 地方裁判所の裁判員はAIによる再犯可能性結果を参考意見の一つとして、被告人へ刑期等の判決を下す。

※ 警察での犯罪捜査や保護観察官でも閲覧は可能である。

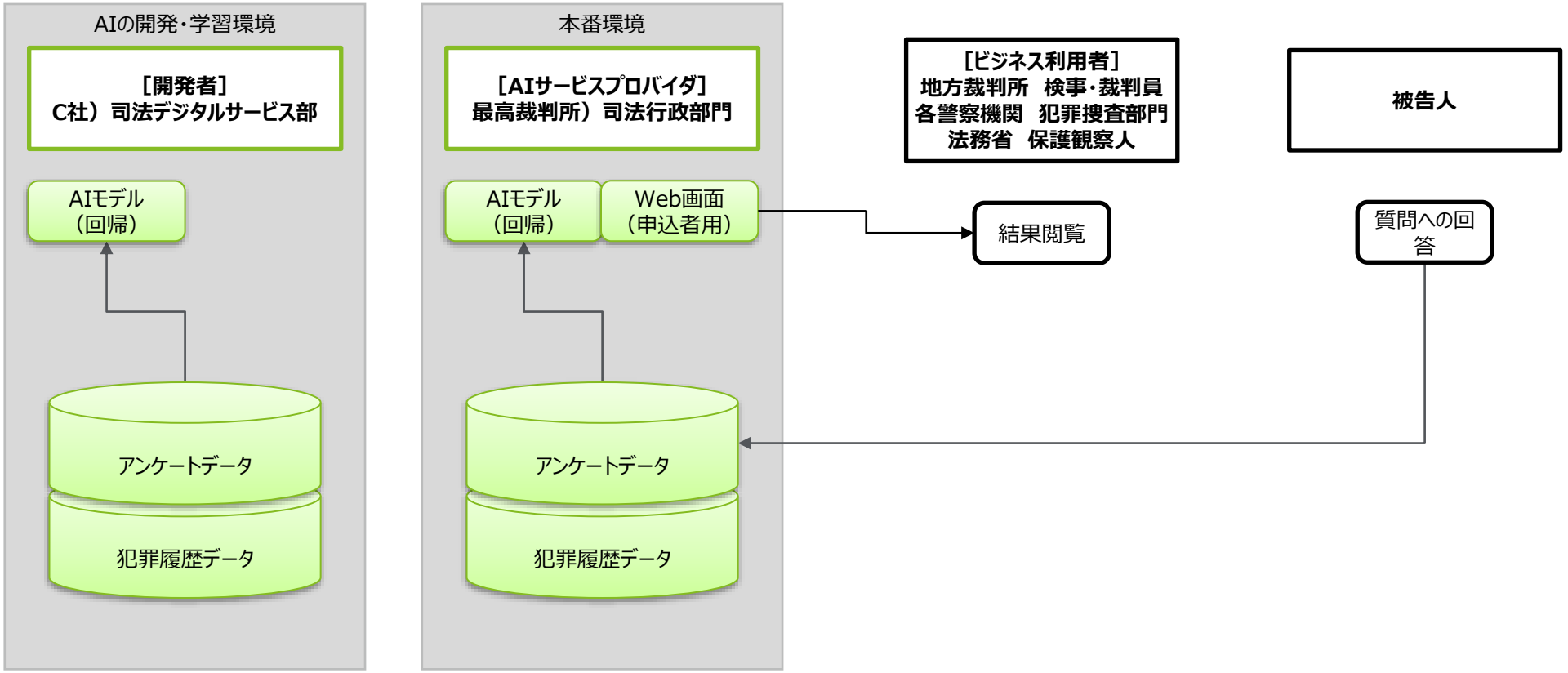
開発を請け負ったC社は、法務省が保有する刑事・民事両方の犯罪履歴データベースを用いて、各地方裁判所共通で用いるAIモデルを作成する。AIモデルの学習はC社の開発環境にて実施され、定期的に犯罪履歴データベース等の情報を学習データとして追加してAIモデルを再学習する。

AIサービスプロバイダは最高裁判所の司法行政部門であり、AIモデルのアップデータは同部門が検討し、必要に応じてC社に指示を行う。実際に用いるAIモデルはWeb上で各利用者（検察・裁判所・警察・保護観察人）のPCから接続し閲覧できる。

ケーススタディの概要（Case06：再犯可能性の検証AI）

- システムの全体概要 -

AIシステム	C社）司法デジタルサービス部	AIモデルの開発・学習を行う
AIサービスプロバイダ	最高裁判所）司法行政部門	AIモデルをWebから必要な利用者に向けて配信する
ユーザー	地方裁判所）検事・裁判員 各警察機関）犯罪捜査部門 法務省）保護観察人	AIモデルの判断結果をWebから閲覧して情報を得る ※主となる利用目的は各裁判所での判決の検討である



ケーススタディの概要（Case06：再犯可能性の検証AI）

- AIサービスの入出力 -

【AIサービスに使用するデータ】

データ	本番/学習	収集方法	データ管理者（管理場所）	個人情報の有無
質問の回答内容	学習	被告人への質問（必須ではない）	C社サーバー	有（要配慮個人情報含む）
過去の犯罪履歴	学習	過去の犯罪履歴データ	法務省データベース	有（要配慮個人情報含む）

【AIサービスからの出力内容】

AIサービス利用者	検察・裁判所・警察・保護観察人
出力結果の内容	10段階評価による再犯予測
出力方法（画面／アクチュエータ等）	Web上で被告人や犯罪捜査対象者の情報を入力すと、再犯可能性のスコアを閲覧できる
期待精度（正解率／誤差等）	75% ※危険度が高いと評価した人物が実際に再犯をした割合
利用者判断の有無	有（最終判断は各利用者の責任において実施される）
根拠情報の出力	アンケート回答の注目部分、類似した過去の判断事例
安全性のリスク有無	有（最終判断は各利用者の責任であるが、誤判断によって再犯増加等のリスクは発生し得る）
外部AIへの連携	無
外部AIへの連携方法・プロトコル	無

