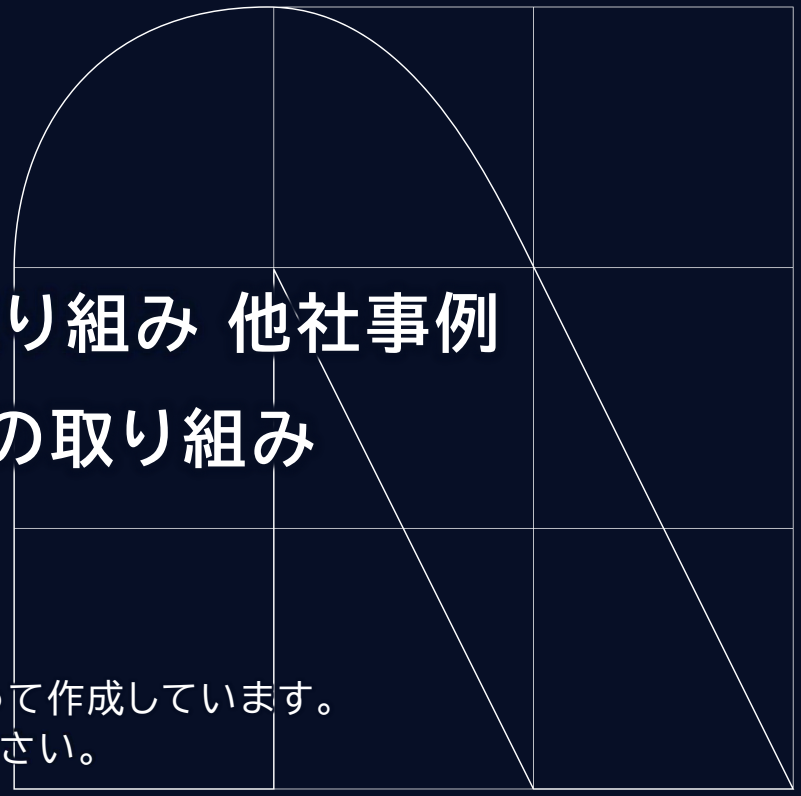
生成AI × サステナビリティレポート 2025

サステナビリティ領域での生成AI技術活用とAIガバナンスの重要性

CONTENTS



※本レポートは2025年10月16日時点で閲覧したWeb情報などを元にNTT DATAが主となって作成しています。
本レポート内の情報を引用する場合、その他お問い合わせについては以下からご連絡ください。
https://www.nttdata.com/jp/ja/contact-us/

1）サステナビリティ | NTT DATA

Chapter 0

はじめに

2024年は、身の回りの技術とサステナビリティの関連に着目したホワイトペーパーを公開し、大きな関心を集めました。2025年はその取り組みを継続、さらには発展させ、2024年に取り上げたいくつかの主要技術について、最新の知見や業界動向を踏まえて情報をアップデートするとともに、サステナビリティの観点からその活用方法や運用のあり方を再考していきます。2025年は、5つのテーマを扱う予定です。本ホワイトペーパーでは、そのうちの1つとして「生成AI（Generative AI）」を取り上げます。急速に普及する生成AIに焦点を当てつつ、その安全で責任ある活用を支える「AIガバナンス」の重要性に注目します。

AIガバナンスとは、AIの設計・開発・利用において、透明性、公平性、安全性、説明責任といった価値を確保し、技術のリスクを適切に管理するための制度・仕組みの総称です。生成AIは、業務効率化や新しい価値創出を可能にする一方で、意図しない差別的な表現やバイアスの介入、著作権侵害、誤情報の出力など、多岐にわたるリスクが指摘されています。こうした技術のリスクとどう向き合い、適切に制御・管理していくかは、持続可能な技術活用を考える上で極めて重要な課題です。

本ホワイトペーパーでは、生成AIに関連する社会的な課題の概観に加え、最新の技術・政策動向を整理しながら、AIガバナンスの実践的アプローチについて紹介します。

2024年ホワイトペーパー：生成AI×サステナビリティレポート-サステナビリティ領域での生成AI技術活用-

NTT DATAのサステナビリティ経営1）

NTT DATAはサステナブルな社会の実現に向けて「Planet positive, Prosperity positive, People positive」の3つの柱で取り組んでいます。

また、NTT DATAは解決すべき社会課題と、当社の事業における重要性を評価し、サステナビリティ経営として取り組むべき優先テーマとして13個のマテリアリティを設定しました。

本ホワイトペーパーで取り上げる生成AIにおけるAIガバナンスの取り組みは、マテリアリティのひとつである「責任あるテクノロジーの利用とAI倫理」の一環です。

［図0-1］NTT DATAの13のマテリアリティ

Chapter 1

生成AIの概要

技術動向と社会・環境との接点

進化する生成AI技術の現在地

生成AIは、テキスト・画像・音声・動画などの新しいコンテンツを自動生成する人工知能技術の総称です。近年、特に大規模言語モデル（LLM：Large Language Model）の発展により、自然な文章の生成や対話が可能となり、業務や日常生活に急速に浸透しています。2025年8月には生成AIの代表的なサービスであるChatGPTの全世界のユーザー数が7億人を超えました[1)]。また、生成AI業界のスタートアップ企業の中でも注目されているAnthropic社が、2025年秋にアジア太平洋地域（APAC）で初の拠点となる東京オフィスを開設し、AIアシスタント「Claude」の日本語版を提供開始すると発表しました[2)]。

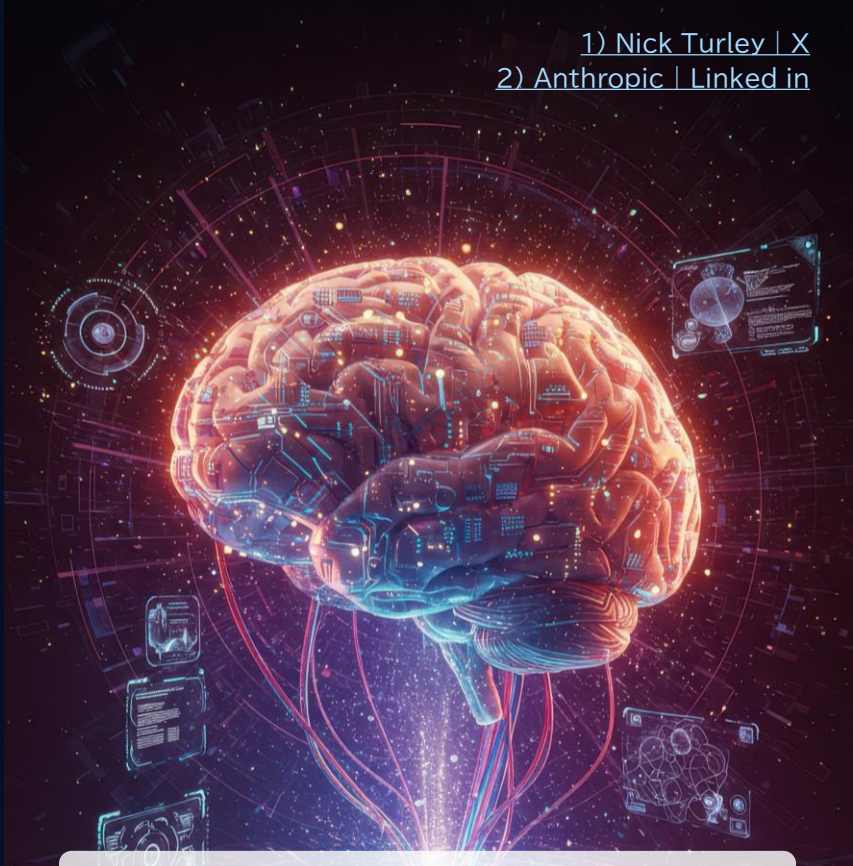
2024年のホワイトペーパーでは、生成AIの基礎的な仕組みやLLMの構造、Transformerアーキテクチャなどについて紹介しましたが、技術はこの1年で大きく進化しています。

たとえば、マルチモーダルモデル※1の進展が著しく、複雑な業務を統合的に支援できるようになってきています。

また、生成AIの軽量化や高速化が進んだことで、従来はクラウド環境での実行が前提だったモデルが、モバイル端末やエッジデバイス上でも動作可能になりつつあります。端末側の処理能力やメモリ容量の向上もこの流れを後押ししています。

さらに、ファインチューニング※2やLoRA※3といった個別最適化技術の進展により、特定業務や企業独自の言語スタイルに対応したカスタムモデルの構築が容易になっています。これにより、汎用モデルを業務特化型へと転換し、現場のニーズに即した高度な応答が可能となることで、実運用への導入が加速しています。

1) Nick Turley | X
2) Anthropic | Linked in



※1 マルチモーダルモデル
テキスト・画像・音声・動画を組み合わせて処理するAI

※2 ファインチューニング（fine-tuning）
すでに大規模な一般データで事前学習されたモデルを、特定の目的や業務に合わせて追加で学習させるプロセス

※3 LoRA（Low-Rank Adaptation）
既存の大規模モデルの一部パラメータだけを学習対象にすることで、計算コストとデータ量を大幅に削減しながら効果的なファインチューニングを可能にする技術

1) AIを用いた持続可能な未来シナリオ分析の共同実験を開始｜NTT DATA

生成AIとサステナビリティとの関連性

ここでは、生成AIとサステナビリティとの関連性について見ていきます。なお、本ホワイトペーパーの主要テーマはAIガバナンスですが、ここでは生成AIを取り巻くサステナビリティ課題を理解するために、環境面や社会面といったガバナンス以外の観点からも整理します。

