RCM
RiskChainModel

2021年6月

リスクチェーンモデル（RCModel）ガイド Ver 1.0

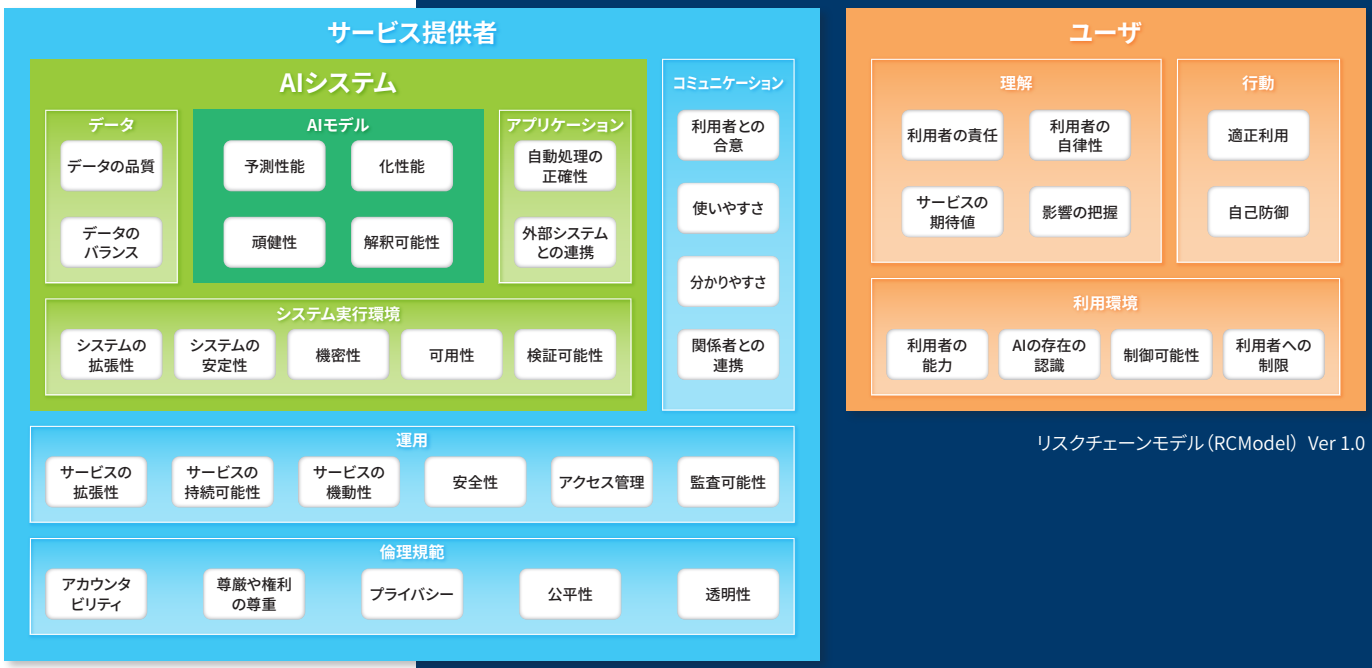
東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット
AIガバナンスプロジェクト

リスクチェーンモデル(RCModel)概要

AIを使ったシステムやサービスは、信頼性や透明性の確保が重要とされていますが、具体的にどのようにすればよいのでしょうか。

世界各国では様々な原則やガイドライン作り、チェックリストなどが作成されています。その原則を実践に落とし込んでいく方法として、東京大学の研究グループが開発したリスクチェーンモデル(RCModel)を紹介します。

RCM
RiskChainModel



リスクチェーンモデル（RCModel）Ver 1.0

RCModelの目的 ①

～AIサービスの「重要なリスク」は何ですか？～

AIサービスの開発・利活用するにあたって重要なリスクは何でしょうか？それはAIサービス自体によって異なります。「採用AI」や「ローン審査AI」のように個人を評価するようなAIであれば「公平性」が重要になりますし、「インフラの外観検査AI」のように画像等の非構造なデータを扱うAIは「頑健性」や「変化への対応」が必要となります。

