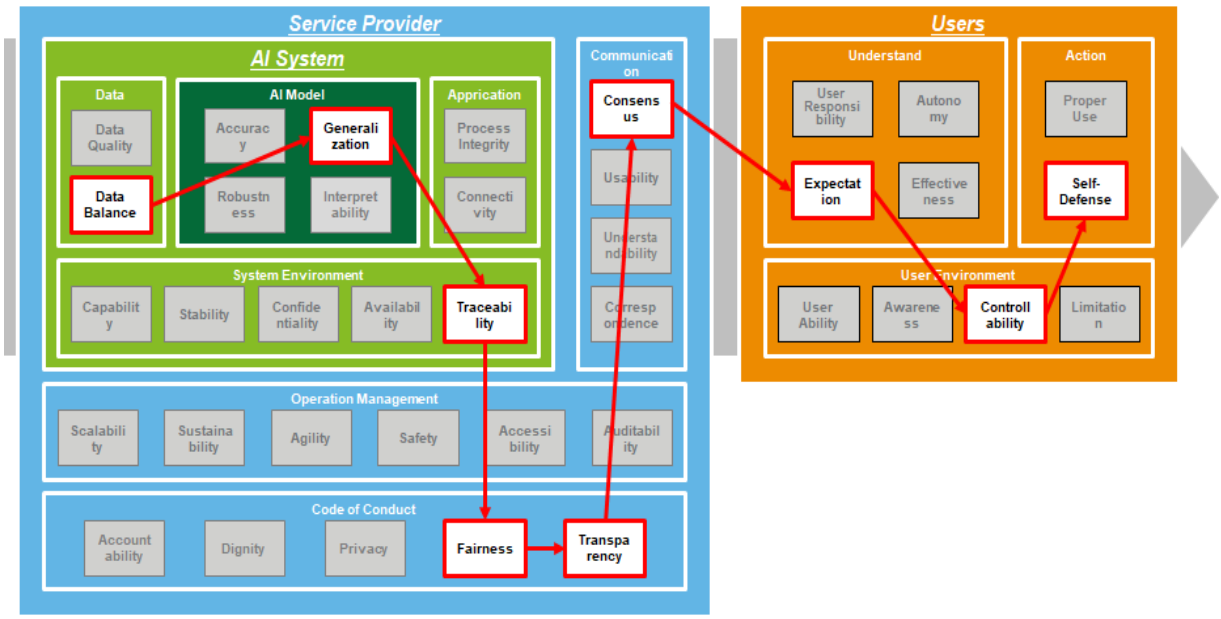
AI サービスのリスク低減を検討するリスクチェーンモデルの提案

松本敬史（有限責任監査法人トーマツ/東京大学未来ビジョン研究センター）
江間有沙（東京大学未来ビジョン研究センター/理化学研究所革新知能統合研究センター）

2020 年 6 月 4 日



リスクコントロールモデルの構成要素とリスクチェーン例

この政策提言は、以下の予稿論文に基づいて作成したものです。
松本敬史、江間有沙、「4N2-OS-26a-02 AI サービスのリスクコントロールを検討するためのモデル提案」、2020 年度人工知能学会全国大会、2020 年 6 月 12 日、熊本県熊本市
https://confit.atlas.jp/guide/event/jsai2020/subject/4N2-OS-26a-02/tables?cryptoId=

本政策提言は、議論の土台となるリスクアセスメントとコントロールのフレームワークを提供しています。今後、他のステークホルダーとの共同研究の積み重ねを通して、フレームワークを体系化していく予定です。

また、本政策提言は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の成果の一つです。

要旨

近年、人工知能（AI）サービスや製品の社会実装が拡大する一方で、AI の信頼性に係る問題が発生しており、AI サービス提供者は多岐にわたるリスクへの備えが求められる。そこで、本政策提言では様々な AI サービス提供の形態の存在を踏まえつつ，AI サービス提供者が自らの AI サービスに係るリスクコントロール検討に資するモデル「リスクチェーンモデル（Risk Chain Model: RCModel）を提案した。信頼性の高い AI サービスの実現に RCModel が役立てられることを期待する。

■リスクチェーンモデルの概要

１）リスク構成要素の整理と構造化

AI サービスの提供にあたりリスク要因になり得る要素（以下、構成要素）は数多くある。それらを RCModel では、(1) AI システムの技術的構成要素、(2) サービス提供者の行動規範（ユーザとのコミュニケーションを含む）に係る構成要素、(3) ユーザの理解・行動・利用環境に係る構成要素に構造化した。

２）リスクシナリオの識別とリスク要因となる構成要素の特定

不公平な判断、制御不能による事故等、AI サービスに係るリスクシナリオを識別した。そして優先して検討すべきリスクシナリオについて、リスク要因となる構成要素を特定した。

３）リスクチェーンの可視化とリスクコントロールの検討

リスクは構成要素単体では十分に低減することが難しいため、AI サービス提供者はリスクシナリオに関連する構成要素の関係性（リスクチェーン）を可視化することで、段階的なリスク低減の検討が可能になる。それによって、リスク要因の所在や効果的かつ効率的なコントロールを検討できる。

■リスクチェーンモデルを用いた今後の AI サービスの開発や実装への提言

提言 1：リスクシナリオと構成要素の理解促進

サービス提供者は、自らの AI サービスの構成要素を適切に把握する必要がある。また、AI をめぐる社会的な事例や事件にも注意を払い、重要なリスクシナリオを認識するべきである。

提言 2:：リスクチェーンモデルを用いた適切なリスクコントロールの推進

サービス提供者は、RCModel のリスクチェーンを可視化してリスクコントロールを検討すべきである。必ずしもすべてのリスクを低減する必要はなく、リスクの大きさや技術的難易度、費用対効果、継続性等を考慮して適切なコントロールを企業内に構築すべきである。

提言 3: ステークホルダー間での対話の促進とアップデート

AI サービス提供者、AI 開発者、利用者の間での対話促進に RCModel を用い、リスク許容度の明確化、リスクシナリオの作成、リスク構成要素の構造化、リスクコントロールモデルの検討、それぞれの責任範囲に関する共通理解を適宜アップデートする体制を構築するべきである。

1. 問題の所在と目的

人工知能（AI）サービス[1]の社会実装が拡大する一方で、AI の信頼性に係る問題（不公平な判断、制御不能による事故等）が発生している。さらに AI（特に深層学習）による判断は一定ではなく変動すること、データ提供や AI モデル開発を外部委託している場合は AI サービス提供者だけでリスクの充分な低減が難しいこと、ユーザの悪用・誤用により AI の性能が劣化・悪化する場合があることなど、AI サービス提供者は多岐にわたるリスクへの備えが求められる。