生成AIは、業務効率化や情報アクセスの向上にとどまらず、適切に活用することで社会・環境への貢献も期待される技術です。

たとえば、業務効率化の面では、製造業での生産性向上や、部品設計の最適化支援、医療分野でのカルテ作成支援や患者向けのセルフケア支援など、様々な分野での導入が進んでいます。情報アクセスの向上に関しては、多言語対応や障がい者支援、学習支援などを通じて、言語や身体的な条件を問わず多様な人々が社会的活動や学びに参加できる環境づくりが進められており、誰もが取り残されない包摂的な社会の実現に寄与する可能性があります。

また、企業の持続可能な経営判断を支援するために、生成AIを活用した長期的な未来シナリオ分析の研究・開発も進められています。これにより、従来は人手に依存していた持続可能な未来に向けた戦略的意思決定を、より効率的・網羅的に実施できることが期待されています[1)]。

一方で、生成AIの開発や運用には依然として大量の計算資源と電力が必要であり、それに伴う温室効果ガスの排出量増加は、環境負荷の面で重大な課題とされています。特に大規模言語モデルの学習には、従来のAIよりもはるかに多くの電力量を消費することが報告されており、再生可能エネルギーの導入や、エネルギー効率の高いモデル設計が求められています。

加えて、フェイクニュースやディープフェイクといった情報の信頼性の低下、著作権侵害、アルゴリズムによるバイアスの再生産といった問題も深刻化しています。たとえば、ディープフェイクの悪用事例として、国政・地方選挙において偽画像や動画を拡散することで有権者を誘導する事例が発生しています。政治家が発言しているかのように見せかけたフェイクコンテンツが拡散されることは、民主主義の根幹を揺るがしかねない問題として警戒されています。

2025年には生成AIを悪用して多数の通信回線契約を不正に取得したとして、中高生が逮捕される事件が発生しました。このような事例は、生成AIの利便性が犯罪行為の手段としても利用されうる現実を浮き彫りにしており、対策の必要性が高まっています。

さらには、企業における情報漏洩リスクも顕在化しています。ある企業では、従業員がChatGPTに機密性の高い業務データをアップロードしたことで情報が外部に漏洩する事故が発生し、生成AIツールの利用が禁止されました。同様のリスクは他の企業にも及ぶ可能性があり、業務における生成AIの取り扱いには一層の注意とガバナンス体制の整備が求められています。

こうした課題を受けて、欧州をはじめとする各国でAI規制の整備が進んでおり、倫理に配慮した設計方針やシステムの透明性確保が今後ますます求められていくと考えられます。

AIガバナンスの重要性

前述した通り、生成AI活用によって様々なことが便利になる一方、誤情報やバイアスに関するリスク、情報漏洩リスク、著作権侵害リスクなどといった負の側面も顕在化しています。これらのリスクに対して、企業・政府・社会が連携し、AIの設計・開発・利用において信頼性や公平性を確保するための枠組みとして「AIガバナンス」の整備が不可欠です。

AIガバナンスには、以下のような構成要素が挙げられます。

・倫理原則・・・透明性・公平性・安全性・説明責任・非差別などを確保するための価値基準

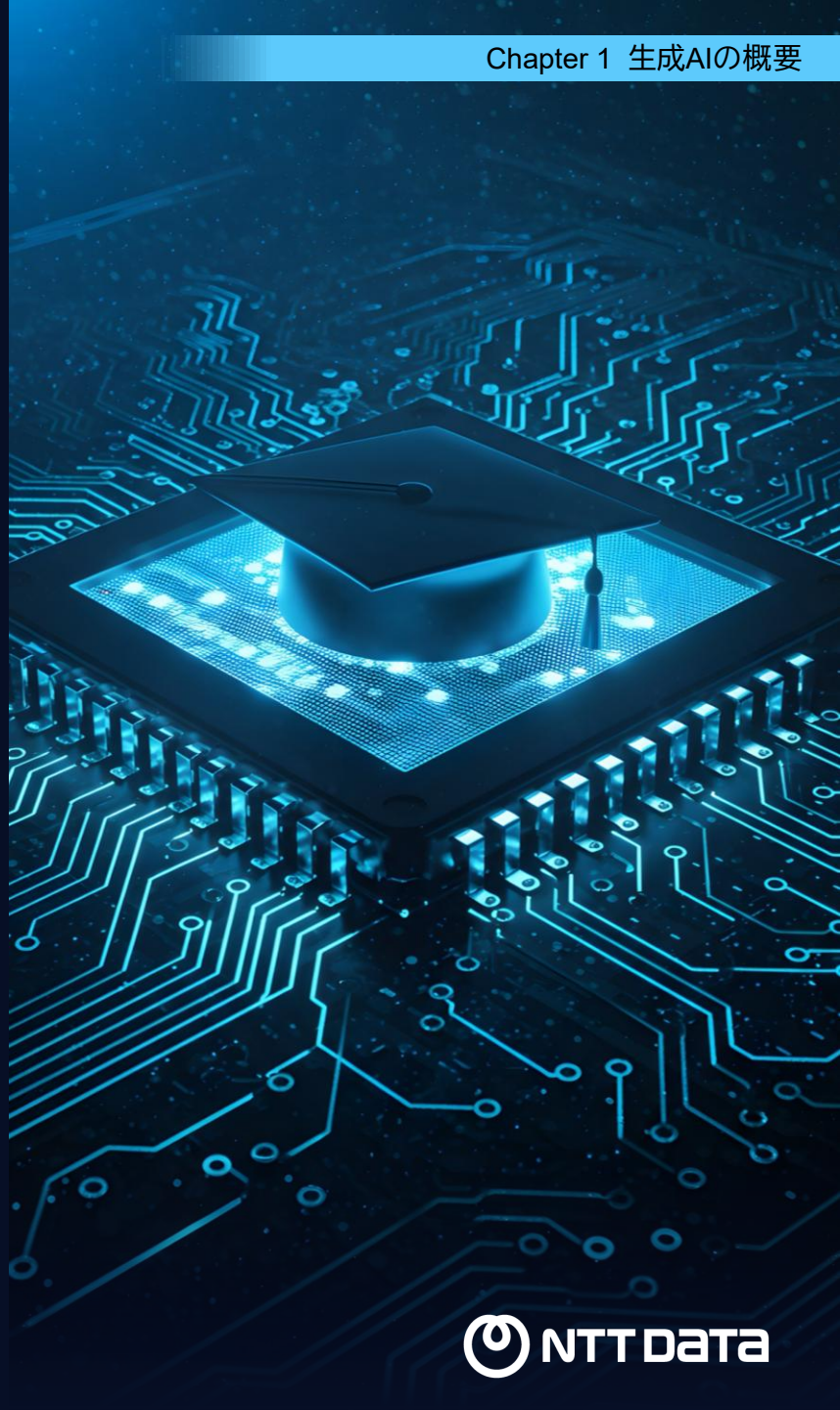
・法規制・・・各国・地域によるAIに関する法整備

・体制整備・・・組織内でのAI委員会設置、リスク評価、利用ポリシー・運用ルールの策定など

・技術対策・・・学習データのクレンジング、悪意ある生成指示の検出、出力のバイアス検出など

・監査・モニタリング・・・AIの利用状況や影響を継続的にチェックし、改善につなげる仕組み

これらの構成要素をバランスよく整備・運用することで、AIの利便性を最大限に活かしつつ、社会的な信頼を損なうことのない持続可能で責任あるAIの活用が可能になります。



1) Generative AI Market Size | Business Research Insights
2) AI Governance Market Size | Business Research Insights

Chapter 2

生成AIのトレンド

マクロ動向

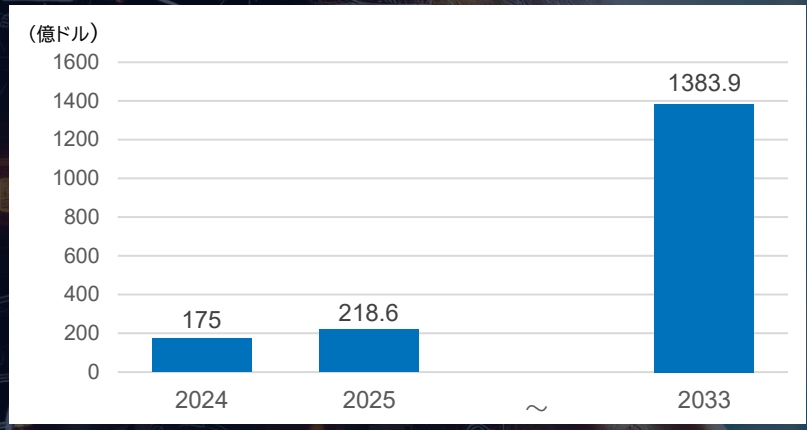
生成AIの産業横断的な急拡大に伴い、AIガバナンス市場も着実に拡大

生成AI市場は2033年に世界で約1,384億ドル[1)]、
AIガバナンス市場は2032年に約267億ドル[2)]の規模と試算

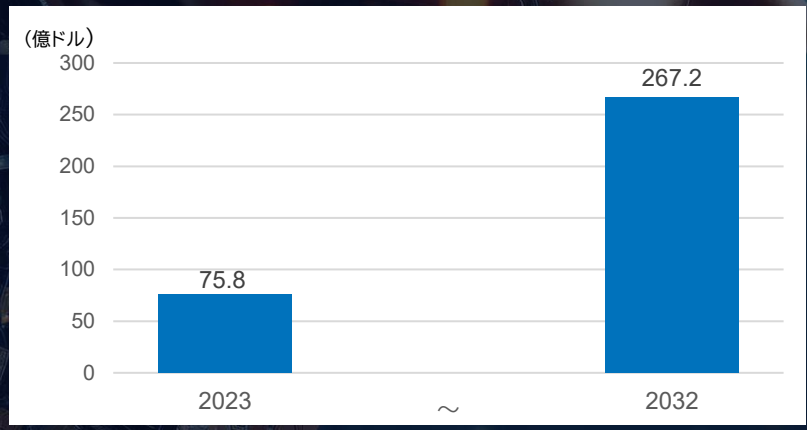
まず生成AI市場を見てみましょう。生成AI市場は、高い期待とともに急成長しており、活発な投資と技術革新が続いています。Business Research Insightsの調査[1)]によると、2024年の世界市場は約175億ドルと評価され、2033年には1,383億9,000万ドルに達する見込みです（CAGR 約24.9%）。