RCM
RiskChainModel

リスクアセスメント＆リスクコントロール
- リスクシナリオの検討・評価 → コントロールの整備 -

重要なリスクシナリオの検討

- 「実現すべき価値・目的」を阻害する「リスクシナリオ」を検討 -

Step2

実現すべき価値・目的		サービス要件と関連テクノロジー			リスクNo.	リスクシナリオ	
1	再犯の減少	1-1	予測性能の確保	■ AIの予測精度	R001	誤った判断	学習によりAIの予測性能が劣化してしまい、誤った判断が行われてしまう
		1-2	継続的な指導	■ AIの予測結果の一貫性	R002	不安定な判断	学習の都度、AIの判断が大きく変更されることで一貫性のない対応が行われてしまう
		1-3	適切な判断	■ アンケート項目 ■ 情報管理	R003	説明可能性	最終判断の根拠について、正しい説明ができない
2	各利用者での適正利用	2-1	利用範囲の明確化	■ 利用範囲の定義 ■ アクセス管理	R004	目的外利用	想定していない目的にAIモデルが利用されることで特定の個人に不利益を与える
		2-2	利用者での誤判断の防止	■ AIの説明可能性 ■ AIの公平性	R005	過度なAI依存	AIの判断に依存することで誤った最終判断が行われる
					R006	不適切な入力による誤判断	被告人が再犯可能性を下げるために、意図的にアンケートへ事実と異なる情報をインプットさせる
					R007	公平性の棄損	特定の人種／性別に対して、明らかに不公平な予測結果を生じさせる
3	業務効率化・負担軽減	3-1	適度な予測レベル	■ 出力制御	R008	過剰な判断	AIが誰に対しても過度にネガティブな結果を出力することで、ユーザーの負担が過大になってしまう
		3-2	負担の少ないデータ収集	■ 効率的なデータ収集	R009	データ収集の負担の増加	人手で収集する学習データが膨大になり、ユーザーの負荷が過剰になってしまう
4	社会的責任	4-1	アカウンタビリティ	■ 説明可能性 ■ 検証可能性	R010	外部に向けた説明	判断プロセスについて、外部から説明が求められた際に十分な説明を行うことができない
		4-2	情報管理	■ セキュリティ確保 ■ データ管理	R011	AIシステムのハッキング	学習データや予測モデルが不適切に更新され、特定の人物やグループに対して誤った判断が行われる
					R012	風評被害の発生	AIモデルの判断傾向について不適切な解釈等が公開され、特定の人物やグループの名誉を棄損する
					R013	プライバシー保護	個人情報の取扱を誤ることによって個人情報保護法の違反が行われる

重要なリスクシナリオに対するコントロールのサマリー

- 各リスクチェーンの検討結果を集約 -

Step5

実現すべき価値・目的		リスクNo.	リスクシナリオ	不確実性	環境変化	利用者起因	RC	コントロールのサマリー		
								AIシステム	サービスプロバイダ	ユーザー
1	再犯の減少	R001	誤った判断	○			●	予測精度の確保 実行結果の記録	期待精度の説明 再学習	期待精度の把握 必要な知識の確保 最終判断
		R002	不安定な判断	○	○		●	データ分布変化の把握 判断根拠の出力	判断根拠の変更の把握 再学習	判断悔過の変更の把握
		R003	説明可能性	○		○	●	判断根拠の出力	判断根拠の分かりやすさ	必要な知識の確保 最終判断
2	各利用者での適正利用	R004	目的外利用			○		データ保護	アクセス管理 目的内利用	データの取扱
		R005	過度なAI依存	○		○	●	※R003と同じ	※R003と同じ	※R003と同じ
		R006	不適切な入力による誤判断	○		○	●	外部データの自動連携 学習データの検証	使いやすいUI 判断精度・異常値の検証 再学習	正確なフィードバック
		R007	公平性の棄損	○			●	データの偏り 特徴量の公平性	公平性の留意点の整理 判断根拠の可視化	公平性リスクの理解 観察計画の検討
3	業務効率化・負担軽減	R008	過剰な判断	○		○	●	判断根拠の出力 実行結果の記録	検出レベルの検討 再学習	最終判断
		R009	データ収集の負担の増加			○		データ収集方法の簡易化	費用対効果の検討	負担のフィードバック
4	社会的責任	R010	外部に向けた説明	○			●	データ理解の記録 モデル性能の記録 実行結果の記録	開示する情報の整理 アクセス権管理 監査対応	
		R011	AIシステムのハッキング					セキュリティ管理	原因調査・改善	ユーザー側環境の保護
		R012	風評被害の発生					データ保護	職業倫理の教育	職業倫理の教育
		R013	プライバシー保護					データ保護	法令順守の教育	法令順守の教育 データの取扱

ステークホルダー毎の役割を整理

- 各コントロールをステークホルダー別に整理 -

Step5

- 責任者 -
最高裁判所長官

- 実現すべき価値・目的の検討
- リスクコントロール方法の承認
- 対外的な説明

- AIサービスプロバイダ -
最高裁判所）司法行政部門

- 期待精度の説明
- 判断根拠の変更の把握
- 公平性の留意点の整理
- 判断根拠の分かりやすさ
- 検出レベルの検討
- 再学習
- 費用対効果の検討
- 目的内利用
- 開示する情報の整理
- 性能監視
- アクセス権管理
- 監査対応
- 障害の原因調査・改善
- 職業倫理の教育
- 法令順守の教育

C社）司法デジタルサービス部

- 予測精度の確保
- 判断根拠の出力
- 特徴量の公平性
- データ分布変化の把握
- データの偏り
- データ理解の記録
- データ収集方法の簡易化
- 使いやすいUI