そこで本政策提言は、2 章で AI の信頼性をめぐる原則やガイドなどの先行研究や事例を紹介したのち、3 章で AI サービス提供者が AI サービスに係るリスク対応（コントロール）を検討するに資するモデルとしてリスクチェーンモデル（Risk Chain Model: RCModel）を提案する。4 章では採用 AI を事例として RCModel のケーススタディを行う。5 章では RCModel の使い方に関する提言を行ったのち、終章では今後の課題や展望について概説する。

2. AI の信頼性をめぐる原則やガイドの必要性

2-1. AI の価値をめぐる原則と実践ガイドの概要

AI サービスが提供すべき価値を体系的にまとめた先行研究としては、産学官民から出されている原則の類型化[2]や、開発段階からの倫理的な観点に配慮するための論点整理[3]などがある。

また、原則を実践的なガイドへと落とし込む試みもすでに始まっている。シンガポール個人情報保護委員会[4]、欧州委員会[5]やドバイ政府[6]などは AI がもたらすリスクの自己評価シートを提供している。日本でも総務省情報通信政策研究所が AI 利活用ガイドライン[7]を公開している。

---

[1] 本政策提言においては人工知能（AI）を用いたサービスを議論対象として扱う。ここでの「AI サービス」とは AI モデルの判断を活用する役務提供を指し、AI スピーカー等の製品も含む。

[2] Anna Jobin, Marcello Ienca & Effy Vayena：The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1, 389-99, 2019.

[3] IEEE：Ethically Aligned Design First Edition, 2019 や Jessica Morley, Luciano Floridi, Libby Kinsey & Anat Elhalal：From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, Science and Engineering Ethics, 2019 など。

[4] Personal Data Protection Commission Singapore：Implementation and Self Assessment Guide for Organisations (ISAGO), 2020. PDPC の自己アセスメント第 1 版の内容の日本語解説に関しては、「AI と倫理に一石、シンガポールの戦略」（江間有沙、『日経 BP ムック　まるわかり！AI 開発 2019 人材戦略』、日経 BP 社、2018、154-9）等を参照。

[5] High-Level Expert Group on AI (HLEG) of European Commission：Trustworthy AI Assessment List(Pilot Version), 2019.最後に、129 項目（https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai）にわたる質問リスト（Trustworthy AI Assessment List(Pilot Version)）が掲載されている。

[6] Smart Dubai：Ethical AI Toolkit, 2018.　ウェブサイト（https://www.smartdubai.ae/initiatives/ai-principles-ethics）にツールキットが提供されている。

[7] 総務省 AI ネットワーク社会推進会議：報告書 2019, 2019.

シンガポール個人情報保護委員会（PDPC）は 2020 年 1 月に Model AI Governance Framework[8]を策定した。本フレームワークは Internal governance structures and measures（組織のガバナンス）、Determining the level of human involvement in AI-augmented decision-making（人間の介在度の定義）、Operations management（運用管理）、Stakeholder interaction and communication（ステークホルダーとのコミュニケーション）で構成され、各項目において企業（Master Card、GRAB、Facebook 等）での実例を紹介しながら説明を行っている。フレームワークと併せて、セルフアセスメント[4]及び企業におけるガバナンスの事例集[9]を公開している。

日本では総務省 AI ネットワーク社会推進会議が 2019/8 に AI 利活用ガイドライン[7]を公開している。「適正利用」「適正学習」「連携」「安全」「セキュリティ」「プライバシー」「尊厳・自立」「公平性」「透明性」「アカウンタビリティ」の 10 原則を定義し、AI サービスプロバイダ／ビジネス利用者／データ提供者／消費者的利用者の夫々が留意すべき事項を定めている。

2-2. 原則をプラクティスに落とす上での課題

これらの原則やガイドラインを参照することによって、AI サービスの信頼性に係る一般的な論点を広範に把握することができる。しかし、AI サービス毎に検討すべきリスクは異なるため、原則やガイドラインを単純なチェックリストとして使用しても重要なリスクにフォーカスがされない恐れがある。さらに運用コストが肥大化して AI サービスの導入効果が十分に得られないことが考えられる。

また、AI サービスの信頼性に係るリスクは、学習済 AI モデルのみではなく学習データ・入出力・利用者・利用環境等の複数の要因によって顕在化し得る。例えば、AI の公平性に係るリスクの場合、データのバイアス・アルゴリズムのバイアス・利用者判断のバイアスのそれぞれがリスク要因となり得る。それぞれのリスクは関連しあっており、時として、リスク同士がトレードオフの関係にあることもある。そのため、リスク対応は複数の要因を包括的に検討することが望まれる。

米 Google 社や Facebook 社のようにデータの取得・AI モデルの開発・サービス提供の全てを自社で実現している場合には、複数のリスク要因に対して最適なアプローチを実施することが可能である。しかし、日本においてはサービスの提供者・AI モデルの開発者・実行環境の提供者・データの提供者が異なるケースが多く、AI サービス提供者が複数のリスク要因を包括的に検討するためには、関係者間で対話するための枠組みが必要となる。

そこで、本政策提言では、AI サービス提供者が関係者（AI システム開発者や利用者）と対話し最適なリスク対応を検討するために、AI サービスの信頼性に係るリスク要因の関連性を可視化できるモデルを提供する。

2-3. リスクアセスメントとコントロール

AI サービス様々な領域に展開されているため、AI サービスの信頼性を確保するためには、まずはサービス享受に伴って発生するだろうリスクの種類や量、誰がリスクを被るのかなどについて整

---

[8] PDPC: Model AI Governance Framework Second Edition, 2020.

[9] PCPC: Compendium of Use Cases: Practical Illustrations of the Model AI Governance Framework, 2020.