ではAIガバナンス市場はどうでしょうか。同社の調査[2)]によると、2032年には267億2,000万ドルに成長する見通しで（CAGR約15.8%）、その背景には各国の規制強化やAI倫理への関心の高まりがあることが考えられます。生成AIの急速な普及に伴い、誤情報の拡散やバイアス混入といったリスクが顕在化し、これに対応するAIガバナンスの需要も高まっています。

以上のように、生成AI市場の成長が新たなリスク対応のニーズを生み出し、それに応じてAIガバナンス市場も成長している構造が見て取れます。



［図2-1］生成AI市場見通し（世界）
グラフは、Business Research Insightsのデータ[1)]を基にNTT DATA作成



［図2-2］AIガバナンス市場見通し（世界）
グラフは、Business Research Insightsのデータ[2)]を基にNTT DATA作成

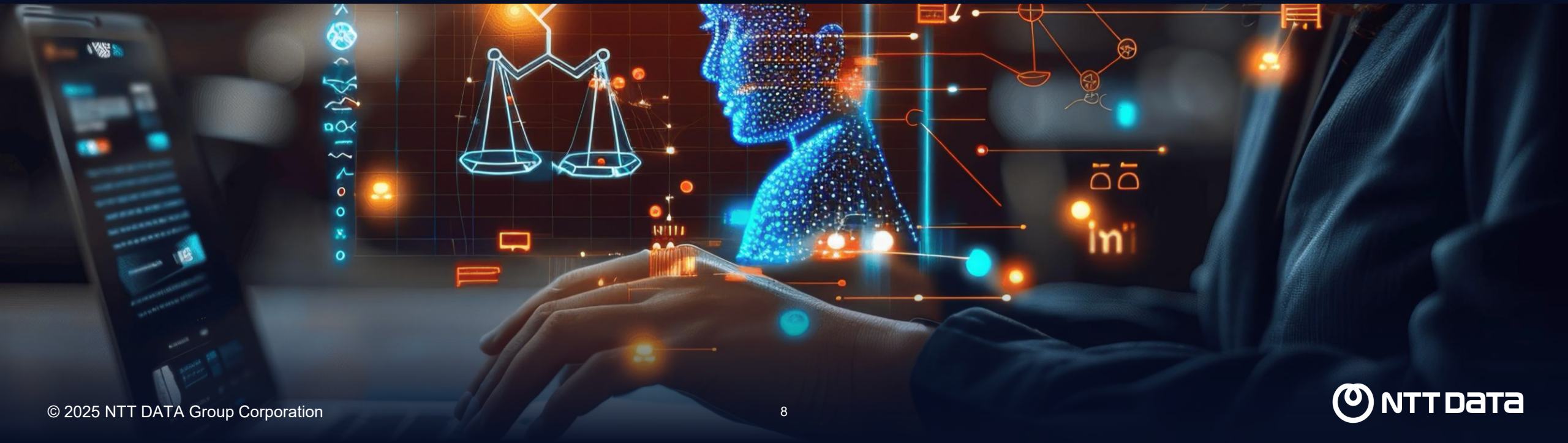
ユースケースの拡大
多様化する生成AIの活用と責任ある実装への課題

1) Generative AI Market Size | Business Research Insights

商業領域を起点に産業・個人へと活用領域が拡大中[1)]

生成AIの活用は商業、産業、個人の各領域で急速に広がっています。商業利用では、広告や商品説明、チャットボットなどを自動生成することで、顧客一人ひとりに合わせた情報発信や対応が可能となり、顧客体験の向上と業務効率の改善に貢献しています。特にeコマースやメディア業界では、多様なニーズに柔軟かつ大規模に対応できる仕組みとして注目されています。産業分野では、製品設計や製造プロセスの最適化、保守の予測といった場面で生成AIが導入されており、生産性の向上やコスト削減、開発スピードの加速に寄与しています。さらに個人利用の面でも、文章作成やデザイン、音楽の生成といった創作活動や、学習支援ツールとしての利用が広がり、専門知識がなくても高度なアウトプットを生み出せる環境が整いつつあります。

こうしたユースケースの広がりは生成AIの可能性を示す一方で、用途の多様化と高度化に伴い、リスクの管理や運用ガイドラインの整備といったAIガバナンスの重要性が一層高まっています。



1) 総務省｜令和6年版 情報通信白書｜生成AI
2) 総務省｜令和6年版 情報通信白書｜生成AI 関連データ

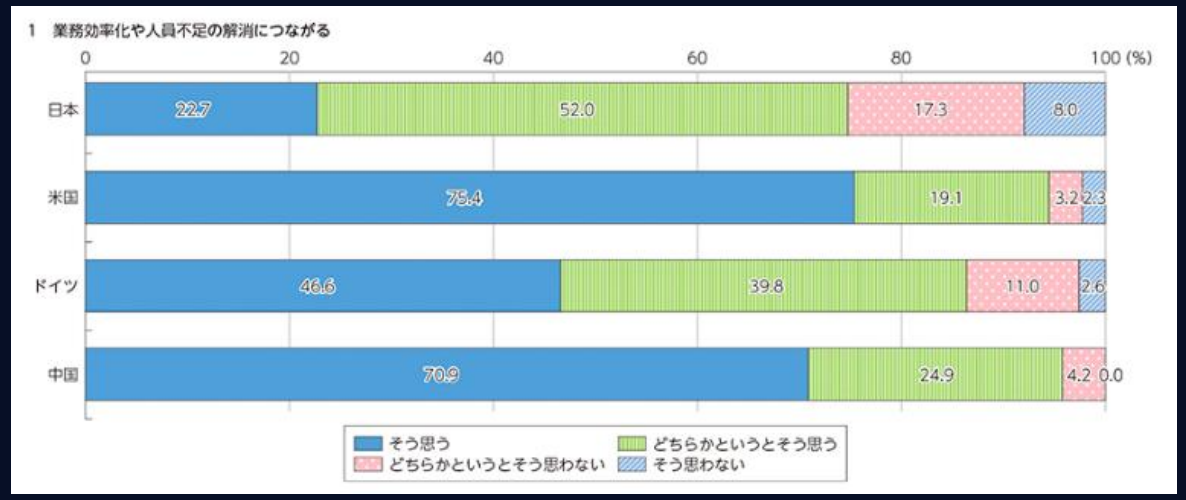
生成AIがもたらす業務改革への期待と倫理的な懸念

総務省が実施した、各国企業を対象とした業務における生成AIの活用状況に関する調査[1)]によると、多くの企業が生成AIに対して高い期待を寄せる一方で、その負の側面についても強い懸念を抱いていることが明らかになりました。

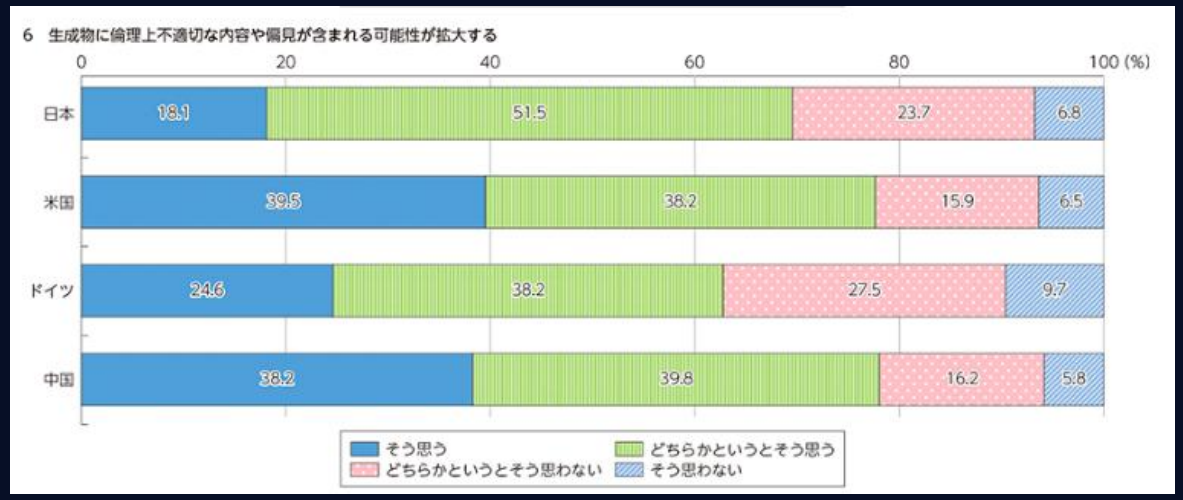
具体的には、「業務効率化や人手不足の解消につながる」「斬新なアイデア/新たなイノベーションがうまれる」といった問いに対し、「そう思う」「どちらかといえばそう思う」と回答した企業は、いずれも70～90%程度となっており、生成AIの可能性を高く評価している様子がうかがえます。

一方で、「著作権などの権利を侵害する可能性が拡大する」「生成物に倫理上不適切な内容や偏見が含まれる可能性が拡大する」といった問いに対しても、60～80%程度の企業が懸念を示しており、高いリスク意識を持っていることが分かります。