RCModelは、まず「そのAIサービスにとって、重要なリスクシナリオ」を識別します。それも全AIサービスを統一に評価する何百項目の膨大なチェックリストなるものではなく、「そのAIサービスが重視とする価値・目的」を定義し、その「価値・目的」の実現を阻害する「リスクシナリオ」を検討するアプローチをとっています。

RCModelの目的 ②

～そのリスクはAIモデルだけで十分に対応し続けられますか？～

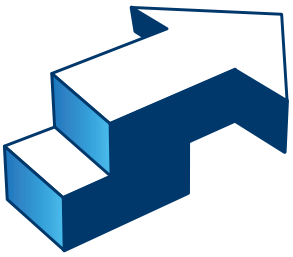
AIモデルは全ての「重要なリスクシナリオ」へ十分に対応し続けることができるのでしょうか？現実的には非常に困難です。AIモデル自体が不確実性を持つため、期待通りの結果を返さない可能性があります。たとえ開発時点では優れた予測性能を持っていたAIモデルであっても、環境の変化によって正しく予測できなくなる可能性があります。また利用者の誤用・悪用等が原因となってAIモデルの予測性能が劣化することも考えられます。

RCModelは、識別した「重要なリスクシナリオ」に対して、AIモデルだけではなく関連するテクノロジー (データやシステム基盤等)、サービス提供者、ユーザに渡ってリスク軽減策(リスクコントロール)を考えることで、AIモデルだけでは十分に対応できないリスクシナリオについても、AIサービス全体で対応できるリスクマネジメント計画を検討します。

RCModelの構造

～AIシステム／サービス提供者／ユーザの三層構造～

RCModelは、
(1) **AIモデル**（精度や頑健性等）及び**AIシステム**（データ、システム基盤、他の制御機能）
(2) **サービス提供者**（行動規範、サービス運用、ユーザとのコミュニケーション）
(3) **ユーザー**（理解、利活用、利用環境）
の3層に構造化しています。



RCModelの方法

それではRCModelの利用方法について説明します。サンプルケース「採用AI」を参考にしながら、各Stepの内容を見ていきましょう。

Step1　AIサービスが「実現すべき価値・目的」の定義

最初に下記のようなAIサービスの特徴を踏まえて「実現すべき価値・目的」を定義します。これは「無人化の実現」「魅力的な提案」「人間よりも高度な検知」等のビジネス目的だけではなく、「不適切な利用の防止」「安全性の確保」等の企業の社会的責任も含まれます。この検討の過程で「価値」同士がトレードオフの関係になることも発生します。その場合には優先すべき価値・目的を検討することが、リスクを考える上で重要になります。

AIサービスの特徴（例）	• AIサービスの利用目的 • システムの概念図 • 使用するアルゴリズムやデータ • 開発者と利用者の役割分担 • 開発方法や学習頻度

■サンプルケース「採用AI」P.2

ケーススタディの概要 (Case01：採用AI)

A社グループのグローバル各社における人材採用部門が、エントリーシートの書類選考を判断する際の参考情報として使用されるAIサービスである。

A社AI開発部門は、ビジネス利用者であるA社人材採用部門（海外グループ会社を含む）より過去のエントリーシートデータと合否判定（内定の判定）結果を受領し、機械学習（分類モデル）で合否を判定する学習モデルを作成している。

【実現すべき価値・目的】

■ 人材採用レベルの維持・向上

■ 採用活動に係るコストの削減

■ 海外グループを含めたサービス提供

■ 企業の社会的責任(公平性のある採用活動)

【期待精度】

適合率（Precision）：書類選考で合格と判断した中で、面接を経て内定を出した割合で評価し、70%を期待値として設定している。

※書類選考で不合格とした候補者については、その後人事部側の判断で書類選考を合格とする人数が少ないため、「再現率（Recall）：書類選考でAIが不合格と判断し、最終的に内定も出さなかった割合」は検討に十分なデータ量が取れないことから、Precisionを評価指標としている。