C社）ITサービス部

- 実行結果の記録
- モデル性能の記録
- データ保護
- セキュリティ管理
- パフォーマンス保守
- ユーザー側環境の保護

- ユーザー -
地方裁判所）検事・裁判員
各警察機関）犯罪捜査部門
法務省）保護観察人

- 最終判断の実施
- 予測結果の変化の把握
- 過検出時に観察計画を見直し
- 期待精度の把握
- 公平性リスクの理解
- 必要なリテラシーの確保
- 正確なフィードバック
- データの取扱
- 代替運用

被告人/観察対象者等

RCM
RiskChainModel

リスクコントロールの検討
- リスクチェーンを用いたコントロール検討の詳細 -

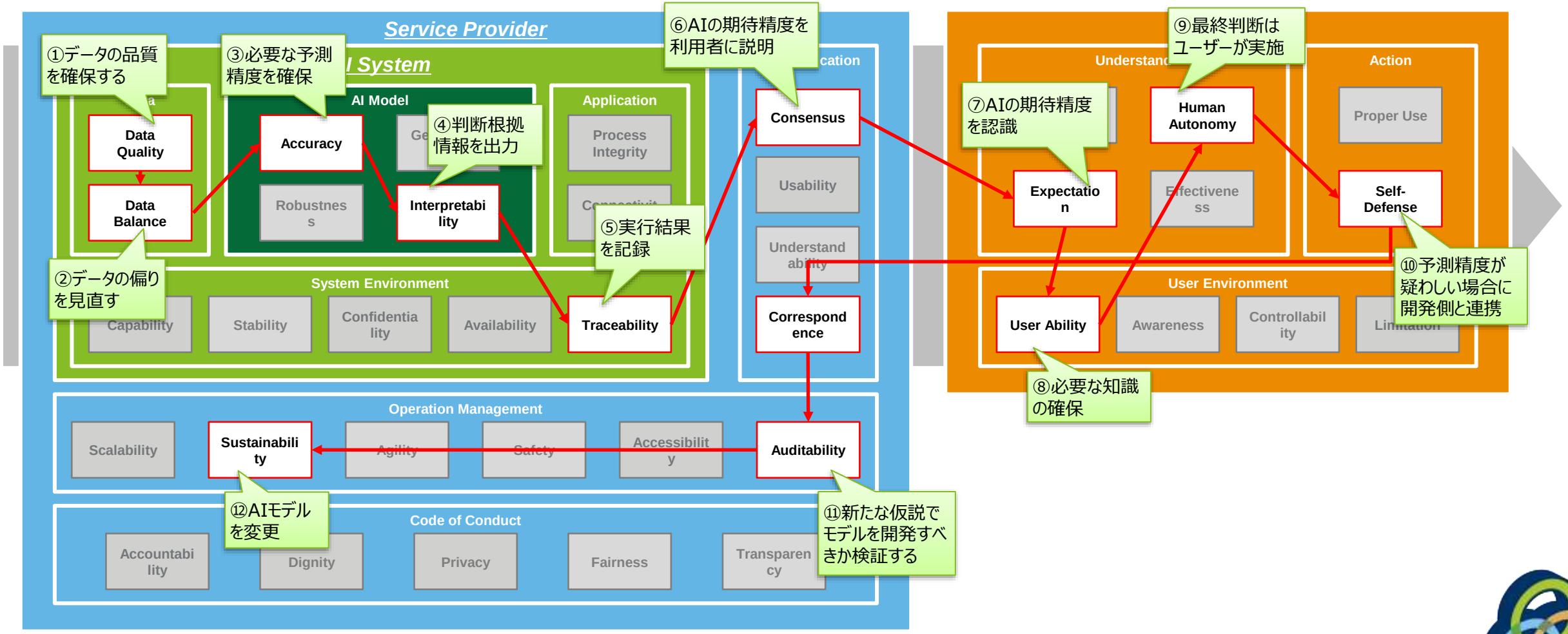
重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討

- 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化 -

Step3

R001

誤った判断
学習によりAIの予測性能が劣化してしまい、誤った判断が行われてしまう



リスクチェーンに従ってリスクコントロールを検討

- リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討 -

Step4

R001	**誤った判断** 学習によりAIの予測性能が劣化してしまい、誤った判断が行われてしまう

コントロールの内容		
AIシステム （C社司法デジタルサービス部）	**サービスプロバイダ （最高裁判所 司法行政部門）**	**ユーザー （検事・裁判員／警察／保護観察人）**
①【Data Quality】誤って収集されたデータを修正し、品質を確保する（C社司法デジタルサービス部） ②【Data Balance】データ偏りを整える（C社司法デジタルサービス部） ③【Accuracy】必要な予測精度を確保する（C社司法デジタルサービス部） ④【Interpretability】判断根拠情報を出力する（C社司法デジタルサービス部） ⑤【Traceability】実行結果を記録する（C社司法デジタルサービス部）	⑥【Consensus】AIの期待精度を利用者に説明する（司法行政部門） ⑩【Correspondence】予測精度が疑わしい場合に開発側と連携する（司法行政部門） ⑪【Auditability】新たな仮説でモデルを作成すべきか検討する（司法行政部門） ⑫【Sustainability】AIモデルを変更する（司法行政部門＋C社司法デジタルサービス部）	⑦【Expectation】AIの期待精度を認識する（ユーザー） ⑧【User Ability】勉強会・事例共有を実施し、最終判断に必要な知識・リテラシーを確保する（ユーザー） ⑨【Human Autonomy】最終判断はユーザーが実施する（ユーザー） ⑩【Self-Defense】予測精度が疑わしい場合に開発側と連携する（ユーザー）