これらの結果から、生成AIは業務改革の有力な手段として注目される一方で、その活用には法的・倫理的リスクへの対策やガバナンス体制の整備が不可欠であることが分かります。実際に、多くの企業が利活用を進める中で、リスク管理や倫理的な配慮に向けた取り組みを始めており、バランスの取れた運用体制の構築が進められています。



［図2-3］生成AIに対する企業向けアンケート結果①[2)]



［図2-4］生成AIに対する企業向けアンケート結果②[2)]

1) 広島AIプロセス
2) フレンズグループ｜広島AIプロセス
3) AI事業者ガイドライン｜経済産業省
4) 人工知能関連技術の研究開発及び活用の推進に関する法律(AI法)の概要｜内閣府

生成AIを巡る社会課題と政策動向

生成AIの社会的・倫理的な課題への懸念の高まりに伴い、AI政策における国際的な枠組みづくりが進められています。

2023年にはG7広島サミットで「広島AIプロセス[1)]」が採択され、国際的なAIガバナンスの方向性が共有されました。同年12月には、G7首脳声明において「広島AIプロセス包括的政策枠組み」が承認されました。この枠組みでは「高度なAIシステムを開発する組織向けの広島プロセス国際行動規範」が示され、高度なAIシステムの導入前にはリスクを特定、評価、軽減するために適切な措置を講じることや、電子透かしなどを用いた信頼性あるコンテンツ認証・来歴追跡の仕組みを導入することなどが盛り込まれています。

AI政策に関する国際的な議論は、G7のみならず、他の国際協議の場においても広がりを見せています。とりわけ2024年5月に日本主導で立ち上げた「広島AIプロセス・フレンズグループ[2)]」では、国際的なAIガバナンスに関する多国間対話が進展しています。このグループは、安全、安心で信頼できるAIのグローバルな実現に向け、広島AIプロセスの精神に賛同する国々の自発的な枠組みです。G7諸国を超えて参加が広がり、50以上の国・地域からの賛同を得て活動が進められています。2025年2月に東京で開催された初の対面会合には各国政府に加え、有識者、国際機関、民間企業も参加し、各国の政策、AIの機会とリスク、グローバルガバナンスの在り方など活発に意見交換が行われました。また、マルチステークホルダーによる協力の必要性が確認されるとともに、AI開発者による報告枠組みへの参加を促すパートナーズコミュニティの立ち上げも公表され、「安全・安心で信頼できるAI」の実現に向けた具体的な取り組みが進められています。

前述のとおり、国際的な枠組みづくりが進む中で、各国においても独自の政策や規制の整備が進められています。以下では、各国の主な取り組みを紹介します。

日本

2024年に経済産業省と総務省が共同で「AI事業者ガイドライン[3)]」(第1.0版)を策定し、生成AI特有のリスク(ハルシネーション※1、情報漏洩、バイアスの再生産など)への対応策を提示しました。2025年3月には国際ルールとの整合を踏まえた第1.1版も公表され、AI開発・提供・利用の各段階におけるリスク対応の共通指針が示されています。また同年、「人工知能関連技術の研究開発及び活用の推進に関する法律(AI法)[4)]」が成立しました。この法律は、日本のAI開発・活用は遅れており、多くの国民がAIに対して不安を感じているという現状を背景に制定されました。イノベーションの促進と、AI活用に伴うリスクへの対応を両立するため、新たな法律が必要だという考えに基づいています。基本的な施策として、研究開発の推進や国際的な規範策定への参画、事業者などへの指導・助言・情報提供などが示されています。これは、AI活用を規制するものではなく、国際指針に則り、イノベーション促進とリスク対応を両立し、最もAIを開発・活用しやすい国をめざすことを目的としています。

※1 ハルシネーション
AIが実際には存在しない情報や誤った内容を、あたかも正しい事実であるかのように生成してしまう現象

アメリカ

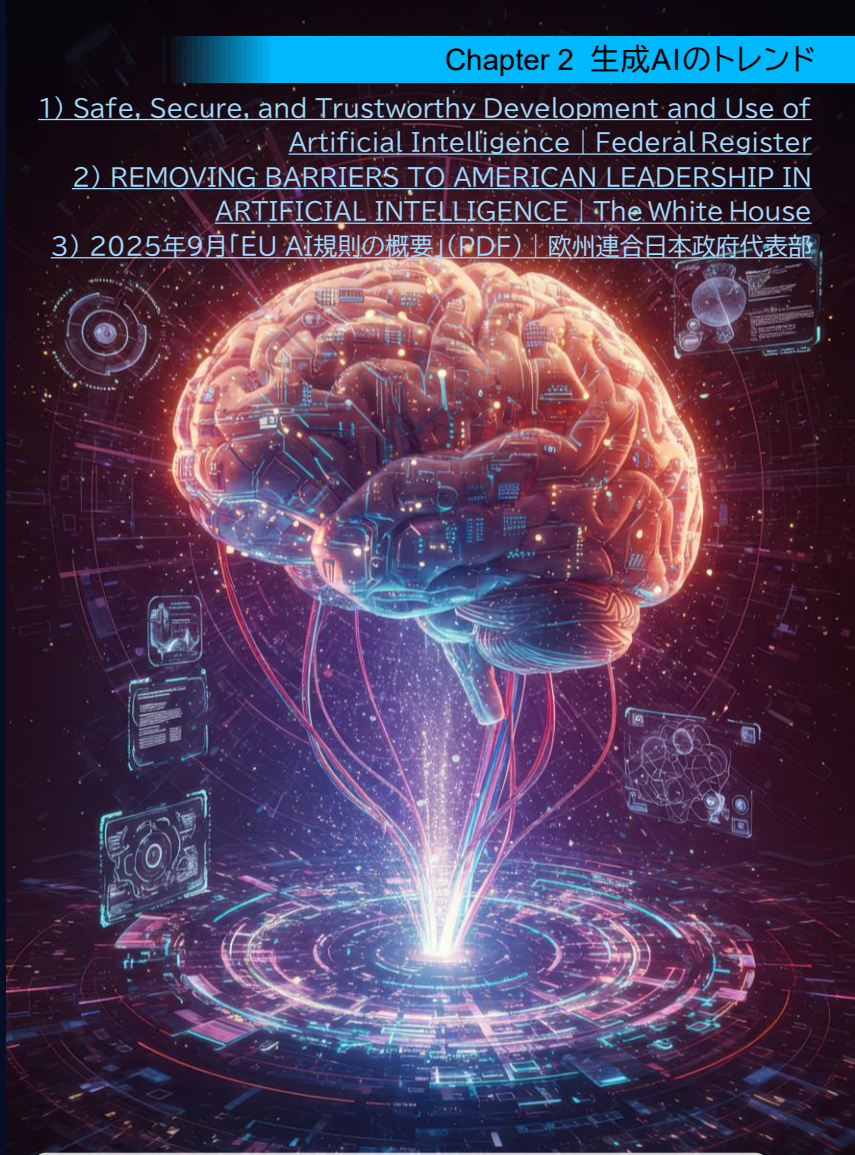
2023年10月に、AIの安心・安全で信頼できる開発と利用に関する大統領令[1)]が発令され、AIの安全性の確保や公平性と公民権の推進、プライバシーの保護などに関する指針が示されました。しかし、2025年にトランプ政権が発足すると、こうした方針に大きな転換が見られました。トランプ政権はAIに対する規制緩和を指示する大統領令[2)]を発表し、前政権下で導入されたAIの安全性や倫理性に関する政策の見直しが進められました。イノベーションの促進と経済競争力の強化を重視し、民間主導でのAI開発を推進する姿勢が強調されています。

EU

世界に先駆けて生成AIを含む包括的な法制度である「AI規制法[3)]」を策定しました。2021年4月に法案が公開され、2024年5月に成立、8月に発効されました。本規則はAIシステム全般を規制対象とし、その中でも汎用AI※1モデルについては特別な義務が課されており、2025年以降、数年間にわたり段階的に施行される予定です。たとえば、汎用AIモデルを開発・提供する事業者に対しては、透明性の確保、学習に使用したコンテンツの概要公開、AI生成コンテンツであることを明示する措置などが義務付けられています。特に高い影響力を有する汎用AIモデルについては、一層厳格な評価義務やリスク管理体制の構築が求められています。また、違反した場合には高額な罰金が科されることも特徴です。

このように、各国は生成AIの急速な発展に対応すべく、独自の法制度やガイドラインを整備しています。今後、企業や開発者はこれらの動向を的確に捉え、国際的なルール形成に対応しながら、責任あるAI活用を進めていくことが必要です。

1) Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence | Federal Register
2) REMOVING BARRIERS TO AMERICAN LEADERSHIP IN ARTIFICIAL INTELLIGENCE | The White House
3) 2025年9月「EU AI規則の概要」(PDF) | 欧州連合日本政府代表部



※1 汎用AI（General Purpose AI）
広範なタスクを学習・実行可能で他のAIシステムに統合可能なAI
生成AIもこれに含まれる

AIガバナンスに関連する製品・サービス

これまで述べてきた通り、生成AIの急速な導入拡大に伴う、倫理的な課題や誤用リスクの顕在化、法規制の強化、そして社会的責任への関心の高まりが加速しています。こうした背景から、AIガバナンスに関連する製品・サービスの需要も拡大しています。以下で主要なソリューションの概要を一部紹介します。