A社人材採用部門は申込者（新卒・中途両方）のエントリーシート（電子ファイル）を学習モデルに読込み、自身のPC上（ブラウザ上）でAIの判断結果（合否判断）を確かめられる。なお、出力画面には合否判断のみではなく、エントリーシートにおいて判断に影響したキーワードがハイライトされて表示される。人材採用部門担当者はAIの判断を参考情報としながら、合否の判断を設定し、人材採用部門長の承認を得て、申込者に合否の通知を行う。

蓄積された本番データは採用・不採用の判定結果が入力されたタイミングで適宜学習データとして追加され、日次で学習モデルを更新して本番環境に反映する。ただし、学習プロセスでクロスバリデーションによるテストの結果、正解率が70%を下回る場合には本番環境のAIモデルを自動更新しない。AIモデルは過去1年分のバージョンを保存している。

2

Step2 重要なリスクシナリオの検討

実現すべき価値・目的を定義したら、その価値・目的の実現を阻害するリスクシナリオを検討していきます。例えば「無人化の実現」であれば「環境変化による精度劣化」「異常動作」等のサービス維持を阻害するリスクシナリオが考えられます。「魅力的な提案」であれば「戦略商品を誰にでも提案してしまうリスク」等のビジネス戦略への悪影響に係るようなリスクシナリオが考えられます。実現すべき価値・目的への影響を踏まえて、様々な関係者と検討を行い、リスクシナリオごとに優先度を検討します。