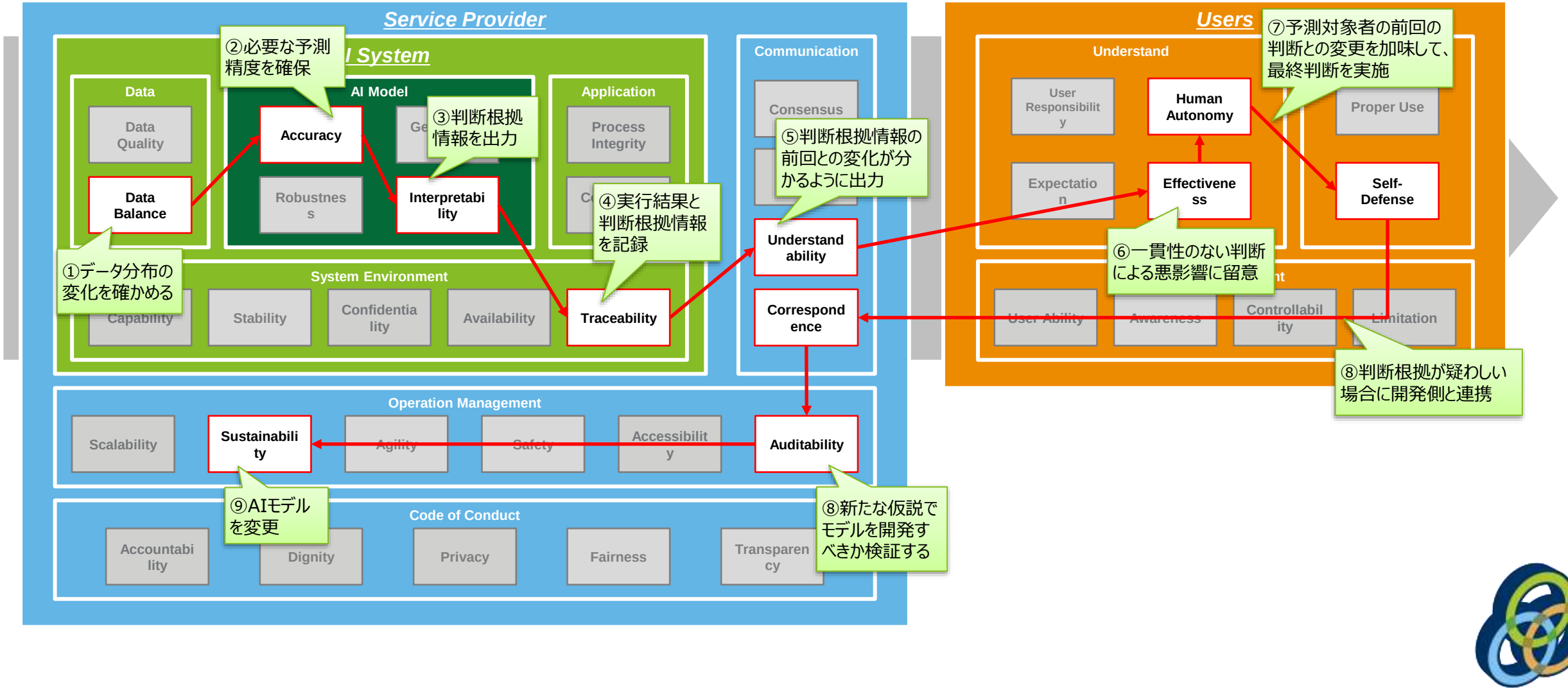
重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討

- 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化 -

Step3

R002

不安定な判断
学習の都度、AIの判断が大きく変更されることで一貫性のない対応が行われてしまう



リスクチェーンに従ってリスクコントロールを検討

- リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討 -

Step4

R002	**不安定な判断** 学習の都度、AIの判断が大きく変更されることで一貫性のない対応が行われてしまう

コントロールの内容		
AIシステム （C社司法デジタルサービス部）	**サービスプロバイダ （最高裁判所 司法行政部門）**	**ユーザー （検事・裁判員／警察／保護観察人）**
①【Data Balance】データ分布の変化を確かめる（C社司法デジタルサービス部） ②【Accuracy】必要な予測精度を確保する（C社司法デジタルサービス部） ③【Interpretability】判断根拠情報を出力する（C社司法デジタルサービス部） ④【Traceability】実行結果と判断根拠情報を記録する（C社司法デジタルサービス部）	⑤【Understandability】判断根拠情報の前回との変化が分かるように出力する（司法行政部門） ⑧【Correspondence】判断根拠情報が疑わしい場合に開発側と連携する（司法行政部門） ⑨【Auditability】新たな仮説でモデルを作成すべきか検討する（司法行政部門） ⑩【Sustainability】AIモデルを変更する（司法行政部門＋C社司法デジタルサービス部）	⑥【Effectiveness】一貫性のない判断による悪影響を考える（ユーザー） ⑦【Human Autonomy】予測対象者に対する前回との変更点を踏まえて、最終判断を実施する（ユーザー） ⑧【Self-Defense】判断根拠情報が疑わしい場合に開発側と連携する（ユーザー）