［表2-1］AIガバナンス関連ソリューション

項 目	概 要
AIモデルのライフサイクル管理	AIモデルの設計から運用、監視、更新、廃棄までの全ライフサイクルを一貫管理する。説明性や公平性のチェック、データドリフト※1監視などAIガバナンス機能を備え、責任ある運用を実現する。
AIリスク評価・コンプライアンス対応支援	AIシステムのリスクを多角的に評価し、各国・地域の規制や業界ガイドラインへの準拠を支援する。使用用途の分類、影響度評価、リスク低減策の策定に加え、AI規制法などの適合性チェック、コンプライアンス文書の整備、監査対応支援を含む。これにより、法的・倫理的な要求に即したAIの導入と運用を実現する。
バイアス検知と説明可能なAI（XAI）	AIモデルの出力に対する根拠や偏りを可視化し透明性と公平性を確保するツール。特徴量の影響度や入力変化による結果の変動を分析し意思決定の仕組みを明らかにする。意図しない差別やバイアスの早期発見・是正を可能にし、説明責任とAIシステムの信頼性向上に寄与する。
AIガバナンス・コンサルティングサービス	AIの倫理的かつ持続可能な活用を目的に、方針策定から運用体制構築、教育・研修、人材育成、規制対応支援まで一貫して支援するサービス。業種や成熟度に応じたカスタマイズが可能で、社内ガイドライン策定、リスクアセスメント、説明責任確立など重要課題への対応を支援する。

Chapter 3、4では、AIガバナンスの概念をより具体的に捉えるため、実際の製品・サービスの活用や取り組み事例を紹介します。

※1 データドリフト
モデルの学習時のデータ分布と、実運用時の入力データ分布が変化する現象

1) Azure AI Content Safety – AI コンテンツ モデレーション | Microsoft Azure
2) EdChat: Pioneering Safe Generative AI in South Australian Classrooms with Azure AI Content Safety

Chapter 3

生成AI×サステナビリティ

サービス・取り組み 他社事例（AIガバナンス関連）

Azure AI Content Safety
～生成AIの安全性を守る高度なガードレール～[1)]

【1. 事例概要】

企業名：Microsoft

地域：海外（アメリカ）

【2. 背景・目的】

生成AIの急速な普及に伴い、有害または不適切なコンテンツの生成といった社会的なリスクが顕在化しています。このようなリスクを抑制し安心して生成AIを活用するために、 Azure AI Content Safety が開発されました。

Azure AI Content Safety は、暴力、憎悪表現、性的内容、自傷行為の有害な入力と出力コンテンツを検出し、ブロックするツールサービスです。複数の有害カテゴリによる分類と、重大度レベルに基づくしきい値設定により、具体的なユースケースに適した責任あるAIポリシーの遵守を支援します。また、カスタムカテゴリを使用した独自のコンテンツフィルターの作成も可能です。

さらに、Azure OpenAI などの他の Azure AI 製品と組み合わせて利用することで、責任あるAIツールを組み込んだ包括的なソリューションを開発することができます。こうした技術基盤は Microsoft が推進する「責任あるAI」の理念に則って設計されており、透明性やプライバシー保護にも十分配慮されています。

【3. 事例詳細】[2)]

以下で、Azure AI Content Safety の活用事例を紹介します。

（1）内容

南オーストラリア州教育省は、 Microsoft と連携して生成AI搭載の教育チャットボット「 EdChat 」を開発・導入しました。この取り組みの目的は、生成AIを活用して教育の質を高め、生徒が将来必要とするスキルや知識を身につけられるよう支援することにありました。導入にあたっては、有害または不適切なコンテンツを検出・ブロックできる Azure AI Content Safety を EdChat に組み込み、安全性を重視した仕組みを整備しました。

2023年に実施された8週間のトライアルには、8つの中等学校から約1,500人の生徒と150人の教師が参加し、教育現場における生成AI活用の可能性が検証されました。

（2）効果

EdChat は、「生徒を有害なコンテンツにさらすことなく学習を支援できる」と評価され、教育現場での使用が安全で適切であることが確認されました。Azure AI Content Safety による安全対策機能により、教師は過度な監視の負担を負うことなく教育活動に集中できる環境が実現しました。教育省の担当者は、 EdChat の成功には安全性の確保が欠かせず、本サービスが「導入時から欠かせない要素であった」と述べています。

こうした成果を踏まえ現在では多くの学校で導入が進んでおり、 EdChat は教育の質向上と生徒の保護を両立させた事例となっています。

Amazon Bedrock Guardrails[1)]
～アプリケーション要件と責任あるAIポリシーに合わせてカスタマイズされた保護手段を実装～

【1. 事例概要】

企業名：Amazon Web Services（AWS）

地域：海外（アメリカ）

【2. 背景・目的】[2)]

生成AIモデルは幅広いトピックに関する情報を生成できますが、その応用には関連性の維持、有害なコンテンツの回避、個人を特定できる情報などの機密情報の保護、ハルシネーション（幻覚）の軽減などの課題があります。AWSの生成AIサービスAmazon Bedrockの基盤モデルには組み込みの保護機能がありますが、これらはモデル固有であることが多く、組織のユースケースや責任あるAIの原則に完全に合致しない可能性がありました。そこで、AWSはAmazon Bedrock Guardrails（以下、「Guardrails」）の提供を行っています。これはセーフガードの導入、有害なコンテンツの防止、主要な安全性基準に対するモデルの評価を支援するものです。Guardrailsを使用すると、ユースケースと責任あるAIポリシーに合わせてカスタマイズされた生成AIアプリケーションのセーフガードを実装できます。複数のユースケースに合わせて複数のガードレールを作成し、それらを複数の基盤モデルに適用することで、ユーザーエクスペリエンスを向上させ、生成AIアプリケーション全体で安全性のコントロールを標準化できます。

- ✓自動推論によりAIのハルシネーション（幻覚）を最小限に抑え、最大99％の精度で正しいモデル応答を特定
- ✓業界をリードするテキストおよび画像コンテンツのセーフガードにより、お客さまが最大88％の有害マルチモーダルコンテンツをブロックできるように支援

1) 生成AIデータガバナンス – Amazon Bedrockのガードレール – AWS
2) Amazon Bedrock Guardrailsを使用したモデルに依存しない安全対策を実装する | Amazon Web Services
3) Amazon Bedrockのガードレールが、新しい機能により、生成AIアプリケーションの安全性を強化 | Amazon Web Services
4) 株式会社BTM様のAWS生成AI活用事例 | Amazon Web Services

【3. 事例詳細】

以下でGuardrailsの活用事例を紹介します。

【Grab（シンガポール）】[3)]

（1）内容

シンガポールの多国籍タクシーサービスであるGrabは、生成AIアプリケーションの安全な利用を実現するため、Guardrailsを活用しています。

（2）効果

Grab社内でのベンチマーキングの結果、Guardrailsは他のソリューションと比較してクラス最高レベルのパフォーマンスを発揮しました。Guardrailsを利用することで、Grabは責任あるAIの実践に沿った堅実なセーフティ機能を備えることができます。特にガードレール機能は、AIを活用したGrabのアプリケーションに対する新たな攻撃からGrabとお客さまを保護する役割を果たしています。これにより、データプライバシーを守りながら、多様な市場におけるアプリケーションの安全な運用を実現しています。

【株式会社BTM（日本）】[4)]

（1）内容

DX推進事業やITエンジニアリングサービスを展開するBTMは、システム調査の効率化を目的に、Amazon BedrockとStrands Agentsを活用して自動で調査する仕組みを構築しました。その際、日本語に対応したGuardrailsを活用し、関係のない問い合わせをブロックする機能も導入しました。

（2）効果

本取り組みにより、システム調査にかかる時間は従来の半日から最短10分に短縮されました。定性的な改善としては、日本語対応したGuardrailsの利用により、無関係な問い合わせを効果的にブロックできるようになりました。これらにより、エンジニアはより本質的な開発業務に集中できるようになりました。

Chapter 4

生成AI×サステナビリティ NTT DATAの取り組み

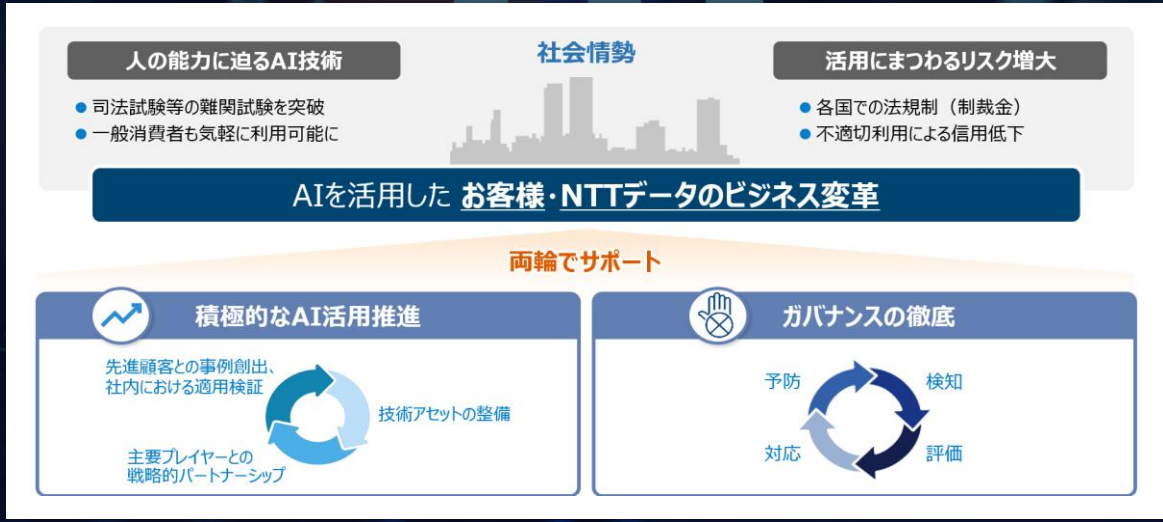
生成AIへの取り組み[1)]