理する必要がある[10]。リスクの種類や関係者が吟味されたら、次にリスクの影響度と発生可能性を加味した上で重要なリスクシナリオを特定し、リスクシナリオに関連するリスク要因を網羅的に把握する（リスクアセスメント）。その上でリスク要因間の関係性を認識して、効果的かつ効率的なリスク対応（リスクコントロール）を実装する。

サイバー攻撃に対するリスクコントロールを検討する方法として、スイスチーズモデル[11]等の「多層防御」の考え方がある。単一のリスク対応策では十分にリスクを低減できない可能性があるため、複数のリスク対応策を多層化することによって十分なリスク低減を試みる。米国安全保障局（NSA）の多層防御戦略[12]では「人的要素」「技術要素」「運用要素」の各層においてリスク対応策を実現することを提唱している。

また企業がリスクコントロールを実装する際に参考となる考え方として、内部統制報告制度（J-SOX）がある[13]。J-SOX では財務報告に係る重要なプロセスにおいてリスクシナリオを検討し、リスクの低減（コントロール）を図る。例えば、「販売プロセス」において「架空売上の計上」というリスクシナリオを検討する場合、(1)受注：与信のチェック、(2) 出荷：実在性の確認、(3)売上・請求：得意先による検収の確認、(４) 債権回収：債権残高の確認と、複数の業務においてリスクコントロールを実施する。各コントロールは単体ではリスクを十分に低減できるとは限らないため、複数のコントロール（「架空売上の計上」リスクシナリオでは四段階）を導入してリスク低減を図る。このような観点から本モデルでも、段階的なリスク低減をモデルに適用する。

3. リスクチェーンモデルの概要

前述のとおり、AI サービスの信頼性に係るリスク対応は複数のリスク要因を包括的に検討することが求められる。本モデルでは、AI サービスにおける各リスク要因の関係性（リスクチェーン）を可視化する。

リスクチェーンモデルを活用するには以下のフレームワークに従って検討を行う。まず構造化されたリスクチェーンモデルの構成要素を確認し（3-1)、個別の AI サービスに係るリスクシナリオを識別して、取り組むべきシナリオの優先順位をつける（3-2)。次に、各リスクシナリオにおいてリスク要因となる構成要素を特定して（3-3)、構成要素の関係性（リスクチェーン）を可視化する（3-4)。最後に各構成要素で適切なコントロールを実装する(3-5)。以下、それぞれの項目ごとに概説する。

---

[10] 金融機関などリスクを積極的に引き受ける領域においては、組織の目的達成のためには進んで受け入れるべきリスクの種類と量を「リスクアペタイト」として表現するリスクアペタイトフレームワーク（RAF）の構築が行われてきていた。参考として「リスクアペタイト・フレームワークの構築」（大山剛、2015、中央経済社）など。

[11] J. Reason, E. Hollnagel, J Paries : REVISITING THE « SWISS CHEESE » MODEL OF ACCIDENTS, 2006.

[12] NSA : Defense in Depth, 2010.

[13] 企業会計審議会（FSA）：財務報告に係る内部統制の評価及び監査に関する実施基準, 2019.

3-1. リスク要因となる構成要素の構造化

AI サービスを構成し、かつリスク要因になり得る要素（以下、構成要素）は数多くあり、各国のガイドラインでは独自の検討によって構成要素を定義している。各国のガイドラインを元にして、Fairness（公平性）や Privacy（プライバシー）などいくつかに類型化している研究もある[14]。

しかし、同一の表現であってもガイドラインごとに表現している範囲や内容が異なる場合がある。例えば「Transparency」という表現について、日本の AI 利活用ガイドライン[7]では「Traceability（検証可能性）」と「Explainability（説明可能性）」を含めており、European Commission の Ethics Guideline for Trustworthy AI[5]では「Traceability」「Explainability」に「Communication」を加えており、シンガポールの Model AI Governance Framework[8]では「Explainability」と「Transparency」を別の構成要素として扱っている。このように各ガイドラインで使用する表現の対象が異なることから、AI サービス提供者はどのガイドラインを参考にすれば良いか判断が難しい。

また、リスク対応を検討するためには、構成要素は AI システム（技術）、AI サービスプロバイダ（運用）、あるいはユーザのいずれに起因するものなのかを把握できる形に分類されていることが望ましい。例えば「Explainability（説明可能性）」を実現する上では、AI モデルの判断根拠を可視化する「Interpretability（解釈可能性）」と、AI サービスプロバイダによって人が理解できる表現にする「Understandability（理解可能性）」によって実施すべき対応が異なる。

そこで本研究では、国内外で発行されている AI 倫理やガバナンスに関するガイドラインで言及されている論点を整理し、AI サービスに係るリスク要因となる構成要素を（1）AI システム、（2）AI サービスプロバイダ、(3)ユーザの三つの層に分類してモデルを作成した（図１）。第一の層（AI システム）は技術的な層であり、AI モデル、データ、ルールベースのアプリケーション、システム実行環境に係る構成要素が含まれる。第二の層（AI サービスプロバイダ）は AI システムを包含しながらも利用者に向けたサービス運用の層であり、行動規範（Code of Conduct）、オペレーションマネジメント、コミュニケーションが含まれる。第三の層（ユーザ）は AI サービスの直接の利用者であり、理解・行動・利用者環境を含む。構成要素の整理に用いたガイドラインリストは附録 1、各構成要素の定義は附録 2 を参照されたい。

---

[14] Yi Zeng, Enmeng Lu, Cunqing Huangfu：Linking Artificial Intelligence Principles, 2018. https://arxiv.org/abs/1812.04814

ただし、これらの構成要素は社会的な事件への考察や新たなテクノロジーの開発によって、構成要素の解釈が変化し、新たな構成要素が追加される可能性がある。また、事例によっては構成要素の配置が変更される可能性もある。本研究では、現段階での考察をもとに構成要素を各層に配置している。

