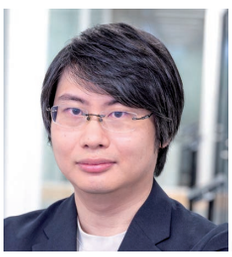
5 生成 AI のガバナンス

生成AI活用における リスクマネジメントのポイント

生成 AI は、その便利さが強調される一方で、その出力に含まれる誤りやバイアスに関するリスクや、生成指示からの情報漏洩リスク、学習データや生成物の著作権侵害リスクなど、利用にあたって様々なリスクがあることが指摘されている。本稿では、生成 AI に関するリスクの特徴と、各国で進む規制の動向を紹介し、これらのリスクに対するマネジメントのポイントを解説する。



株式会社 NTT データグループ
技術革新統括本部システム技術本部
D&I 技術部　AI ガバナンス室
室長　**伏田 享平** 氏、課長　**安部 裕之** 氏、課長代理　**鈴木 賢一郎** 氏

生成 AI に対するリスクマネジメントの必要性

2022 年の後半からデータを生成する AI は急激に進化しており、画像生成 AI や、文書生成 AI は広く使われるようになった。これらの AI は、人間の生成物と比しても劣らない出力をすることから、その利便性に注目が集まっている。

一方で生成 AI には、出力に誤りやバイアスを含むリスクや、フェイクニュースのような悪用できるコンテンツが容易に生成できるリスクがある。加えて、入力（プロンプト）からの情報漏洩リスク、学習データや生成物に対する著作権侵害のリスクなど、様々なリスクがあることが指摘されている。これらのリスクに対して、一部の地域・国家レベルでも規制を強化するする動きがある。そのため、生成 AI の利用にあたっては、規制への対応とあわせてこれらのリスクを適切にマネジメントする必要がある。本稿では生成 AI のリスクの特徴と各国の規制動向を紹介し、リスクマネジメントのポイントを解説する。

生成 AI のリスクの特徴

従来の AI では、特定の目的のために都度学習データの収集と学習アルゴリズムの選定をし、学習を行い、構築したモデルにデータを入力して推論結果を得ていた。これに対して生成 AI は、多くの場合、事前学習済みの大規模モデルが生成指示を受け付け、指示に応じた様々な媒体のコンテンツを生成する点が大きく異なる。従来の AI と比較して、生成 AI はモデルの動作が見えず、何が生成されるのか生成指示を与えてみないとわからない。そのため、これまでの AI 以上に内部の挙動を把握することが困難なブラックボックスとなっている。

生成 AI の利用にあたっては、先にも述べたとおり様々なリスクがある。これらのリスクを企業として留意すべき観点から整理すると以下の 4 つに大別される。

・サービス提供者の責任を問われるリスク

生成 AI の出力は必ずしも正確ではないことが知られている。提供する AI サービスの出力結果に何らかの瑕疵があり、お客様へ損害を与えた場合には、サービス提供者として責任を問われる。また、悪意を持ってフェイクニュース（テキスト、画像）やマルウェアのような正確ではない情報を生成し、その結果社会的混乱が生じた場合、サービス提供者の社会的信用も大きく毀損される。

・権利侵害リスク

生成 AI は様々な著作権侵害の可能性が指摘されており、それにともなう訴訟も提起されている。生成 AI は、入力次第では学習したデータをそのまま再現して出力できる場

合がある。特にその出力に学習を許諾していないデータが含まれていた場合には、権利侵害・契約違反等に基づく差し止め請求等が発生することが考えられる。実際にアメリカでは複数の画像生成AIに対して訴訟が提起されている。また、提供する生成AIサービスの出力結果が他者の著作権や肖像権、商標らを侵害した場合にも、訴訟を提起されることが考えられる。

・情報漏洩リスク

生成AIにはユーザーが生成指示として入力した情報を再学習するものがある。このような生成AIに生成指示として機密情報が入力され、再学習に利用された場合には、他のAI利用者にその機密情報が出力されてしまう情報漏洩のリスクがある。また、生成指示に個人情報を入力した場合、再学習の有無に関わらず入力したこと自体が個人情報の第三者提供に該当するとみなされ、個人情報保護法違反を問われるリスクがある。