NTT DATAでは、生成AI活用によるビジネス変革を推進しています。

この変革を実現するために、「積極的なAI活用の推進」とAI活用におけるリスクへの対応として「ガバナンスの徹底」に、両輪で取り組んでいます。

「積極的なAI活用の推進」では、生成AI技術を最大限に活用し、ビジネスや社会における価値を創出します。その一環としてたとえば、LITRON®[2)]やtsuzumi[3)]といった生成AIソリューション・プラットフォームの提供、OpenAIとのグローバルでの戦略的提携[4)]、グループ約20万人の全社員を対象とした生成AIの人財育成フレームワークの開発・展開[5)]など、多面的な取り組みを進めています。

「ガバナンスの徹底」では、生成AI活用における透明性や公平性、安全性を確保するために、検知、評価、対応、予防の4つのプロセスを継続的に実施しています。



［図4-1］NTT DATAの生成AIへの取り組み[1)]

1) 生成AI(Generative AI)｜NTT DATA
2) LITRON®｜NTT DATA
3) tsuzumi｜NTT DATA
4) OpenAIとのグローバルでの戦略的提携を開始｜NTT DATA
5) グローバル約20万の社員を対象とした生成AIの人財育成フレームワークを整備｜NTT DATA

AIガバナンス構築の全体像

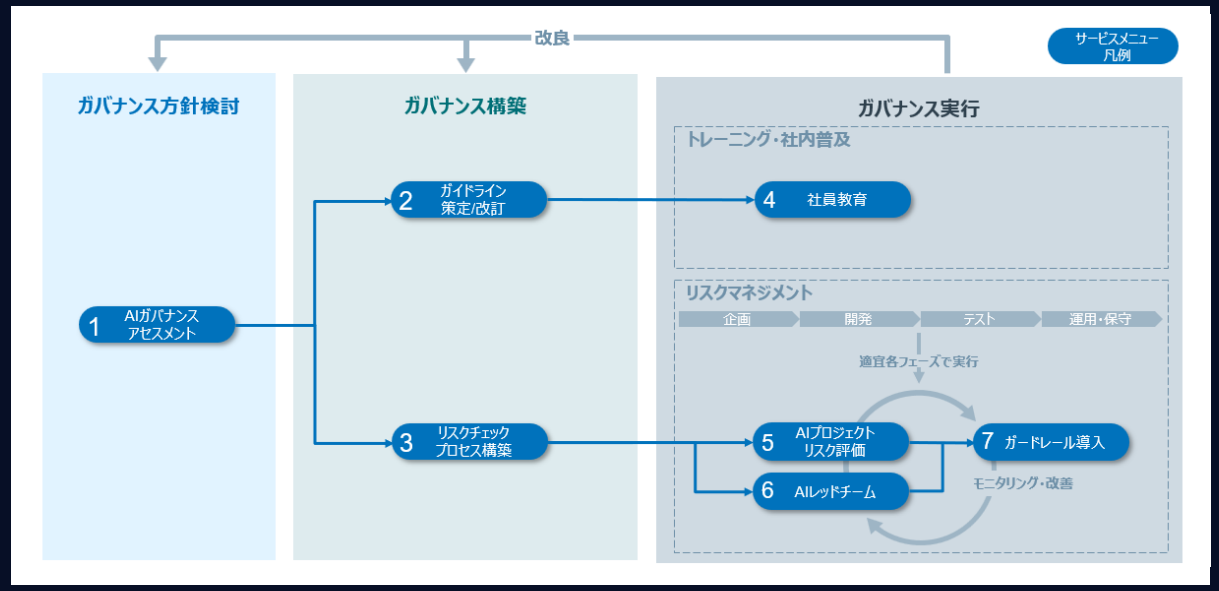
～コンサルティングサービスを通じて安心・安全なAI活用を支援～

1) AIリスク診断から対策実行・運用まで支援する「AIガバナンスコンサルティングサービス」を提供開始 | NTT DATA

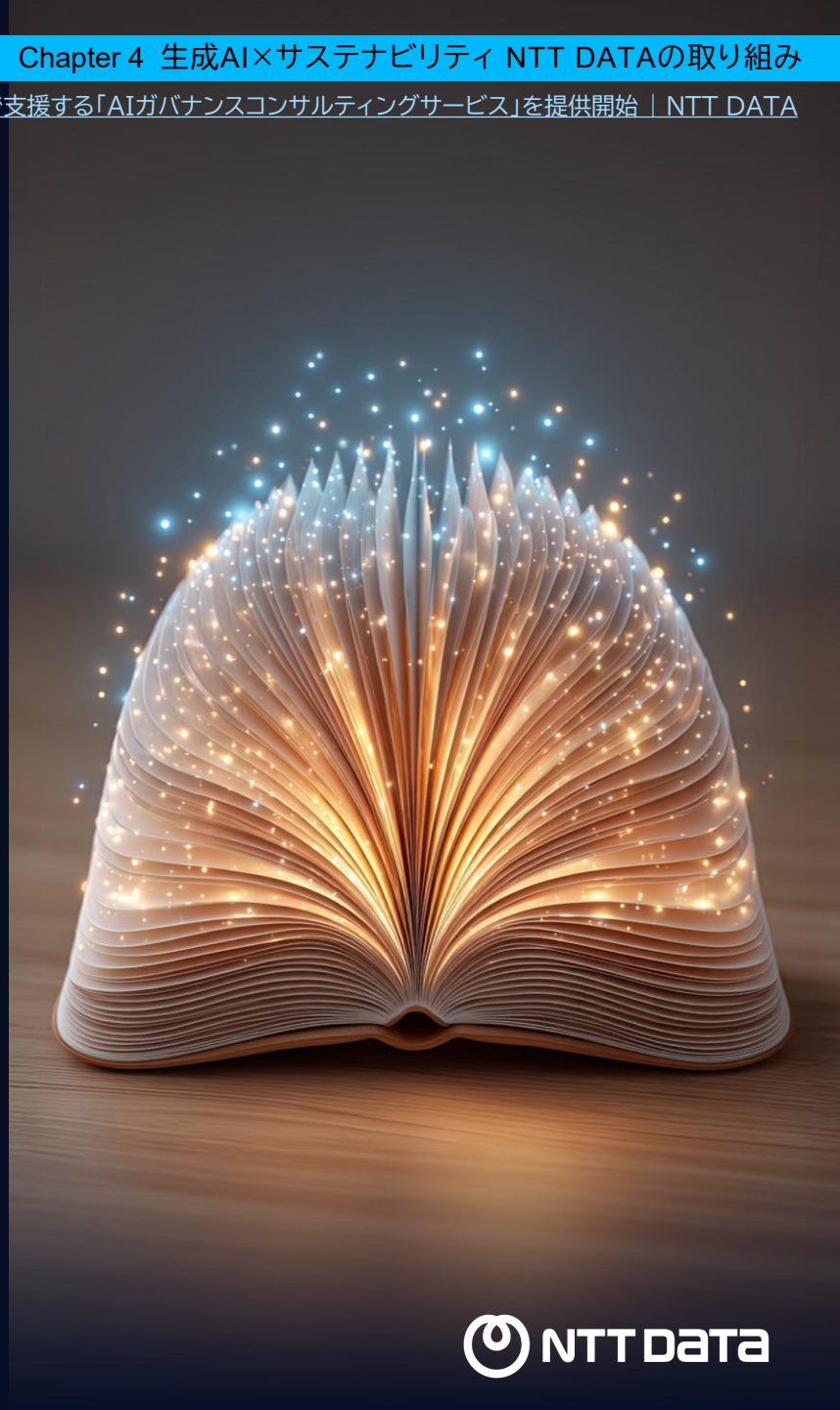
生成AI活用に伴うリスクの可視化と対応の実践

それでは、生成AIを組織で活用する際に求められるガバナンスの全体像と、NTT DATAが提供するAIガバナンスコンサルティングサービス[1)]のアプローチを見ていきましょう。

生成AIのリスクに対応するには、技術面だけでなく、組織のルールや運用面からの取り組みも必要となります。NTT DATAが提供するAIガバナンスコンサルティングサービスは以下の7つのステップを軸に構成されており、リスクの検知から実行、運用定着に至るまでを包括的にサポートします。この仕組みは、AI導入の初期段階（企画・設計）から運用・改善フェーズまで、プロジェクト全体にわたって活用可能です。



［図4-2］NTT DATAのAIガバナンスコンサルティングサービス　全体像



①AIガバナンスアセスメント

まず対象組織のAIガバナンス状況を評価し、現在のAIリスク管理体制の弱点やリスクを明示します。その上で強化が必要な領域を特定し、優先度の高いタスクを整理して何から取り組むべきかを具体的なアクションプランとして提示します。

②ガイドライン策定・改訂

次に最新の国内外のAI規制や技術の進化を踏まえ、AIリスクやあるべき姿を業務の観点から定義し、組織のルールを策定します。その上で、各ガバナンス観点で実行すべき行動を整理し、AIリスクと対応方針を明文化します。既存の社内ガイドラインがある場合は国内外の動向との整合を図り、必要に応じて改訂します。

③リスクチェックプロセス構築

すべてのAI・データ活用案件で同一基準によるリスク評価を行えるようにし、判断の一貫性を確保します。また、AIリスク確認プロセスを確立し、未確認リスクを見逃さない仕組みを構築します。さらに、プロジェクトが増えても標準化されたプロセスにより、効率的かつ適切なリスク管理を実現します。加えて、現場でもリスクを理解し、日常的に意識する環境を整えます。

④社員教育

制度やルールの定着に向け、勉強会や教育コンテンツを通じてAIガバナンスの基本概念と重要性を社員に教育するとともに、社員が日常業務でAIリスクを意識して行動できるよう促します。最新の法規制や業界ガイドラインも反映し、組織に合った効果的な教育体制を構築します。