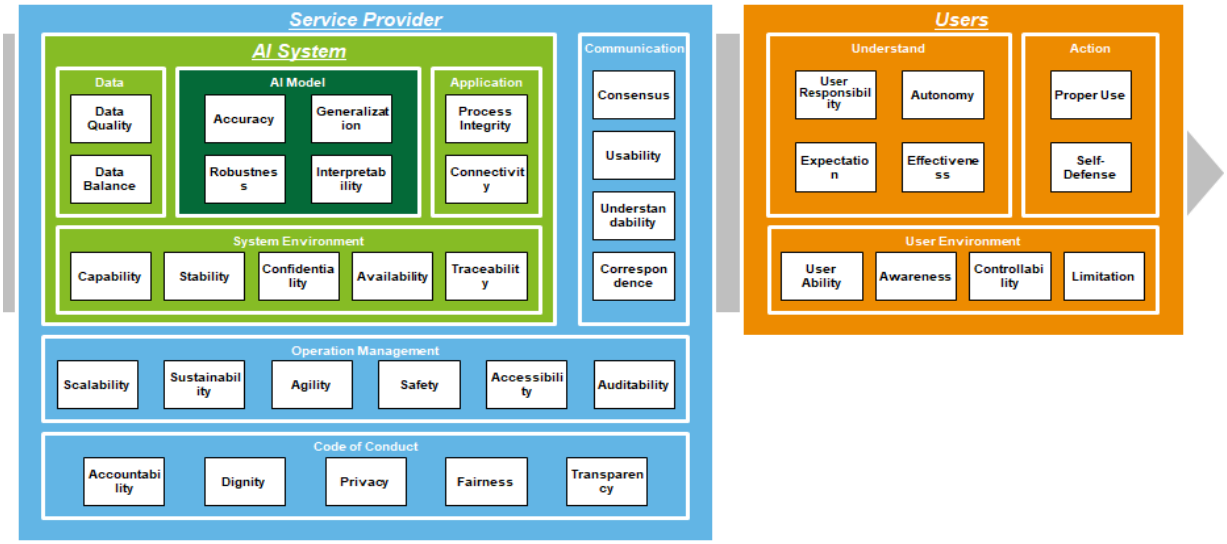


図 1 モデルの構成要素と構造

3-2. リスクアセスメントとシナリオの作成

AI サービスの信頼性に係るリスクは、サービスの品質だけではなく倫理や法令遵守に係る内容も含まれるため、リスクシナリオの一般化やチェックリスト化は難しい。そのためリスクシナリオの検討にあたっては AI サービスに係る人たちが、様々な観点からシナリオを作成することが求められる。作成にあたっては、(1) 各国のガイドライン（附録 1）で示されている一般的なリスク観点に加えて、(2) AI サービスの信頼性を損ねるような事件や事故などを参照しながら、(3) 技術、法務、営業など多様な業種、属性の人たちからなるグループを作り議論する必要がある。一部企業では、社内ではなく社外で有識者による委員会を作って検討することも行われている[15]。

表 1 では、(1) 報告書で示されている懸念や(2) 具体的な事件や事故を事例として、AI の信頼性を損ねるようなリスクを例示している。このような観点から、各 AI サービスに係るリスクシナリオを識別するが、全てのリスクシナリオに対応することは難しい。そのため、各リスクシナリオの影響度と発生可能性を鑑みて、優先的に取り組むべきリスクシナリオを特定する[16]。

---

[15] 本政策提言の目的は後述するリスクチェーンモデルの枠組みを提案することであるため、リスクシナリオ導出方法やアセスメントの方法、企業ガバナンスの検討は対象外とする。ただし、日本国内においても、提供する AI サービスのリスクなどの評価をする委員会は組織されており、例えば、富士通は AI 倫理外部委員会を設けており、ABEJA も「Ethical Approach to AI」(EAA)を設立している。

[16] 日本内部監査協会 : リスク評価手法の内部監査での 25 の活用事例 ～内部監査での活用方法・改善提言のための確認事項～, 2016, http://www.iiajapan.com/pdf/kenkyu/a03_1611.pdf

表 1　AI サービスをめぐるリスクの一般事例

リスク	(1) 報告書で示されている懸念	(2) 具体的な事件や事故
公平性に係るリスク	特定の属性を持つ利用者グループに対して、不当にネガティブな判断を行う[17]	AI による人材採用において女性に対して極端にネガティブな判断を行った[18]
頑健性に係るリスク	データに微細なノイズが含まれることで著しく誤った判断を行い、利用者に不利益・被害を与えるリスク[19]	自動運転車が道路標識を認識する際に、人間の目でも判断できないような微細なノイズが含まれることで、著しく誤った判断を行う危険性が指摘されている[20]
説明責任に係るリスク	AI の判断結果に対して判断根拠を説明できないことで、トラブル発生時に十分な説明責任を果たせないリスク[21]	医療における AI の活用において、医師が「ブラックボックス」である AI の判断結果を正しく解釈し、患者に十分な説明ができるかが懸念されている[22]
適正利用に係るリスク	利用者が不適切な利用を行うことで、AI サービスの性能が劣化し、別の利用者に対して不利益を与えるリスク[23]	チャットボットとの会話において特定のユーザが多くの差別発言を行うことで、チャットボット自体が問題発言を多く行った[24]

3-3. リスク要因となる構成要素の特定

特定したリスクシナリオ（3-2）について、リスク要因となる構成要素(3-1)を識別し、構成要素間の関係性（チェーン）を可視化する（3-4)。リスクチェーンを可視化することで、AI サービス提供者はどの構成要素に対するコントロールを検討すればよいかの判断をつけられる。リスクシナリオと構成要素を紐づけるためには、リスクシナリオに起因するリスク要因を分解し、附録 2 の定義を元に関係する構成要素を選ぶ。

---

[17] 総務省：AI 利活用ガイドライン[7]に「⑧ 公平性の原則」が含まれる。

[18] Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women, Fortune, https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/

[19] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy : Explaining and Harnessing Adversarial Examples, 2014, https://arxiv.org/abs/1412.6572

[20] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, Dawn Song : Robust Physical-World Attacks on Machine Learning Models, 2017, https://arxiv.org/abs/1707.08945v3

[21] 総務省：AI 利活用ガイドライン[7]に「⑩ アカウンタビリティの原則」が含まれる。

[22] 日本医師会 学術推進会議 : 第IX次 学術推進会議 報告書 人工知能（AI）と医療, 2018, http://dl.med.or.jp/dl-med/teireikaiken/20180620_3.pdf

[23] 総務省：AI 利活用ガイドライン[7]に「① 適正利用の原則」が含まれる。

[24] TIME : Microsoft Takes Chatbot Offline After It Starts Tweeting Racist Messages, 2016.