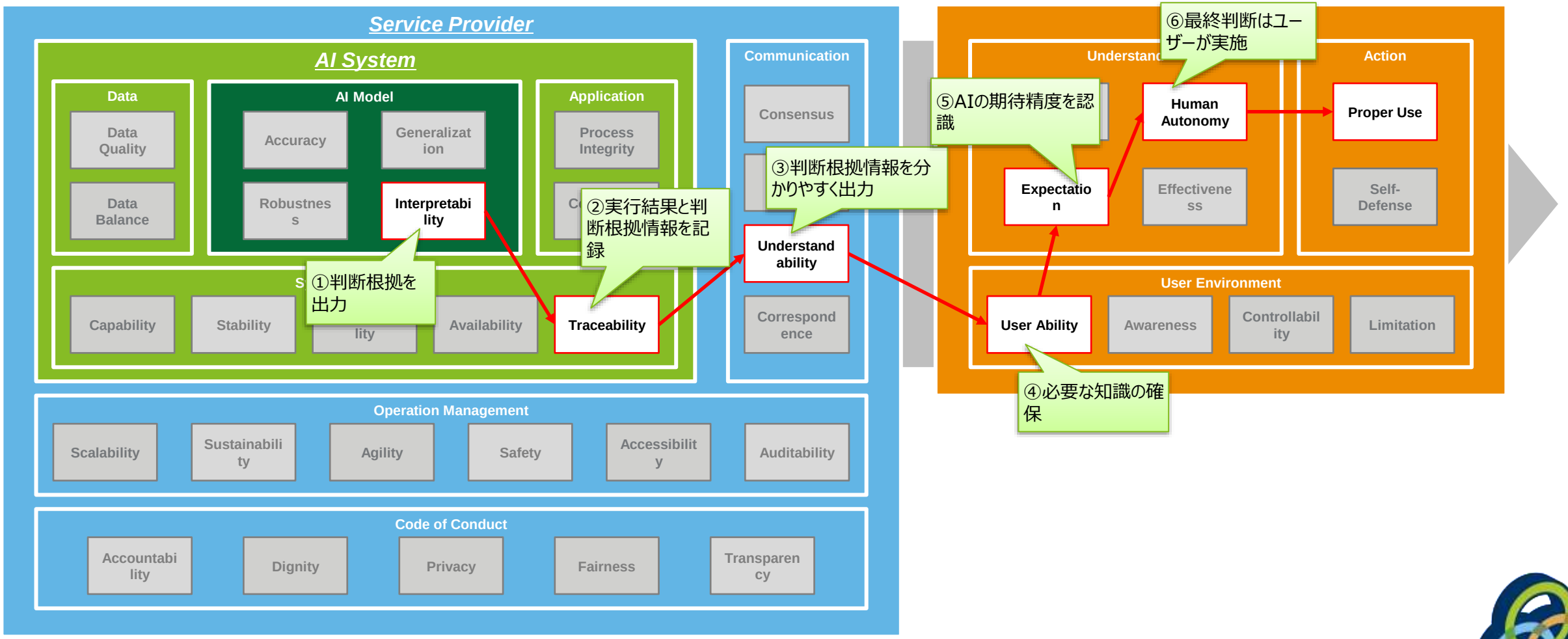
重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討

- 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化 -

Step3

R003

説明可能性
最終判断の根拠について、正しい説明ができない



リスクチェーンに従ってリスクコントロールを検討

- リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討 -

Step4

R003	**説明可能性** 最終判断の根拠について、正しい説明ができない

コントロールの内容		
AIシステム （C社司法デジタルサービス部）	**サービスプロバイダ （最高裁判所 司法行政部門）**	**ユーザー （検事・裁判員／警察／保護観察人）**
①【Interpretability】判断根拠情報を出力する（C社司法デジタルサービス部） ②【Traceability】実行結果と判断根拠情報を記録する（C社司法デジタルサービス部）	③【Understandability】判断根拠情報を分かりやすく出力する（司法行政部門）	④【User Ability】勉強会・事例共有を実施し、最終判断に必要な知識・リテラシーを確保する（ユーザー） ⑤【Expectation】AIの期待精度を認識する（ユーザー） ⑥【Human Autonomy】最終判断はユーザーが実施する（ユーザー） ⑦【Proper Use】最終判断を適切に実施する（ユーザー）

重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討

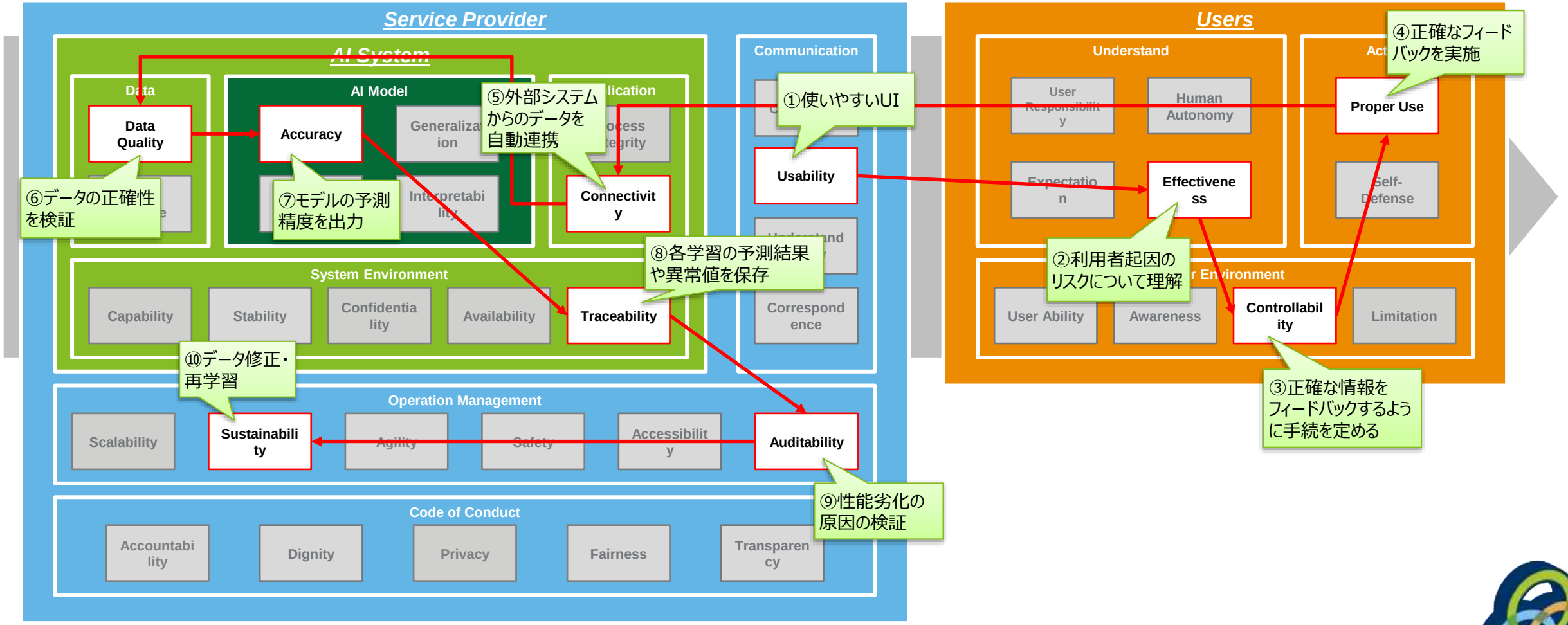
- 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化 -