・法令等への抵触リスク

生成AIは新しい技術であり、その利用の仕方次第では既存の法令や規制に抵触する場合がある。例えば、生成AIによる診療サービスを当局の許可を得ずに提供すれば、法令違反となるおそれがある。

各国規制の検討状況

先に示したように、生成AIの利用に関して様々なリスクが指摘されるなか、世界各国で法規制やガイドラインなどAIの開発や利用に対して統制をかける動きが進んでいる。2023年のG7広島サミットでは広島AIプロセスとして生成AIを議論する場が設置され、12月に首脳合意がされた（図1）。ここでは日米とEU、およびG7の動向を紹介する。

① 日本

2022年までは各省庁が定めたガイドラインをもとに、各事業者の活動はその自主性にゆだねられる「ソフトロー」の路線であった。2023年に入り生成AIの登場により国内外の状況が大きく変化したなか、政府にAI戦略会議が設置され、戦略の見直しが行われている。具体的には内閣府、総務省、経産省が連携して事業者向けの新たな統合ガイドラインの策定が行われている。総務省と経産省のガイドラインをベースに、生成AIなど新しい要素も取り込んだうえで、1つにまとめるもので、2023年度内に公開される予定である。

また、2023年12月に自民党から出された提言では、AI事業者に対してガイドラインの遵守を求める法制化について言及されている。

② アメリカ

2023年8月にバイデン大統領がOpenAIやマイクロソフト、グーグルなど7社のトップと会談を行った。会談では生成AIの急速な発展を受け、開発企業がサービス発売前の段階で外部専門家による検証やリスク評価を通じ、AIの安全性や信頼性を確保する自主的な取り組みを進めることで合意した。2023年10月末にはAIの安全性に関する大統領令が発行され、そのなかでAIに対する法規制について言及があった。

③ EU

EUのAI規制法は2021年4月に法案が公開され、2023年12月9日に大筋合意がされた。今後は加

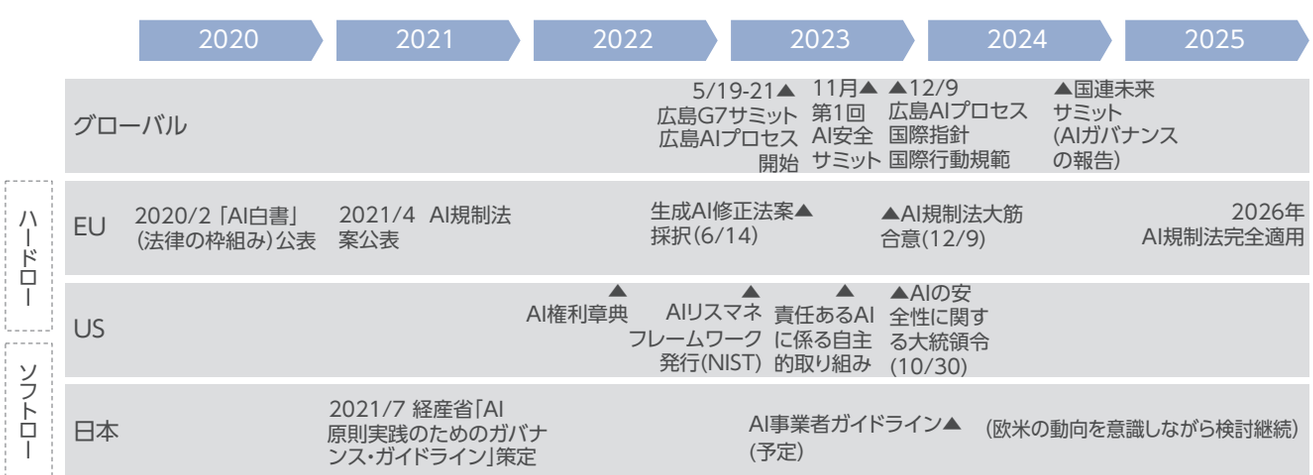


図1　世界の規制動向

盟国とヨーロッパ議会による正式な承認を経て成立し、2026年にも完全適用されるといわれている。

EUのAI規制法は、「リスクベース・アプローチ」により、倫理リスクの観点からAIを「受容できないAI」「ハイリスクAI」「透明性義務をともなうAI」「極小リスクAI」の4つにカテゴライズし、それぞれに対策を義務づけている。生成AIは様々なユースケースが想定されることから、「Foundation Model」としてこの枠組みとは別に、モデルの提供者に対して運用上の義務と透明性義務の規制、学習データの開示義務など、生成AI特有の規制を課している。違反した場合は最大で3,500万ユーロ（約55億円）あるいは 全世界売上高の7%の大きい方の金額の制裁金が科される。