3-4. リスクチェーンの可視化

その後、各層（AI システム/AI サービスプロバイダ/ユーザ）においてリスクシナリオに関係する構成要素について、リスク要因が顕在化する順序を踏まえて結線していく（リスクチェーンの作成）。リスクチェーンは必ずしも一方向かつ単線であるとは限らず、AI システムが起点となるとも限らない。リスクチェーンを共通基盤として関係者間での検討を蓄積しブラッシュアップすることによって、AI サービスに係るリスクについての議論を深化することが重要である。

3-5. リスクチェーンを用いたコントロールの検討

リスクチェーンを可視化することで、AI サービス提供者は各構成要素を関連づけて効果的かつ効率的なリスク低減策（コントロール）を検討することが可能となる。全ての構成要素において対応策を実現できればリスクをより低減することが可能と考えられるが、実際には「特定の集団に係るデータが入手困難である」「AI モデルの複雑性が高すぎる」「ユーザ側に十分な判断スキルを持たせることが困難である」等の制約事項によって、一部のコントロールが有効に機能しない／コストが掛かりすぎる／実現が不可能である状況が想定される。そのため、AI サービス提供者はリスクチェーンで登場する全ての構成要素でコントロールを検討する必要はない。リスク低減の効果が高く、かつ費用対効果を含めて管理しやすいコントロールを優先して実現することが望ましい。

また、同一のリスクチェーンが引かれたとしても、AI サービス提供者のサービス体制・責任範囲・支配権・技術難易度等により、どの構成要素で重要なコントロールを導入するかは異なる可能性がある。そのため、多様なステークホルダーの知見を踏まえて有効なコントロールを検討することが重要である。

4. ケーススタディ：採用 AI サービスにおける RCModel 利用

本節では、具体的なケースを元にリスクチェーンモデル（以下 RCModel）の活用例を紹介する。ケースとしては、「企業の人材採用部門がエントリーシートから採用すべき応募者を判断する AI サービスを利用する」（以下「ケース」）を用いる。

4-1. ケースの具体的な紹介

本ケースは特定の AI サービスを対象としたものではなく、RCModel の利用例を紹介するために仮想的に設定しているものである。しかし、シナリオ作成などにリアリティを持たせるため、実際に以下のような状況を設定する。

ケースの概要

- 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。
- AI サービスプロバイダの 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。
- 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。※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合否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。

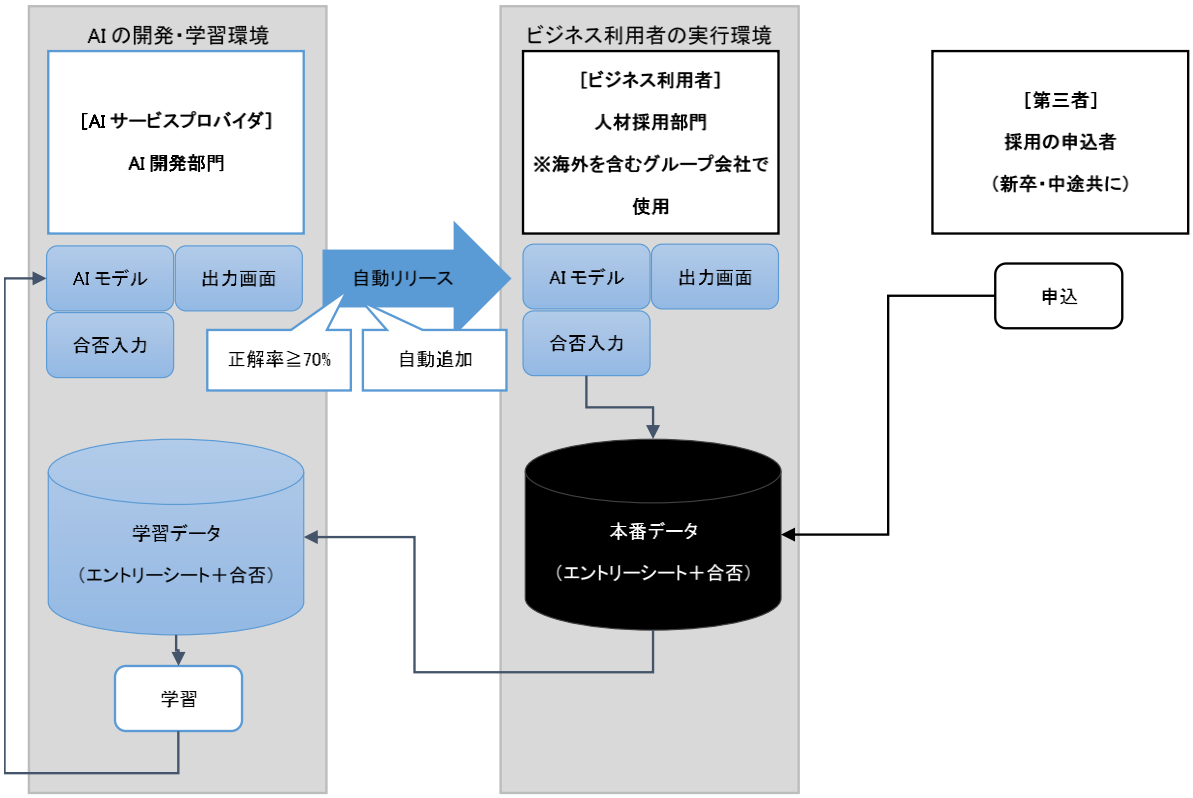
本ケースの AI サービスで使用されるデータ

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
過去のエントリーシートデータ	学習	申込者が A 社グループ人材採用部門に提出したエントリーシートデータ＋合否のラベル	A 社グループ会社人材採用部門責任者（A 社プライベートクラウド環境）	有（要配慮個人情報含む）
最新のエントリーシートデータ	本番	申込者が A 社グループ人材採用部門に提出したエントリーシートデータ	A 社グループ会社人材採用部門責任者（A 社プライベートクラウド環境）	有（要配慮個人情報含む）

AI サービスの出力内容

AI サービス利用者	A 社人材採用担当者
出力結果の内容	合否の判定
出力方法	A 社人材採用部門担当者の端末上で、申込者のエントリーシートを投入すると、書類選考の合否判定が出力される。
期待精度	適合率（Precision）：70% ※書類選考を通過した中で、実際に内定を出した比率
利用者判断の有無	有
根拠情報の出力	エントリーシートの中で判断に強い影響を与えたキーワードがハイライトされる
安全性のリスク有無	無
外部 AI への連携	無

AI サービスと利用者の関係図



4-2. リスクアセスメントとシナリオの作成

本ケースに係るリスクシナリオの識別を行った。まず一般的な観点として、総務省の AI 利活用ガイドライン[7]に沿って本ケースに該当するリスクシナリオを識別した。加えて、採用 AI に関する課題に係る報告書や提言書[25]を参考にして、本ケースのリスクシナリオに追加した。さらに各リスクシナリオについて「影響度」と「発生可能性」を検討した上で、優先度の高いシナリオから並べた結果が表 2 である。