■サンプルケース「採用AI」P.6

リスクアセスメント (Case01：採用AI)
- 実現したい価値･目的 → サービス要件 → リスクシナリオを検討 -

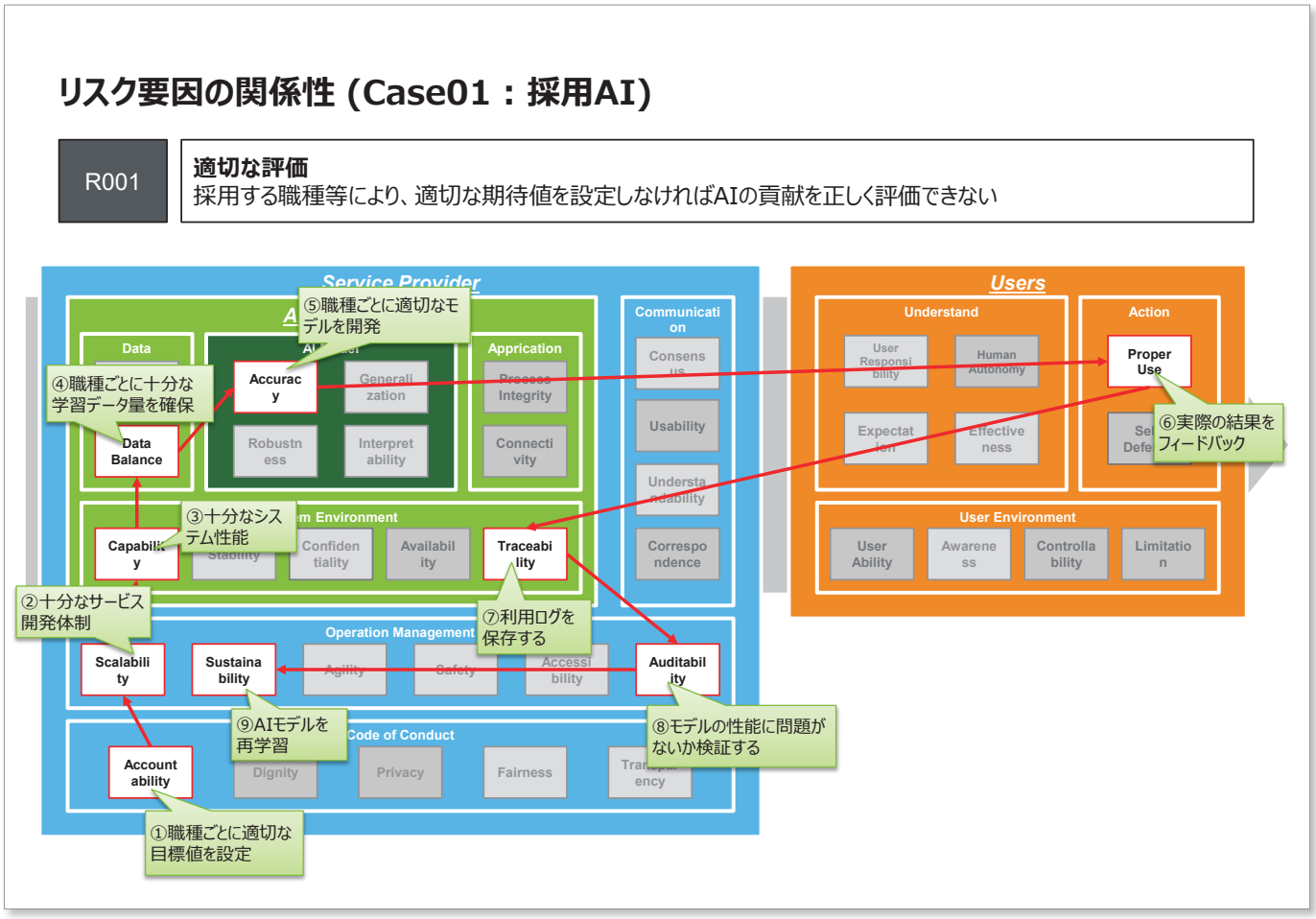
実現すべき価値･目的		サービス要件と関連テクノロジー			リスクNo.	リスクシナリオ	
1	人材採用レベルの維持･向上	1-1	予測性能の維持	■ AIの予測精度 ■ AIの頑健性 ■ AIの説明可能性	R001	適切な評価	採用する職種等により、適切な期待値を設定しなければAIの貢献を正しく評価できない
					R002	予測性能の維持	AIの予測性能が劣化したことで採用レベルが低下する
					R003	ノイズによる影響	エントリーシートで使用される文字情報が若干異なるだけで(句読点の違い等)、AIの判断結果が大きく変化する
					R004	虚偽の申込	申込内容に虚偽が含まれているが、合格と判断してしまう
		1-2	AIサービスと利用者の連携	■ AI依存 ■ AIへのフィードバック	R005	過度なAI依存	人材採用担当者がAIの判断に依存しすぎることで、AIの判断誤りに気が付かない
					R006	誤ったフィードバック	人材採用担当者によるAIへのフィードバック(合否のラベル設定)が不正確なことでAIの性能が劣化する
		1-3	ビジネス環境変化への対応	■ データ分布の変化	R007	人材トレンドの変化	求める人材トレンドの変化にAIの予測傾向が対応できず、採用レベルが低下する
					R008	新たな職種	新たな職種・人材を求める際にAIモデルの予測が妥当でない可能性がある
2	採用活動に係るコストの削減	2-1	適切なサービスコスト	－	R009	コスト超過	サービス維持コストが超過する
3	海外グループを含めたサービス提供	3-1	サービスのローカライズ	■ AIの学習 ■ 個別モデル開発	R010	地域の会社への対応	地域の会社ごとに採用方針や申込者の傾向が異なることで、会社によって採用レベルが確保できなくなる
		3-2	広範囲な開発･サポート体制	■ 開発体制	R011	不十分な開発スピード	充分な開発体制が確保されていないことで、モデルの精度が劣化した際に再学習などの対応が行われない
4	企業の社会的責任(公平性のある採用活動)	4-1	倫理･コンプライアンス遵守	■ AIの判断根拠 ■ AIの汎化性	R012	判断根拠情報の不正販売	AIサービスを何度も利用することで高確率で合格判断を行うキーフレーズ等を特定し、社外へ不正販売する
					R013	公平性	特定の国／地域／人種／性別／年齢に対して不公平な予測結果を生じさせる
		4-2	情報管理	■ データ管理	R014	予測結果の目的外利用	AIの判断結果情報が別の目的に使用されることで、特定の人間に不利益を生じさせる
					R015	風評被害	AIの予測結果情報が外部に流出することで特定の人物に対して風評被害が発生する
					R016	プライバシー保護	AIシステムに蓄積された個人情報が流出する

6

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（AIシステム／サービス提供者／ユーザ）におけるリスク要因を検討し、リスクが顕在化する順番を考えながらリスクチェーン（線）を引いていきます。チェーンは一方向や単線であるとも限らず、どこが起点になるかも扱うリスクや問題によって変化します。