Step3

R006

不適切な入力による誤判断

被告人が再犯可能性を下げるために、意図的にアンケートへ事実と異なる情報をインプットさせる



リスクチェーンに従ってリスクコントロールを検討

- リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討 -

Step4

R006

不適切な入力による誤判断
被告人が再犯可能性を下げるために、意図的にアンケートへ事実と異なる情報をインプットさせる

コントロールの内容		
AIシステム （C社司法デジタルサービス部）	**サービスプロバイダ （最高裁判所 司法行政部門）**	**ユーザー （検事・裁判員／警察／保護観察人）**
⑤【Connectivity】可能なデータを外部システムから自動連携（C社司法デジタルサービス部） ⑥【Data Quality】学習データの正確性を検証する（C社司法デジタルサービス部） ⑦【Accuracy】モデルの判断精度を確保する（C社司法デジタルサービス部） ⑧【Traceability】各学習段階での予測結果や異常値の内容を保存する（C社司法デジタルサービス部）	①【Usability】誤操作が起こりにくい使いやすいUIを準備する(司法行政部門＋C社司法デジタルサービス部) ⑨【Auditability】予測性能の変化や学習データにおける異常値(ラベル誤りと思われるデータ)を検証する(司法行政部門＋C社司法デジタルサービス部) ⑩【Sustainability】データの教師ラベルを修正し、必要に応じてAIモデルを再学習する(司法行政部門＋C社司法デジタルサービス部)	②【Effectiveness】フィードバック誤りがAIモデルの性能劣化に影響することを認識する（ユーザー） ③【Controllability】正確なフィードバックを行うように手続きを整備する（ユーザー） ④【Proper Use】正確にフィードバックを実施する（ユーザー）

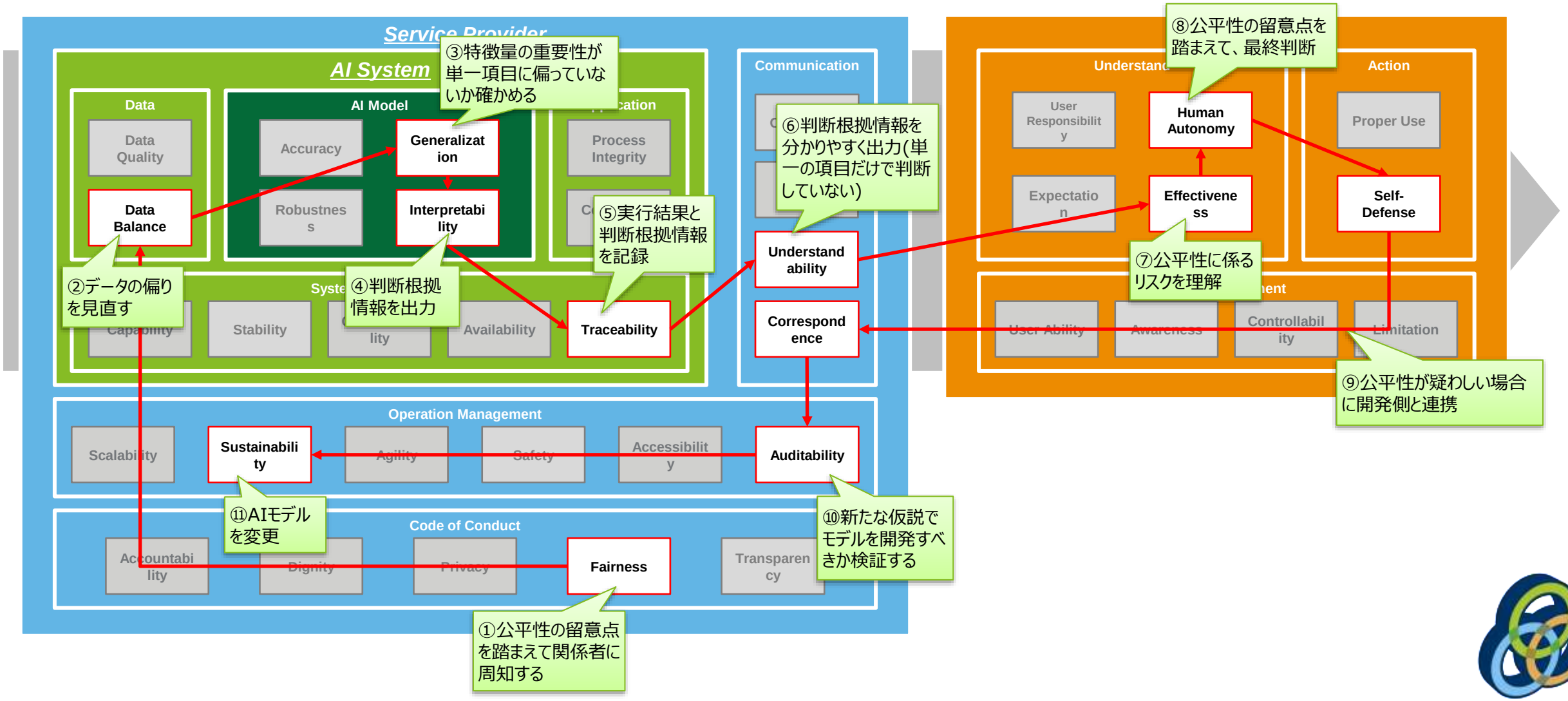
重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討