[25] 本政策提言では HR テックの紹介は行わないが、パーソナルデータ＋α研究会が 2019 年 2 月に公開した「プロファイリングに関する提言案」では採用に関するプロファイリングをめぐる懸念や事例が紹介されている（https://www.shojihomu-portal.jp/nbl20190222）。また、採用 AI をめぐる平等やバイアスに関しては 2018 年 12 月に Upturn が報告書を公開している（Miranda Bogen and Aaron Rieke, Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias, 2018）。

表 2 採用 AI で想定されるリスクシナリオ

No	リスクシナリオ	参考情報
R001	特定の国／地域／人種／性別／年齢に対して、不適切な予測結果を生じさせる	AI 利活用ガイドライン[7]「⑧公平性の原則」 プロファイリングに関する提言案[24]
R002	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ず、適切な最終判断が行われない	AI 利活用ガイドライン[7]「⑦尊厳・自律の原則」
R003	人材採用担当者が AI サービスを何度も利用することで、AI が高確率で合格判断を行うエントリーシートやキーフレーズ等を特定し、社外で人材エージェント等に不正販売する	AI 利活用ガイドライン[7]「①適正利用の原則」(利用者の信頼性)
R004	人材採用担当者による AI へのフィードバック（エントリーシートにおける個人情報の匿名化、合否のラベル設定）が不正確なことで、AI が不適切な判断を行うようになる	AI 利活用ガイドライン[7]「①適正学習の原則」(不正確な学習データ)
R005	地域の会社ごとに採用方針や申込者のパーソナリティ（人種等）にバラつきがあるため、各グループ会社で学習データを用意しなければ適切な予測が行われない	ケース評価者より提起（適切な学習データの分布）
R006	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。	ケース評価者より提起 (頑健性に係るリスク[18])
R007	プライバシー情報の取扱を誤って、漏洩した際に適切な対応が取れず、被害の拡大及び法律違反（個人情報保護法等の違反）が発生する	AI 利活用ガイドライン[7]「⑥プライバシーの原則」

これらのリスクシナリオは 4-1 で紹介した技術や用途で使うものとして導出したものであるため、同じ採用 AI であっても、利用法やサービス提供方法が変化すれば導出されるリスクシナリオや優先順位も異なる可能性がある。

4-3. リスクシナリオごとで構成要素の特定

優先順位を踏まえて検討すべきリスクシナリオを特定したら、リスクシナリオ毎にリスク要因となる構成要素を識別する。本政策提言では、リスクシナリオ「R001 特定の国／地域／人種／性別／年齢に対して、不適切な予測結果を生じさせる」(以下、本リスクシナリオ)に係るリスク要因となる構成要素を識別した。識別した構成要素をリスクチェーンで関連づけする際にはリスクが顕在化する順番を考慮に入れるため、リスク要因を「予防：リスクを予防できない要因」「発見：リスクの実現を発見できない要因」「対応：リスクを発見しても適切に対応できない要因」の 3 段階で識別を行った(表 3)。

予防段階でのリスク要因としては、AI システムの層で「データのバイアスによって公平な判断が行われない（データの偏り：Data Balance)」と「アルゴリズムのバイアスによって公平な判断が行われない（汎化性：Generalization)」が識別された。

発見段階でのリスク要因としては、AI システムの層で「AI の判断根拠が分からない（解釈可能性：Interpretability)」及び「AI の判断結果を検証できない（検証可能性：Traceability)」が識別された。また AI サービスプロバイダの層としては、行動規範（Code of Conduct）として「公平な判断を行う上での留意点が明確になっておらず、判断尺度が人によって大きく異なる（公平性：Fairness)」及び「特定のグループにネガティブな判断をする傾向がある場合にその情報を可視化できない（透明性：Transparency)」が、コミュニケーションとして「ユーザ側での留意点について認識合わせを行わず、ユーザ側で公平な判断が行われない（合意：Consensus)」が識別された。

対応段階でのリスク要因としては、ユーザ（人材採用担当者）の層において、ユーザが理解すべき事項として「AI の判断へ過度に依存することで公平な判断が行われない（人間の自律性：Human Autonomy)」「特定のグループにネガティブな判断をすることを認識していなければ公平な判断が行われない（期待値：Expectation)」が、ユーザ側の環境として「意思決定プロセスが不明確なことで差別的な判断が補正されない（制御可能性：Controllability)」が、ユーザの行動として「最終判断の場において公平な選考が行われない（適正利用：Proper-Use)」が識別された。

表３　各段階でのリスク要因と構成要素の検討

予防	発見	対応
■ 【AI システム】データのバイアスによって公平な判断が行われない（Data Balance） ■ 【AI システム】アルゴリズムの汎化性能が損なわれ公平な判断が行われない（Generalization）	■ 【AI システム】AI の判断根拠が分からない（Interpretability） ■ 【AI システム】AI の判断結果を検証できない（Traceability） ■ 【サービス提供者】公平な判断を行う上での留意点が明確になっておらず、判断尺度が人によって大きく異なる（Fairness） ■ 【サービス提供者】特定のグループにネガティブな判断をする傾向がある場合にその情報を可視化できない（Transparency） ■ 【サービス提供者】ユーザ側での留意点について認識合わせを行わず、ユーザ側で公平な判断が行われない（Consensus）	■ 【ユーザー】AI の判断へ過度に依存することで公平な判断が行われない（Human Autonomy） ■ 【ユーザー】特定のグループに対してネガティブな判断をすることを認識していなければ公平な判断が行われない（Expectation） ■ 【ユーザー】意思決定プロセスが不明確なことで差別的な判断が補正されない（Controllability） ■ 【ユーザー】最終判断の場において公平な選考が行われない（Proper-Use）

4-4. リスクチェーンの可視化

リスクシナリオに係る構成要素と識別された項目を結線したのが図 2 である。前の工程（4-3）で識別された構成要素の順に従い、第一の層：AI システムにおいて「データの偏り（Data Balance）」→「AI モデルの汎化性（Generalization）」→「解釈可能性：Interpretability」→システム環境における「検証可能性（Traceability）」が登場する。次に第二の層：AI サービスプロバイダでは、行動規範として「公平性（Fairness）」→必要な情報の開示（「透明性（Transparency）」）→ユーザとの「合意（Consensus）」とつながる。最後に第三の層：ユーザにおいて、利用者である人材採用担当者の「人間の自律性（Human Autonomy））」→「期待値：Expectation」→「制御可能性（Controllability）」→「適正利用（Proper Use）」としてユーザの行動まで接続される。