⑤AIプロジェクトリスク評価

NTT DATAのノウハウを活用し、AI活用案件の公平性や安全性が十分に担保されているかを評価します。これにより、リスクの存在や重大性を可視化し、倫理的なリスクの影響範囲も明確にします。加えて、評価結果を踏まえた具体的なリスク低減策を提案します。

⑥AIレッドチームによる専門的な支援

AIレッドチームでは最新の技術や法制度の動向を踏まえ、品質・セキュリティ・倫理の観点からAIリスクを評価します。自動評価指標に加え、ユースケースに応じたカスタム指標を用いて客観的に評価します。大量データは自動、少量データは人手で評価し、効率的かつ正確なモデル評価を実施します。また、レポートに基づいた今後の対策を助言します。

⑦ガードレール導入

ガードレールツールを導入することで、AIモデルへの入力と出力を監査し、不適切な内容（例：公序良俗に反する要求や差別的な表現など）に対して、ユーザーに応答を返す前に即時対策を実行します。

1）りそなホールディングスと共創する、AIガバナンス構築の取り組み｜DATA INSIGHT

【ケーススタディ】
りそなホールディングスにおけるAIガバナンス構築の取り組み[1)]

以下では実際の取り組み事例として、株式会社りそなホールディングス（以下、りそなHD）との協業によるガバナンス構築の実践内容を取り上げます。

✓ 背景と目的

りそなHDでは、生成AIを含む先進技術の活用を通じて、顧客サービスの高度化や業務変革を推進しています。一方で、AI活用に伴うリスクを的確に把握し、コントロールする体制の整備が課題となっていました。こうした背景から、りそなHDの業務特性に即したAIガバナンス構築に向けて、NTT DATAと連携して取り組みが開始されました。

✓ ガイドラインの策定（前頁②）

まず、組織内で共通言語として機能し、実務で参照されやすいAIガイドラインの策定を実施しました。策定にあたっては、既存の社内規定との整合性や実務で使用されている用語表現の確認を行い、実務での活用を前提とした構成と情報量を検討しました。また、AIガバナンスを取り巻く技術・社会動向も適宜反映しています。

✓ リスクチェックプロセスの構築（前頁③）

AI関連プロジェクトごとのリスク評価のばらつきを防ぐため、案件横断で適用可能な標準的リスクチェックプロセスを整備しました。これは、NTT DATAが蓄積してきた知見と国や業界団体のガイドライン・規制動向をもとに設計されており、実務上の判断がしやすい項目構成としています。トライアル導入を通じて現場からのフィードバックを取り入れ、継続的な改善も実施しました。

✓ 教育コンテンツの整備（前頁④）

AIリスクへの感度を高めるため、基礎教育コンテンツを整備しました。これは、制度対応に先立ち、AIリスクに初めて触れる従業員の“入り口”として機能することを意図しています。対象は開発者に限らず一般従業員も含み、平易な表現と短時間で理解できる構成になっています。業務に即した具体例も盛り込み、自律的な判断を促すことでガバナンス実効性の向上を図ります。

✓ 今後の展望

本取り組みでは段階的に進めながら、現場に根ざした実効性のあるガバナンス体制を整備することができました。今後も、変化する技術や制度に対応しながら、体制の維持・強化が期待されます。NTT DATAは引き続き、戦略的パートナーとして支援に努めてまいります。

生成AIのリスクを可視化する「AIレッドチーム」[1)]

1) 生成AIのリスクを可視化する「AIレッドチーム」| DATA INSIGHT

ここからは、前述したAIガバナンス構築の全体像のうち、「⑥AIレッドチームによる専門的な支援」の実施内容を詳しく見ていきます。

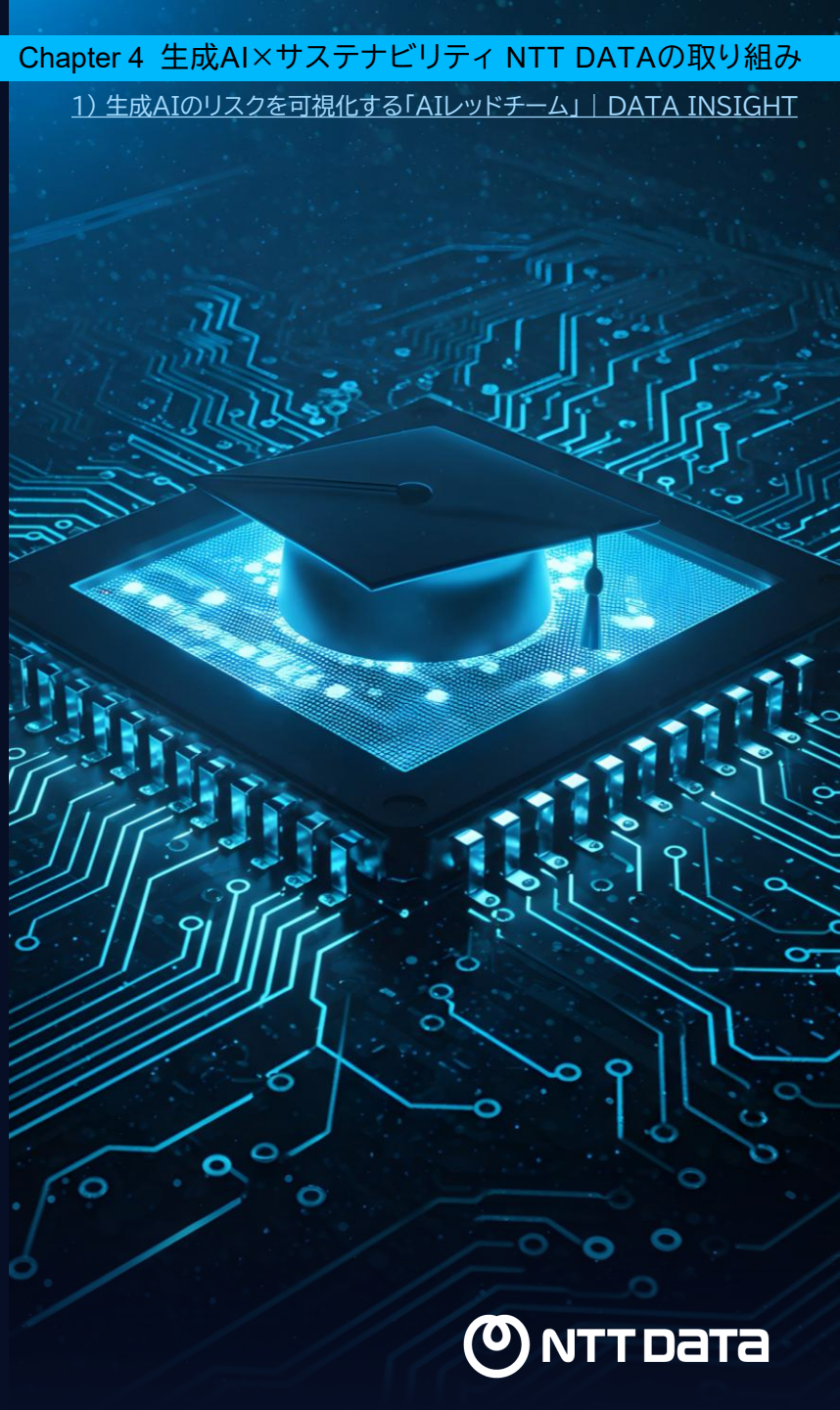
近年、多くの企業において生成AIの活用が広がり、業務支援から消費者向けサービスまで社会実装が加速しています。その一方で、意図しない情報漏えいや偏見を含んだ出力など、ガイドラインやルールだけでは捉えきれないリスクも顕在化しています。こうした状況を踏まえ、生成AI活用に伴うリスクを利用者視点で発見し、評価する手法として注目されているのがAIレッドチームです。ここでは、NTT DATAが提供するAIレッドチームサービスの必要性と特徴を紹介します。

なぜ、いま「AIレッドチーム」が求められているのか？

生成AIは、いまや業務の高度化や効率化にとどまらず、一般消費者向けのサービスにも広く組み込まれ、企業や行政機関の業務プロセスやサービス提供のあり方を大きく変えつつあります。しかし、その利便性の裏側には、「意図しない機密情報の漏えい」や「特定属性に対する偏見を含んだ出力」など、事前に想定しづらいリスクが潜んでいます。

こうした問題に対し、多くの組織ではガイドラインやポリシーを整備することで、生成AIの安全な活用を目指しています。しかし、ルールや制度だけではリスクを完全に防ぎきることは困難です。なぜなら、生成AIは入力の自由度が高く、使われ方が無数に存在するためです。

こうした背景から注目されているのが「AIレッドチーム」という手法です。レッドチームはもともとサイバーセキュリティの分野で発展したアプローチで、攻撃者の視点でシステムやネットワークに模擬攻撃を行い、弱点をあぶりだすものです。この考え方を生成AIに応用したものがAIレッドチームです。生成AIに対してあえて「意地悪な入力」を与えたりルールの隙間を突く問いを試したりすることで、モデルやシステムがはらむリスクを発見し、評価します。