■サンプルケース「採用AI」P.10（リスクシナリオ「R001.適切な評価」の検討例）



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された各要素でリスク低減策(リスクコントロール)を検討します。リスクコントロールは単一では十分にリスクを低減できないかもしれませんが、関連づけを踏まえることで段階的にリスクを低減していくことを試みます。リスク対応策が過剰で非現実的なものにならないように、リスクの大きさと費用対効果を踏まえて最適なリスクコントロールを検討する、というのがRCModelの考え方です。

各層でどのような対策（コントロール）を取るべきかを関係者間（開発プロジェクト、データサイエンティスト、サービス提供者、ユーザ等）で取るべき対策や責任の範囲を考えます。多くのリスクシナリオは複数の層に線が引かれるため、開発者だけではなくサービス提供者や利用者もAIのリスク対策を考える必要性があります。

■サンプルケース「採用AI」P.11（リスクシナリオ「R001.適切な評価」の検討例）

リスクコントロールの内容（Case01：採用AI）

R001	**適切な評価** 採用する職種等により、適切な期待値を設定しなければAIの貢献を正しく評価できない

コントロールの内容		
AIシステム （A社AI開発部門）	サービスプロバイダ （A社人材採用部門）	ユーザー （A社グループ人材採用担当者）
③【Capability】複数のモデルを実行できるシステム環境を用意する（A社情報システム部門） ④【Data Balance】職種ごとに十分な学習データを確保する（A社AI開発部門） ⑤【Accuracy】職種ごとに適切な予測精度を持つモデルを開発する（A社AI開発部門） ⑦【Traceability】利用時のログ情報を保存（A社情報システム部門）	①【Accountability】職種ごとに適切な目標値を設定する（A社人材採用部門/A社グループ人材採用部門） ②【Scalability】複数のモデルをサービス提供できる体制を確保する（A社人材採用部門） ⑧【Auditability】各モデルの性能を定期的に検証する（A社人材採用部門/A社グループ人材採用部門） ⑨【Sustainability】必要な精度を確保するためにAIモデルの再学習を依頼する（A社人材採用部門）	⑥【Proper Use】実際の合否結果をフィードバック（A社グループ人材採用担当者）

11

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

各リスクシナリオについてリスクチェーンによるリスクコントロールを検討したら、社内外の各組織／委託先／利用者を含む各ステークホルダーの役割を整理します。

RCModelを用いて「実現すべき価値・目的」「重要なリスクシナリオ」「各ステークホルダーの役割」に係る一連の検討過程を整理することによって、関係者間で共通認識を持つことが期待できます。