④ G7

2023年のG7広島サミットではグローバルAIガバナンスが主要トピックとして取り上げられ、生成AIに関する議論のための「広島AIプロセス」が創設された。様々な議論の結果、12か条からなる国際指針と、指針に対してAI開発企業が取るべき対策事例を例示した行動規範が取りまとめられた。具体的には、AI関連企業が製品を市場に投入する前に外部の専門家のチェックを受けることや、人が作成したものと区別するためAI生成物への電子透かし導入などが言及されている。

企業としてのリスクマネジメントのポイント

これまでに紹介した生成AIの様々なリスクを効果的にマネジメントする際のポイントを2つ紹介する。

1つ目は、生成AIの利用の「立場」を明確にすることである。具体的には以下の3つの立場がある（図2）。

- 「**モデル開発者**」：データを学習または追加学習しモデルを作成する。
- 「**プロダクト提供者**」：学習済みモデルや外部サービスを利用してアプリケーションを開発、提供する。
- 「**利用者**」：生成指示をして出力される生成物を利用する。

一方で、生成AIのバリューチェーンは「モデル開発」「チューニング」「プロダクトへの組み込み」「生成指示」「出力の利用」の活動から構成される。このバリューチェーンにおいて、自組織の活動がどちらにあたるのか明確にしたうえで、先に示した3つの立場に応じて4つのリスクのどれが該当するのか分析し、該当するリスクに対して、そのリスクを低減/回避する対応策を設計するアプローチが有効である。

もう1つは、リスクに対する免責事項を明確にすることである。例えば、生成AIの出力は100%正しいとは限らない。プロダクト提供者の立場からはその誤りをどちらまで許容いただけるのか、お客様と協議が必要である。もし、誤りが許容されない業務に適用するような場合には、人的確認やフェールセーフの仕組みを導入する必要がある。

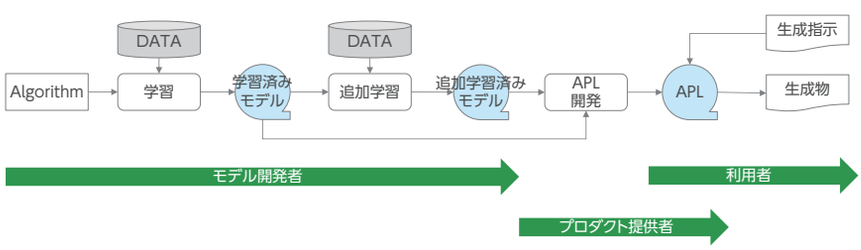


図2　生成AIのバリューチェーン

生成AIは新しい技術であり、関係するステークホルダーごとのリスクとベネフィットがまだ明確になっていない。そのため、リスクへの認識はバラツキが大きく、適切なリスクコミュニケーションが求められる。利用するユースケースに応じて、上記のような正確性に限らず、権利の侵害や情報漏洩、個人情報保護といった観点から、プロダクト提供者であればリスクと軽減策の説明を、利用者であればリスクの理解をすることが重要である。

NTTデータグループでは生成AI利用における3つの立場と4つのリスクについて、社員向けに留意事項と対処方針をまとめたガイドラインを整備している。

今後に向けて

生成AIの出力は人間が生成するのにも劣らないことから、社会に大きな変化をもたらす可能性を秘めている。その反面、利用にあたっては従来のAIよりも広範なリスクがあり、適切なリスクマネジメントが求められる。NTTデータグループではAIガバナンス活動を通して、利用にあたってのリスクを低減しその価値を最大化することで、お客様への安全なAIシステム提供を実現する。