- 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化 -

Step3

R007

公平性の棄損
特定の人種／性別に対して、明らかに不公平な予測結果を生じさせる



リスクチェーンに従ってリスクコントロールを検討

- リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討 -

Step4

R007	**公平性の棄損** 特定の人種／性別に対して、明らかに不公平な予測結果を生じさせる

コントロールの内容		
AIシステム （C社司法デジタルサービス部）	**サービスプロバイダ （最高裁判所 司法行政部門）**	**ユーザー （検事・裁判員／警察／保護観察人）**
②【Data Balance】データの偏りを見直す（C社司法デジタルサービス部） ③【Generalization】特徴量の重要性が単一項目に偏っていないか確かめる（C社司法デジタルサービス部） ④【Interpretability】判断根拠情報を出力する（C社司法デジタルサービス部） ⑤【Traceability】実行結果と判断根拠情報を記録する（C社司法デジタルサービス部）	①【Fairness】公平性の留意点を踏まえ、モデルの判断傾向に問題がないことを確かめる（司法行政部門） ⑥【Understandability】判断根拠情報を分かりやすく出力する（司法行政部門） ⑨【Correspondence】公平性が疑わしい場合に開発側と連携する（司法行政部門） ⑩【Auditability】新たな仮説でモデルを作成すべきか検討する（司法行政部門） ⑪【Sustainability】AIモデルを変更する（司法行政部門＋C社司法デジタルサービス部）	⑦【Effectiveness】公平性に係るリスクを理解する（ユーザー） ⑧【Human Autonomy】判断根拠の公平性を踏まえて、最終判断を行う（ユーザー） ⑨【Self-Defense】公平性が疑わしい場合に開発側と連携する（ユーザー）

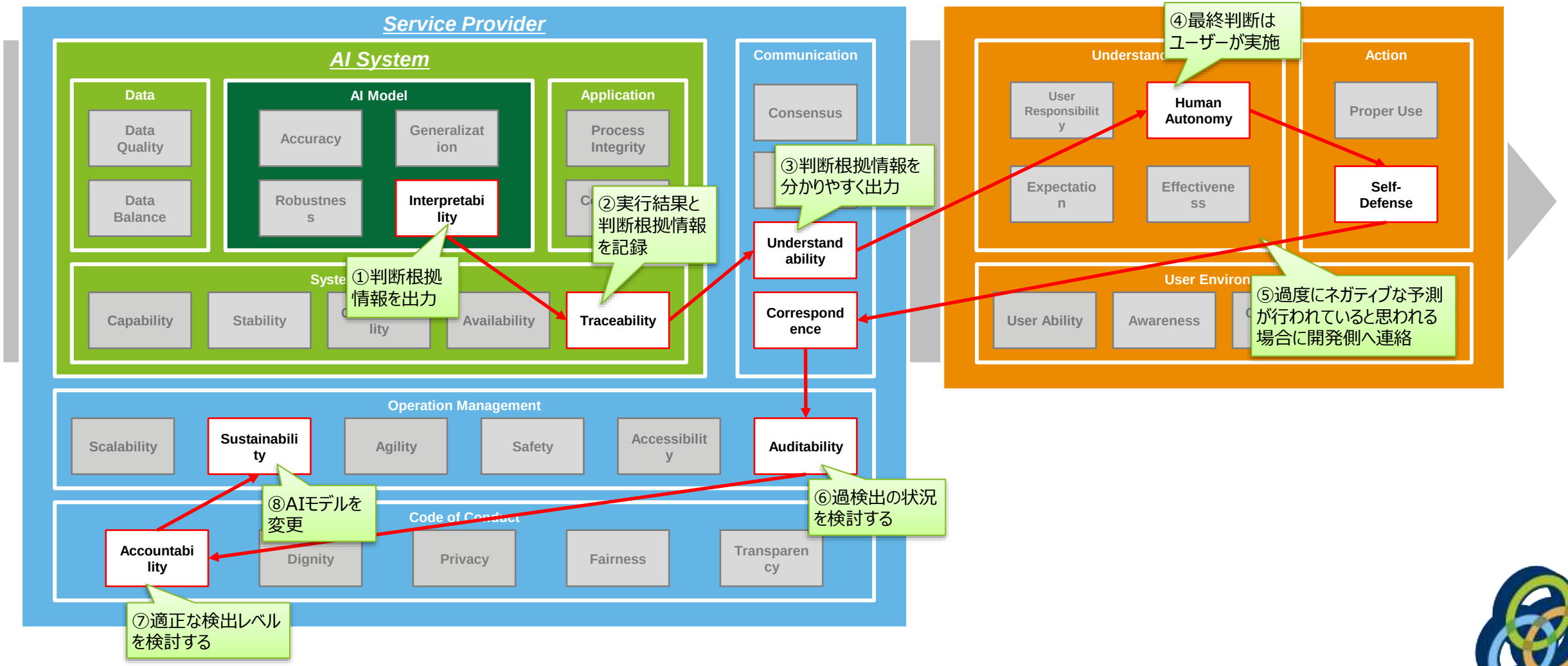
重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討