■サンプルケース「採用AI」P.7-8

リスクアセスメント＆コントロール（Case01：採用AI）
- リスクの複雑性評価 → リスクコントロールを検討 -

実現すべき価値・目的（リスクの影響先）		リスクNo.	リスクシナリオ	技術的難易度	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	人材採用レベルの維持・向上	R001	適切な評価	○			●	個別モデルの開発環境 個別モデルの学習データ 個別モデルの開発	個別モデルの目標精度 個別モデルの開発体制 個別モデルの性能監視	正確なフィードバック
		R002	予測性能の維持	○			●	予測性能の確保 利用ログの記録	予測性能の検証 代替運用の取り決め	代替運用の実施
		R003	ノイズによる影響	○			●	敵対的事例の学習 モデルの頑健性 判断根拠の出力	判断根拠の分かりやすさ 開発チームへの連携 判断根拠の検証	判断根拠の検討
		R004	虚偽の申込	○			●	予測性能の確保 判断根拠の出力	過去の不正事例との検証 利用部門への注意喚起	人間による最終判断
		R005	過度なAI依存	○		○	●	予測精度の確保 判断根拠の出力	AIモデル性能の開示 判断根拠の分かりやすさ	予測性能・リスクの認識 人間による最終判断
		R006	誤ったフィードバック	○		○	●	学習データの検証 学習時の情報の保管	判断精度や異常値の検証 教師ラベルのデータ修正	正確なフィードバック 人事システムからの反映
		R007	人材トレンドの変化	○	○		●	データ分布変化の認識 汎化性能の見直し	データ分布／正解率の検証 再学習	人材トレンド変化の認識
		R008	新たな職種	○	○		●	個別モデルの開発環境 個別モデルの学習データ 個別モデルの開発	個別モデルの開発体制 個別モデルの検証	新たな職種要件の検討
2	採用活動に係るコストの削減	R009	コスト超過						コスト管理	
3	海外グループを含めたサービス提供	R010	地域の会社への対応	○	○		●	個別モデルの開発環境 個別モデルの学習データ 個別モデルの開発	個別モデルの目標精度 個別モデルの開発体制 個別モデルの性能監視	
		R011	不十分な開発スピード					再学習環境の準備 モデル開発者の確保	PJ体制の確保	
4	企業の社会的責任(公平性のある採用活動)	R012	判断根拠情報の不正販売	○		○	●	判断根拠の出力 利用ログの記録	管理責任の合意形成 不正利用の監視 外部弁護士との連携	不正利用の管理責任 ペナルティの理解 担当のローテーション
		R013	公平性	○		○	●	データの偏り モデルの汎化性	公平性ポリシーの検討 ネガティブな判断の開示	AIの判断傾向の理解 公平な最終判断
		R014	予測結果の目的外利用			○		データ保護	アクセス管理 目的内利用	データの取扱
		R015	風評被害			○		データ保護	職業倫理の教育	職業倫理の教育
		R016	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

組織におけるリスクマネジメント（Case01：採用AI）
- 各ステークホルダーにおける役割の定義 -

- 責任者 -
A社）経営者
- 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
- 公平性ポリシーの策定

申込者(データ提供者)

A社）法務・コンプラ
- 社内通報窓口
- 外部弁護士との連携
- 職業倫理の教育

A社）内部監査
- システム内部監査

人材エージェント

- AIサービスプロバイダ -
A社）人材採用部門
- 個別モデルの目標精度
- 個別モデルの開発体制
- 判断根拠の分かりやすさ
- AIモデル性能の開示
- ネガティブな判断の開示
- 再学習
- 管理責任の合意形成
- 代替運用の取り決め
- 個別モデルの性能監視
- データ分布／正解率の検証
- 判断根拠の検証
- 教師ラベルのデータ修正
- 不正利用の監視
- 開発チームへの連携
- 利用部門への注意喚起
- コスト管理

A社）AI開発部門
- 予測性能の確保
- 判断根拠の出力
- 汎化性能の見直し
- データ分布変化の認識
- 個別モデルの学習データ
- 学習データの検証
- 敵対的事例の学習
- モデルの頑健性
- モデル開発者の確保

A社）情報システム部門
- 再学習環境の準備
- 個別モデルの開発環境
- 利用ログの記録
- 学習時の情報の保管
- データ保護

- ユーザー -
A社グループ）人事部
- 不正利用の管理責任
- 人材トレンド変化の認識
- 新たな職種要件の検討
- 予測性能・リスクの認識
- 人事システムからの反映
- 担当のローテーション

- ユーザー -
A社グループ）人材採用担当
- 人間による最終判断
- 公平な最終判断
- 正確なフィードバック
- 判断根拠の検討
- 代替運用の実施
- ペナルティの理解

RCModelの利用タイミングとさらなる展開

RCModelの利用については、なるべくPoCや開発段階で検討を行い、利活用が始まる前に必要なリスクコントロールを実現しておくことが望まれます。

しかし、ビジネス環境や関係する法制度等の変化によって、重要なリスクシナリオ自体が変化する可能性があるため、利活用においても定期的に見直しして、必要なリスクコントロールを見直すことが望まれます。

また、RCModelによる一連の検討過程について、社内外の各ステークホルダー及び関連する有識者・専門家を含めたディスカッションを行うことで、AIに関する社会的・法的な論点・知見を把握し、AIサービスに係る将来課題を事前に検討することが望まれます。