このリスクチェーンは、情報の伝達順を踏まえて、AI システムを起点として単一方向に線を引いているが、線は一方向かつ単線である必要はない。また必ずしも AI システムを起点として線を引く必要もない。

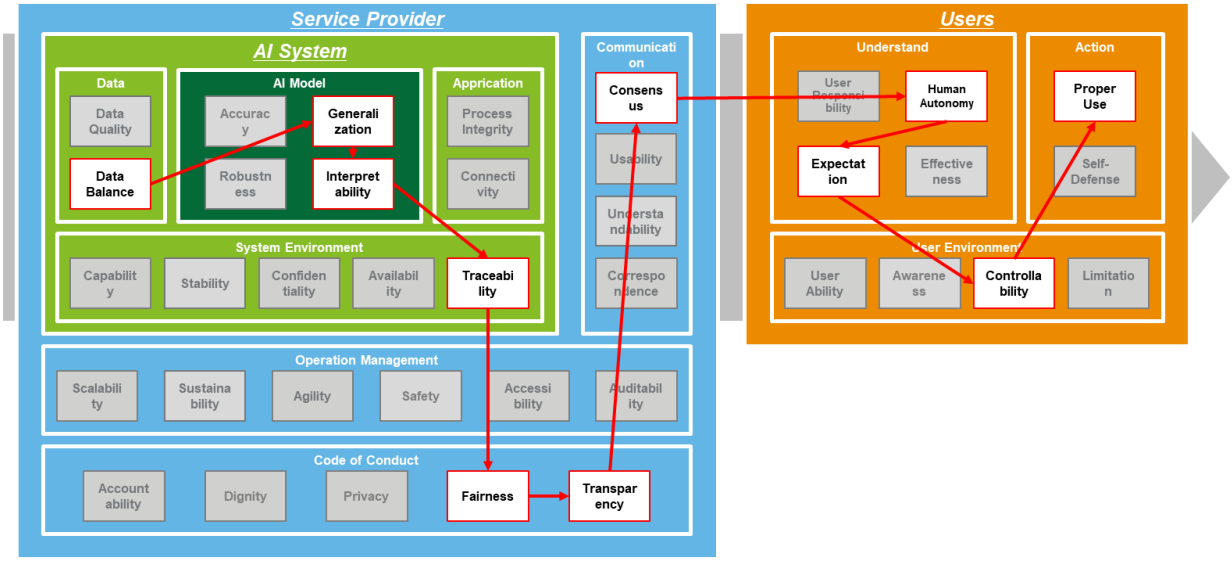


図 2　ケースにおけるリスクチェーンの例

4-5. リスクチェーンを用いたコントロールの検討

AI システム／AI サービスプロバイダ／ユーザにおける各構成要素において、検討されるコントロール案を表 4 に整理した。これらのコントロールを導出するにあたっても、リスクシナリオ作成時で参照した報告書や、社会的な事例や事件の対応などが参考となる。

また、本モデルは全てのコントロールの対策を行うためではなく、ここで上げられた対策のうち、リスク低減の効果が高く、かつ操作しやすい構成要素を検討することを目的として使う。そのためにも、関係者で実現可能なコントロールを特定し AI システム開発者や利用者に対してその理由や経緯を説明するための土台としても使うことができる。

表４　リスクシナリオでの構成要素ごとのコントロールの例

層	構成要素	コントロール
AIシステム	データの偏り(Data Balance)	● 学習データが特定の層（性別等）に偏らないようにデータの割合（男女比等）を整える
	汎化性（Generalization）	● モデルの説明変数から不公平な判断につながるような特徴量（性別，国籍等）を除く
	解釈可能性（Interpretability）	● モデルの判断根拠(特徴量の重要度等)の情報を出力できるようにする
	検証可能性（Traceability）	● モデル学習時の公平性に係る情報（属性別の判断結果の比較等）を保存する ● 利活用時の判断根拠に係る情報（判断結果に強く影響した特徴量等）を保存する
AIサービスプロバイダ	公平性(Fairness)	● 応募者に係る公平性の留意点（性別，国籍等）を整理し，関与者に周知する
	透明性（Transparency）	● 特定のグループにネガティブな判断をする傾向を除けない等，ユーザ（人材採用担当者等）が留意すべき情報を開示する
	合意(Consensus)	● AIサービスの予測精度・留意事項（特定のグループにネガティブな判断をする等)・ユーザの責任（最終判断等）について，ユーザと合意する
ユーザ	人間の自律性（Human Autonomy）	● AIサービス側から提供される留意事項を踏まえ，過度にAIに依存しないように判断する
	影響（Effectiveness）	● 予測精度及び利用者側の留意事項を確認する
	制御可能性（Controllability）	● どのユーザに合否の最終判断の権限があるかを明確にする ● 最終判断時の判断材料として，AIによる判断の採用／不採用を整理する
	適正利用（Proper Use）	● 特定のグループを差別するような判断となっていないか必要な確認を行い，合否を最終判断する

5. 提言

3 章で示した概要を適用して 4 章では仮想的な事例を用いて、具体的な RCModel の使い方を示した。RCModel を用いることによって、AI サービス提供者が自らの AI サービスに係るリスク対応（コントロール）を検討することが可能となる。モデルを用いた今後の AI サービスの開発や実装への提言は以下の通りである。

提言 1：リスクシナリオと構成要素の理解促進

サービス提供者は、自らの AI サービスの構成要素を適切に把握する必要がある。また、AI をめぐる社会的な事例や事件にも注意を払い、重要なリスクシナリオを認識するべきである。

提言 2:：リスクチェーンモデルを用いた適切なリスクコントロールの推進

サービス提供者は、RCModel のリスクチェーンを引いてリスクコントロールを検討すべきである。その際、必ずしもすべてのリスクを低減する必要はなく、リスクの大きさ（影響度や発生可能性）や技術的難易度、費用対効果、継続性等を考慮に入れた適切なコントロールを行う仕組みを企業内に構築すべきである。これにより、AI サービス提供者が最適なコントロールを検討し関係者に対して説明が行えるようになる。

提言 3: ステークホルダー間での対話の促進とアップデート

AI サービス提供者、AI 開発者、利用者の間での対話促進に RCModel を用い、リスク許容度の明確化、リスクシナリオの作成、リスク構成要素の構造化、リスクコントロールモデルの検討、それぞれの責任範囲に関する共通理解を適宜アップデートする体制を構築するべきである。