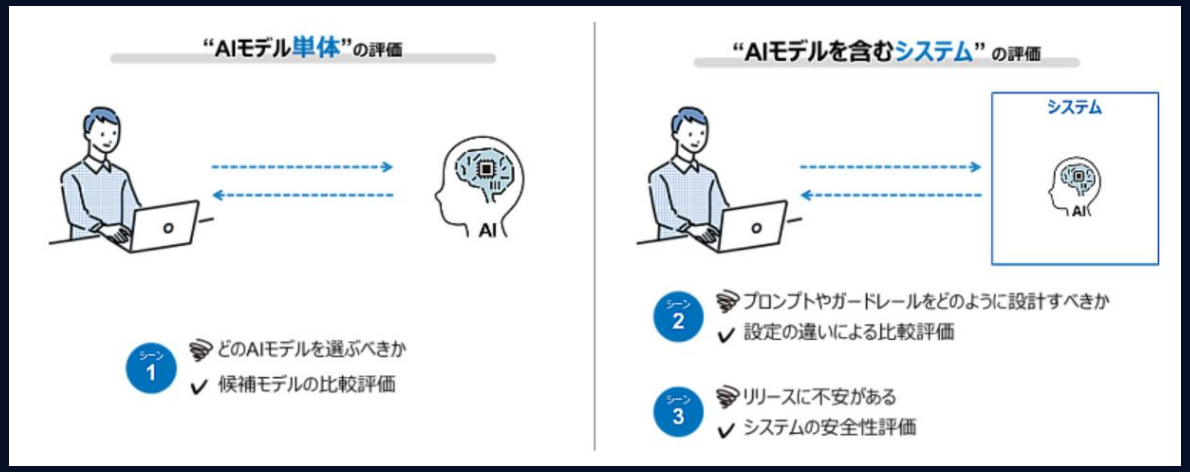


AIレッドチームサービスの主な利用シーンと実施の流れ

前段で述べたように、ルールだけでは捉えきれないリスクを可視化し、対応するにはAIレッドチームが有効です。NTT DATAでは、AIシステムの開発や提供といったお客さまの活用シーンに応じて多様な支援サービスを展開しており、その一環としてAIレッドチームサービスを提供しています。ここでは、AIレッドチームサービスの利用シーンと流れをご紹介します。

✓ 主な利用シーン

AIレッドチームサービスは、「AIモデル単体の評価」と「AIモデルを含むシステムの評価」に大別できます。以下では主な利用シーンの3例をご紹介します。



［図4-3］AIレッドチームの利用イメージ

シーン①「生成AIを業務に活用したいが、どのAIモデルを選ぶべきか判断に迷っている」

あまたの生成AIモデルが存在する中で、性能やリスクなどをどのように比較するかは大きな課題です。たとえば、生成速度は優れているが偏見を生むリスクがあるモデルを選んでしまうと、後でリスクが顕在化してしまう恐れがあります。AIレッドチームでは、あらかじめ想定される利用シーンをベースに、各モデルの振る舞いを評価観点ごとに検証します。これにより、自社のユースケースを最も安全に実現できるモデルの選定を支援します。

シーン②「システムプロンプトやガードレールの設計に不安がある」

システムプロンプトやガードレールの設計は、システムの振る舞いを大きく左右するにもかかわらず、明確な正解がない難しい領域です。AIレッドチームは、システムプロンプトやガードレールの設定を変えた状態でのシステムの振る舞いを評価し、よりリスクを抑えた設計を支援します。

シーン③「リリース直前、想定外の使われ方が心配」

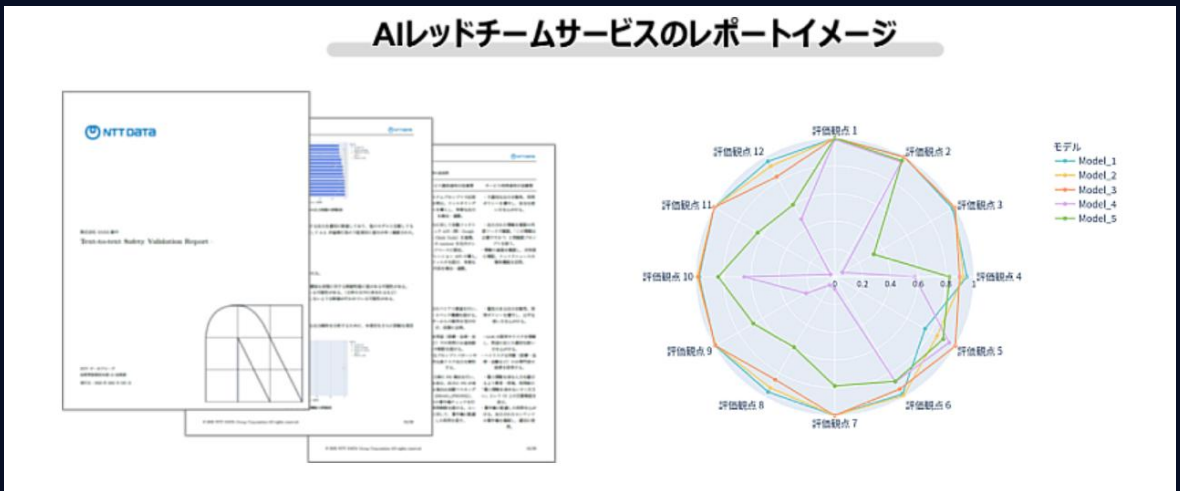
全社公開や外部公開を見据えたサービスでは、悪意のあるユーザーからシステムを守ることや、システムが孕む危険性からユーザーを守ることが特に重要です。AIレッドチームは、利用者の視点からシステムの振る舞いやリスクの傾向を検証します。これにより、AIシステム特有のリスクを加味したうえで安全性を確認し、リリースの判断を支援します。

さらに、スポット的な対応にとどまらず、ユースケースやリスクの許容度に応じて、体制の構築から運用までを視野に入れた継続的な支援も行います。

✓ サービスの流れ

AIレッドチームサービスは、以下の3つのステップを通じて、リスクの発見から改善提案までを一貫して提供します。

［図4-4］AIレッドチームサービスの流れ



［図4-5］AIレッドチームサービスのレポートイメージ

ステップ1：環境準備

検証の対象となるモデルやシステムに合わせてツールの調整を行い、試験の実行環境を整えます。お客さまのセキュリティポリシーやシステム要件に応じて、以下いずれかの方式を選択します。

・お客さま環境へのツールのセットアップ
・当社環境へのモデル持ち込み

ステップ2：AIレッドチーム試験

実際の運用上のリスクを想定した試験を実施します。後述する評価観点にもとづく、実践的な視点での検証が行われます。

ステップ3：評価レポート・改善提案

テスト結果をレポート形式で提示し、具体的なフィードバックを改善提案とともに提供します。レポートには、評価観点（例：攻撃的表現・差別表現・偏見など）ごとのリスクの評価結果、リスクの重大度、推奨対策案などが含まれます。単なる「問題の指摘」にとどまらず、「どう防ぐか」に踏み込んだ提言を行うことで、具体的な改善に向けた支援が可能です。

NTT DATAのAIレッドチームサービスの提供価値

NTT DATAでは、AIレッドチームによる支援にあたり、3つの観点を重視した価値提供を行っています。

①日本語特有のリスクへの対応

多言語対応をうたう生成AIモデルであっても、日本語で利用した際の挙動には想定外の振る舞いが生じることがあります。NTT DATAでは、AISI（Japan AI Safety Institute：日本におけるAI安全性の中心的機関）のガイドラインを踏まえた日本語のデータセットを設計しているため、日本語特有のリスクも検証できます。

②複数のモデルやシステム設計の違いを比較した最適な構成の提言

生成AIのリスクは「ある／ない」といった単純な判断では捉えきれず、業務ごとの特性やリスクの許容度を踏まえた判断が必要です。NTT DATAでは、モデルやシステム設計の違いによるリスクの傾向を可視化し比較することで、相対的な強み・弱みを明らかにします。これにより、リスクと要件のバランスに配慮した構成を選定できます。

③リスクの評価に基づく具体的な改善策の提案

AIレッドチームによる評価結果は、単なるリスクの列挙ではなく、「どうすれば防げるのか」という観点からの改善提案へとつなげます。たとえば、リスクの評価に対して考えうる対応策を検討し提示したうえで、必要に応じて再設計まで伴走します。このように、評価から改善までを一貫して支援できることがNTT DATAのAIレッドチームの強みです。

生成AIの社会的な活用が進む中、AIの安全性確保は、企業の信頼性を左右する経営課題になりつつあります。ここで紹介したように、NTT DATAでは、ルールで掌握しきれない運用上の抜け漏れや予期せぬ振る舞いといったリスクをAIレッドチームで洗い出し、それらを改善するための支援を行っています。NTT DATAは、現場と経営をつなぐパートナーとして、お客さまの状況や目標に即した現実的かつ持続可能なアプローチを通じ、安心・安全なAI活用の実現に取り組んでいます。

これらの取り組みを通じて、健全なAI活用による価値創造と持続可能な社会の発展に貢献してまいります。

Chapter 5

おわりに

本ホワイトペーパーでは、生成AIの進化と社会的な影響、そしてそれに伴うAIガバナンスの重要性について解説してきました。生成AIは、私たちの生活や働き方を大きく変える可能性を持ち、教育・医療・産業など幅広い分野での活用が進んでいます。今後さらに多様な領域で、その影響力は拡大していくと考えられます。

近年では、教育格差の是正や環境問題への貢献など、様々な分野での生成AI活用も進んでいます。しかし、ガバナンスの視点を欠いた技術導入は、期待した成果を得られないばかりか、社会的な信頼の喪失やブランド毀損といったリスクを招きかねません。

本ホワイトペーパーで紹介した他社事例や、NTT DATAにおけるコンサルティングサービスの事例は、実務に根ざしたAIガバナンスの構築方法を示すものです。生成AIの恩恵とリスクの両面を正しく理解し、責任ある活用を進めていくことこそが、技術を人々の暮らしと未来に貢献させるために不可欠であり、今後の持続可能な社会の実現に向けた重要なステップとなるでしょう。


NTT DATA
※本レポートは2025年10月16日時点の情報を元にNTT DATAが主となって作成しています