主な論点

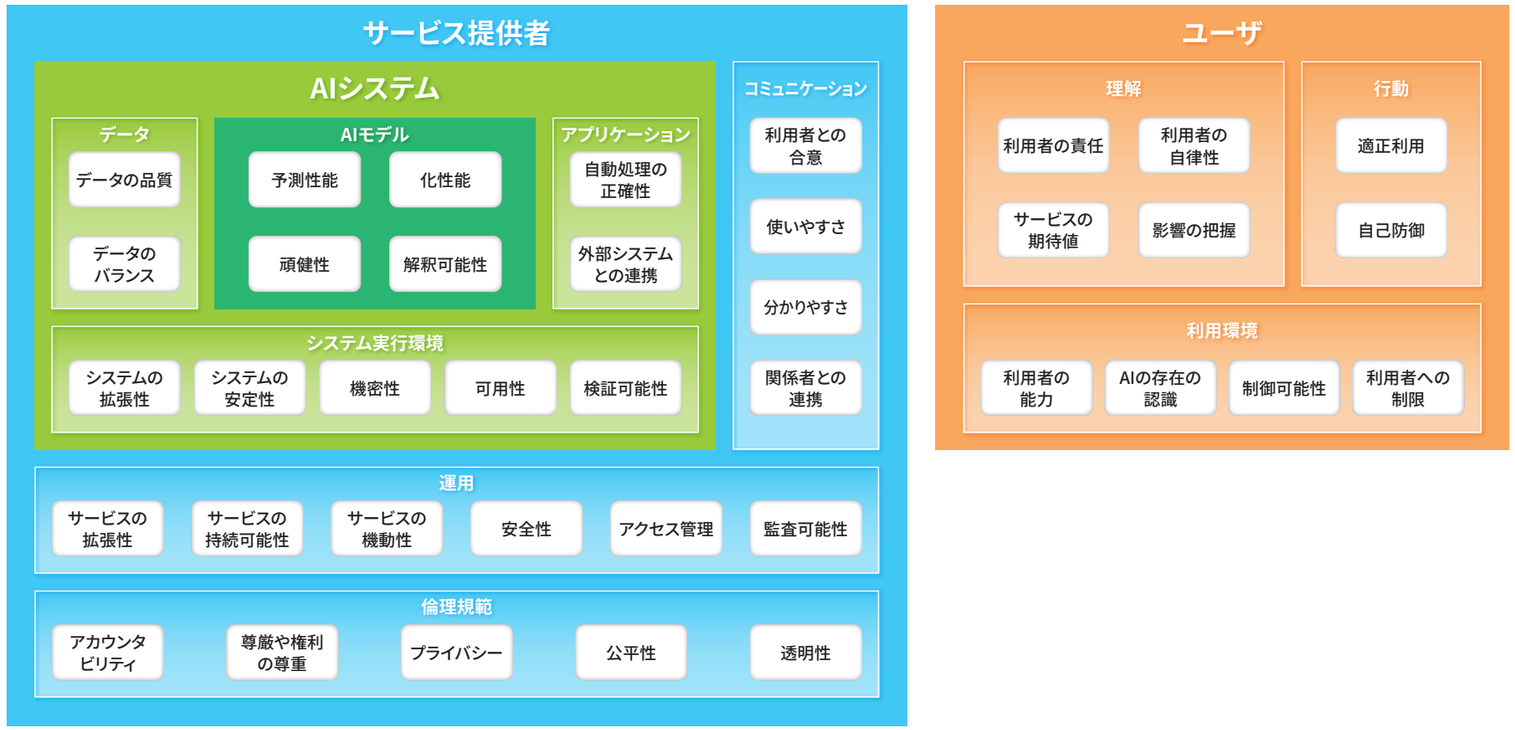
1. 本ケースが扱う分野についてビジネス環境/社会的要請はどのように変化していくか？
2. 上記変化に対してテクノロジー・サイエンスはどのように対応できるか/実現可能か？
3. テクノロジーと人間の関係性はどうなるか？
4. 海外の動きは？
5. 法改正に係る動きはどうなっているか？

RCModelの利用について

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事務局 ifi_tg@ifi.u-tokyo.ac.jp

RCM
RiskChainModel

RCModelの構成要素
(Ver 1.0)