6. 今後の課題と展開

本政策提言は、AI サービス提供者がリスクチェーンを活用して最適なリスクコントロールを検討し関係者と対話、説明するためのモデルを示した。また、実際に採用 AI のケースを用いてモデルの使い方を紹介した。

RCModel では、リスクの構成要素を構造化し（3-1)、リスクシナリオの検討（3-2)、リスクシナリオ毎に要因となる構成要素の特定（3-3)、リスクチェーンの可視化（3-4）とコントロールの検討（3-5）というフレームワークを提供した。このような順番で様々なステークホルダーとともにリスクアセスメントとコントロールを行っていくことは、論点を促進しやすくする観点から重要である。また技術や仕組みのマイナーチェンジがあった場合でも、本フレームワークに立ち戻ることで、シナリオや構成要素、チェーンの結線の仕方などどこを変更すればよいのかを可視化でき、議論の透明化という観点からも有効である。

本政策提言は議論の土台となるリスクアセスメントとコントロールのフレームワークを提供したが、議論の方法やシナリオ導出やチェーン結線、コントロールの検討などの方法論に関しては今後、さらなる事例研究の積み重ねを通して体系化していく予定である。さらには、構成要素に関しても技術の進展や社会的な期待などを加味して定期的な見直しが求められる。また、医療、交通など様々な領域において使われている AI サービスにおいて、どのようなリスクをどの程度なら受け入れられるのかは、社会全体として議論をしていくことも重要となる。これらはリスクアセスメン

トやコントロールのフィードバックを経ながら議論を熟成していくべきものであり、ステークホルダー間だけのコミュニケーションだけではなく、広く社会としての対話の方法も構築をしていくことが今後の課題となってくる。

また、本政策提言は AI サービスや製品のリスクコントロールのみに焦点をあてて検討をした。しかし、AI サービス提供者には想定されていないリスクに対するレジリエンスや継続的な変化への対応力が求められており、組織レベルでのガバナンスの在り方についても検討が必要である。例えば、ポリシーの策定、専門性を持つ人材・体制の確保、開発・運用プロセスの構築、リスクアペタイトの設定等が挙げられる[26]。さらに、リスクの高い領域（交通・社会インフラ・医療・公共機関等）での AI サービスについては、客観的に信頼性を検証するために AI サービスに対する第三者評価のアプローチについても検討が必要である[27]。

AIサービスが信頼されて社会へと普及をしていくためには、AIの開発段階、サービス提供段階、さらにはユーザとのコミュニケーションにおけるガバナンスが不可欠であり、リスクチェーンモデルが、事前そして事後のリスク評価や関係者間の対話のツールとして用いられることを期待する。

謝辞

本稿を執筆するにあたって東京大学未来ビジョン研究センター技術ガバナンスユニットの皆様、有限責任監査法人トーマツの皆様からご助言をいただいた。また、日本ディープラーニング協会公共政策委員会の委員の皆様にも貴重なご意見をいただいた。本研究は「多様な価値への気づきを支援するシステムとその研究体制の構築（JSTRISTEX)」（代表：江間有沙）と「人工知能の倫理・ガバナンスに関するプラットフォーム形成(D18-ST-0008)」（代表：江間有沙）による研究成果の一部である。

[26] デロイトトーマツグループ：AI ガバナンスサーベイ 2019, 2020.

[27] 総務省 AI ネットワーク社会推進会議：報告書 2019, 2019.

附録１：参考リスト

AI 倫理やガバナンスに関するガイドラインは以下の通り（発行機関：参考資料名，国・団体，発行時期）

1. 総務省：AI 利活用ガイドライン, Japan, 2019 年 8 月.
2. OECD：Recommendation of the Council on Artificial Intelligence, OECD, 2019 年 5 月.
3. NIST：U.S.LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools, USA, 2019 年 8 月.
4. IMDA&PDPC：Model AI Governance Framework, Singapore, 2020 年 1 月.
5. HLEG of European Commission：ETHICS GUIDELINES FOR TRUSTWORTHY AI, EU, 2019 年 4 月.
6. The Alan Turing Institute：Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, UK, 2018 年 5 月.
7. 国家次世代 AI ガバナンス専業委員会：次世代 AI ガバナンス原則—責任ある AI の発展, China, 2019 年 6 月.
8. Smart Dubai：AI ETHICS PRINCIPLES & GUIDELINES, Dubai, 2018 年 12 月.
9. IEEE：Ethically Aligned Design First Edition, IEEE, 19-Mar.
10. Partnership on AI：Human - AI Collaboration Framework and Case Studies, Partnership on AI, 2019 年 9 月.

附録 2: 構成要素の定義

領域		構成要素	内容
AI System	AI Model	Accuracy	予測性能
		Generalization	汎化性能（アルゴリズムのバイアス）
		Robustness	頑健性（ノイズへの耐性）
		Interpretability	モデルの解釈可能性
	Data	Data Quality	データの品質、完全性、鮮度
		Data Balance	データの偏り（データバイアス）
	Application	Process Integrity	自動処理の正確性
		Connectivity	外部システムとのプロトコル
	System Environment	Capability	処理パフォーマンス、システムの拡張性
		Stability	安定稼働、誤り補正、反復性
		Confidentiality	機密性
		Availability	可用性
		Traceability	検証可能性、誤り検知
Service Provider	Code of Conduct	Accountability	サービスに係る答責性
		Dignity	利用者の判断・権利の尊重
		Privacy	プライバシーの保護
		Fairness	公平性
		Transparency	透明性、情報の可視化
	Operation	Scalability	変化に対する組織の柔軟性
		Sustainability	持続可能性、サービス品質の維持
		Agility	迅速な開発プロセス・インシデント対応
		Safety	安全性、エラーへの対応
		Accessibility	アクセスコントロール、権限管理
		Auditability	監査可能性
	Communication	Consensus	利用者との合意（目的、役割分担等）、委託先との権利関係
		Usability	使い易さ、人間系への切替、インタラクション
		Understandability	理解可能性、適切な表現（利用者の意思決定を誘導しない）
		Correspondence	利用者・関係者との連携
Users	Understand	User Responsibility	利用者の責任
		Expectation	サービスの期待値、提供範囲
		Human Autonomy	意思決定の自由度
		Effectiveness	サービスに係るリスク

	Action	Proper Use	適切な利用（悪用しないこと）
		Self-Defense	自己防御
	User Environment	User Ability	リテラシー、経験、スキル
		Awareness	AI の存在の認識
		Controllability	支配可能性、行動の選択肢
		Limitation	技術的・法的な制約