- 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化 -

Step3

R008

過剰な判断
AIが誰に対しても過度にネガティブな結果を出力することで、ユーザーの負担が過大になってしまう



リスクチェーンに従ってリスクコントロールを検討

- リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討 -

Step4

R008	**過剰な判断** AIが誰に対しても過度にネガティブな結果を出力することで、ユーザーの負担が過大になってしまう

コントロールの内容		
AIシステム （C社司法デジタルサービス部）	**サービスプロバイダ （最高裁判所 司法行政部門）**	**ユーザー （検事・裁判員／警察／保護観察人）**
①【Interpretability】判断根拠情報を出力する（C社司法デジタルサービス部） ②【Traceability】実行結果と判断根拠情報を記録する（C社司法デジタルサービス部）	③【Understandability】判断根拠情報を分かりやすく出力する（司法行政部門） ⑥【Correspondence】過度にネガティブな予測が行われていると思われる場合に開発側に連携する（司法行政部門） ⑦【Auditability】過検出の状況を検討する（司法行政部門） ⑧【Accountability】適正な検出レベルを検討する（司法行政部門） ⑨【Sustainability】AIモデルを変更する（司法行政部門＋C社司法デジタルサービス部）	④【Human Autonomy】最終判断はユーザーが実施する（ユーザー） ⑤【Self-Defense】過度にネガティブな予測が行われていると思われる場合に開発側に連携する（ユーザー）

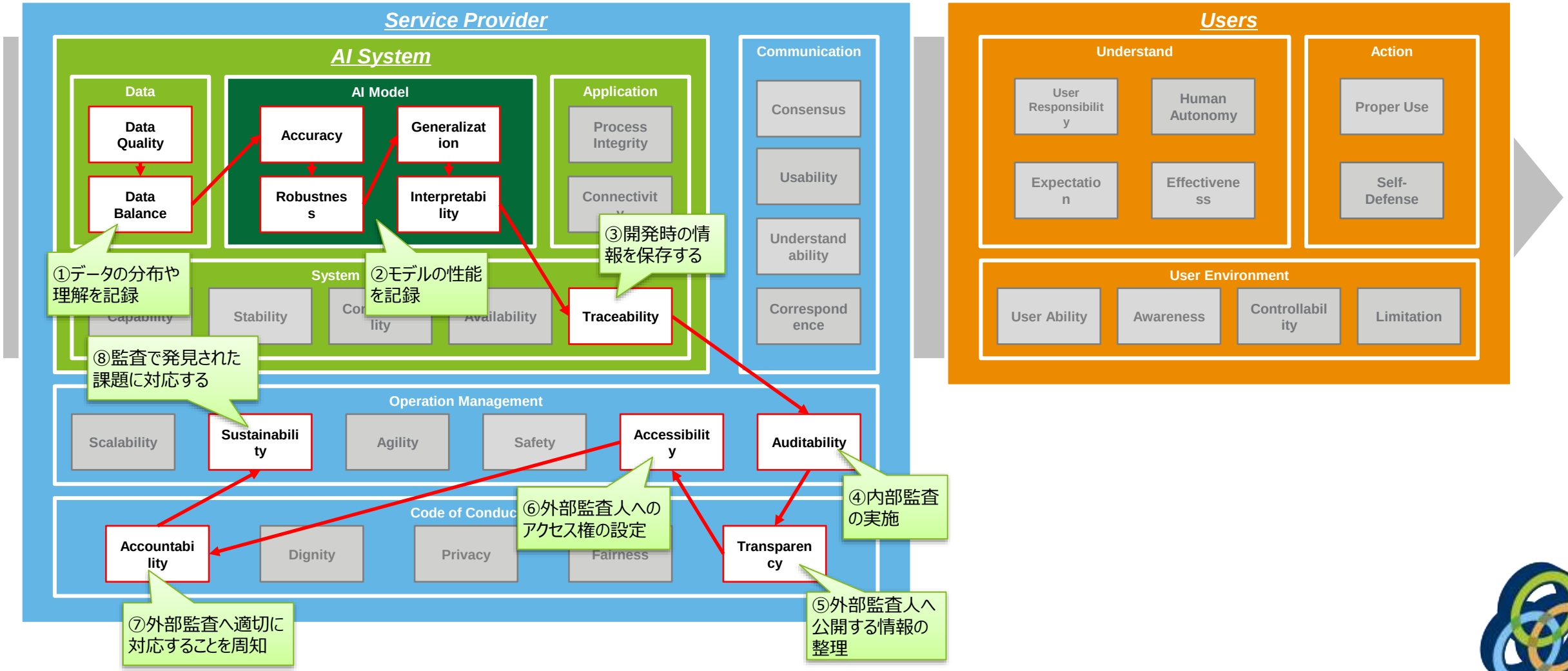
重要なリスクシナリオごとにリスクチェーン(リスク要因の関係性)の検討

- 「リスクシナリオ」ごとにRCModelの各層からリスク要因と関係性(リスクチェーン)を可視化 -

Step3

R010

外部に向けた説明
AIサービス自体の信頼性を対外的に説明できない



リスクチェーンに従ってリスクコントロールを検討

- リスクチェーンで関連づけられた構成要素においてリスク対応策(コントロール)を検討 -

Step4

R010	**外部に向けた説明** AIサービス自体の信頼性を対外的に説明できない

コントロールの内容		
AIシステム （C社司法デジタルサービス部）	**サービスプロバイダ （最高裁判所 司法行政部門）**	**ユーザー （検事・裁判員／警察／保護観察人）**
①【Data Quality】【Data Balance】データの分布や理解を記録（C社司法デジタルサービス部） ②【Accuracy】【Robustness】【Generalization】【Interpretability】モデルの性能を記録（C社司法デジタルサービス部） ③【Traceability】AIの判断結果を記録する（C社司法デジタルサービス部）	④【Auditability】内部監査を実施し事前に対応を行う（司法行政部門） ⑤【Transparency】外部監査人へ公開する情報を整理する（司法行政部門） ⑥【Accessibility】外部監査人へ必要なアクセス権を設定する（司法行政部門） ⑦【Accountability】外部監査へ適切に対応することを周知（司法行政部門） ⑧【Sustainability】監査で発見された課題に対応する（司法行政部門）	