AIシステム

構成要素	内容	主なコントロールの例
AIモデル		
予測性能	十分な予測性能	予測性能の検証
汎化性能	特定の事例に偏ることが少ない判断	汎化性能の検証
頑健性	ノイズ等への耐性	敵対的事例の学習
解釈可能性	モデルの判断根拠	判断根拠情報の出力
データ		
データの品質	データの品質確保	教師ラベルづけの検証 ノイズの補正 特別データの分類(戦略商品等)
データのバランス	データの分布	十分な学習データの確保 データバイアスの検証
アプリケーション		
自動処理の正確性	補助的な自動処理	異常値の補正 異常判断(過検出等)のアラート
外部システムとの連携	外部システム等との連携	データの自動連携 プロトコルの定義・実装
システム実行環境		
システムの拡張性	利用環境の増加等へ対応できる システム環境	複数モデルを実装する環境
システムの安定性	安定稼働するシステム環境	十分なパフォーマンスの確保 IoTのメンテナンス
機密性	機能や情報が保護されたシステム環境	システムの保護 アクセスコントロールの実装
可用性	必要な時に使用できるシステム環境	システム稼働時間の確保 異常時の代替機能への切替
検証可能性	AIサービスを事後検証できるための性質	実行履歴や判断根拠の保存 開発時の学習結果の記録

サービス提供者

構成要素	内容	主なコントロールの例
倫理規範		
アカウンタビリティ	説明責任、利用者の保護	責任関係の明確化 適切な期待値の検討
尊厳や権利の尊重	利用者側の優先権の検討	人間の介在度合の定義
公平性	サービス全体での公平性	公平性を確保する項目の定義
プライバシー	個人情報の適切な管理	プライバシーポリシーの遵守
透明性	AIサービスに係る必要な情報の開示	ステークホルダーごとに開示すべき情報 の整理
運用		
サービスの拡張性	利用環境の増加等へ対応できる サービス体制	サービス提供に十分な体制 必要な言語へのサポート
サービスの持続可能性	持続的なサービスの維持	サービス要件の見直し 再学習・継続学習
サービスの機動性	適時の対応	アジャイル開発の実施
安全性	サービス全体での安全性の確保	システム外の安全対策 利用者の不正利用への対処
アクセス管理	適切なシステムの 利用権限・範囲の管理	適切なアクセス権限の設定 不正な利用者の制限
監査可能性	第三者を含む必要な監視	AIシステムのモニタリング トラブル発生時の検証 不正利用の監視
コミュニケーション		
利用者との合意	利用者との認識合わせ	利用者責任の合意 AIシステムの変更への合意
使いやすさ	AIサービスの使いやすさ	使いやすいUI
分かりやすさ	AIサービスが出力する情報の 分かりやすさ	判断根拠の分かりやすい表現
関係者との連携	利用者に対するリモートでのサポート	利用者のリモート支援 専門家との連携

ユーザ

構成要素	内容	主なコントロールの例
理解		
利用者の責任	利用者責任の理解	利用者責任の合意(最終判断等) AIシステムの変更への合意
利用者の自律性	自律的に判断すべき内容の理解	AIの判断の訂正
サービスの期待値	AIサービスに対する期待精度の理解	期待値(予測精度等)の理解 ネガティブな判断傾向の理解
影響の把握	利用者への影響の把握	AIサービスに係るリスクの理解 ワークスタイルの変化の理解
利用環境		
利用者の能力	サービス利用において 必要な知識・スキルの習得	必要なリテラシーの確保 利用者の教育
AIの存在の認識	AIの存在の認識	AIの存在の明示 利用者への通知
制御可能性	利用者側での制御	マニュアル手続の整備 AI判断の訂正機能
利用者への制限	不正利用等の防止に向けた 利用者の制限	利用者への技術的な制限 法的な制約
行動		
適正利用	正しいサービス利用	適切な目的に沿うサービス利用 教師ラベルの正しい設定
自己防御	利用者自身の保護	不利益が発生した際のクレーム 環境変化に係る事前連携

2021年6月 リスクチェーンモデル（RCModel）ガイド Ver 1.0