SBI Financial and Economic Research Institute

FERI INFORMATION

AI エージェント時代の情報漏洩リスク
―生成 AI による内部露出とデータ統治―

関口　麻衣子（SBI 金融経済研究所 研究員）

AI 活用が露わにするリスク

日本企業において、法人向け生成 AI（Microsoft 365 Copilot や Gemini for Google Workspace 等の法人契約版）は、日常業務を支える業務基盤の一部として浸透している。これらのサービスでは、文書要約、問い合わせ対応、資料やプログラムコードの作成補助、内部調査、稟議支援などに加え、ユーザーの権限範囲内にある組織内のファイル、メール、チャット、会議情報等を参照して回答を生成する利用も広がっている。さらに、こうした生成 AI 機能を基盤として、一連の業務を自律的に進める AI エージェント型サービスの利用も進んでいる。ただし、その利用の程度やそれを支える管理体制の整備には、組織間で差がある。

こうした法人向け生成 AI や AI エージェント型サービスの広がりは、これまで人間による利用を前提としてきたデータ管理や権限設計、説明責任の在り方を、改めて問い直す契機となっている。本レポートでは、法人向け生成 AI を出発点として、業務システムや外部ツールと連携し、業務上の操作にまで広がる利用形態を AI エージェント型サービスとして扱う。

その形態は多様であり、法人向け生成 AI の拡張機能として組み込まれているものから、既存の SaaS（Software as a Service）等の業務サービスにあらかじめ搭載されているもの、独立したサービスとして特定の業務プロセスを遂行するもの、自組織の要件に応じて構築するものまで幅広く含まれる。ユーザーが生成 AI を使っているつもりでも実はエージェント機能が動いているケースや、業務サービスにこうした機能が組み込まれていることを意識せずに利用しているケースも少なくない。本レポートでは、こうしたサービスの導入によって、これまで見えにくかった情報管理上の論点がなぜ顕在化するのかを「データ統治」の視点から整理する。

入力情報は外部に流出するのか

法人向け生成 AI については、主なプラットフォーマーが提供するサービスの利用条件やシステム設計において、概ね以下の点が前提として示されている。

- 生成 AI で参照・処理した情報（入力したプロンプトや処理対象として指定したファイル等）は、外部モデルの学習に使用されない。

- 上記情報は、他の組織に流用・再利用されない。
- 生成 AI が参照・処理できる範囲は、ユーザーに付与されたアクセス権限を超えない。

したがって、法人向け生成 AI について、「プロンプトを入力したり、処理対象ファイルを指定したりしただけで、そこに含まれる情報が当然に外部へ流出する」という理解は、少なくとも現在の主な法人向けサービスの契約・設計を前提とする限り、正確ではない。

もっとも、これは法人向け生成 AI の利用に情報管理上のリスクがないことを意味するものではない。上記の整理は、セキュリティ管理が整備された法人契約版を利用し、未承認の個人向けサービスを業務利用しないこと、入力・参照させる機密情報や個人情報の取扱いに関する組織的なルールが整備・遵守されていることを前提とする。以下では、このような前提の下でもなお残る、組織内の情報管理上の論点を確認する。

内部露出リスクの顕在化

公的機関や業界団体などが指摘しているのは、生成 AI や AI エージェント型サービスが直ちに新たな外部流出経路を生むというよりも、「組織内に既に存在していた情報管理上の脆弱性を増幅し、顕在化し得る」という点である。本レポートでは、データ統治（データガバナンス）を、組織が保有するデータを適切に管理・利活用するための方針、ルール、体制、責任分担を含む枠組みとして捉える。特に AI 利用を念頭に、データの取得・保存・利用・廃棄、アクセス権限の設計、利用目的の管理、二次生成物の扱い、説明責任を含むものとする。

前節で確認したとおり、問題は AI がアクセス権限を超えて情報を取得する点にあるのではない。むしろ、ユーザーに付与されたアクセス権限の範囲内で、組織内の情報が検索・要約・連結され、従来であれば目に触れにくかった情報や、別々の文脈で管理されていた情報が、一つの回答として再構成され得る点にある。本レポートでは、こうした組織外への流出を伴わない情報の露出を、広い意味での情報漏洩リスクとして捉え、「内部露出リスク」と呼ぶ。検討中の M&A、人事情報、経営戦略、資本政策など、厳格な内部管理を要する情報は企業内に多数存在する。これらの情報は、情報システム、共有領域、または従業員端末上のクラウド同期フォルダ等に保存されている場合でも、AI エージェント型サービスの連携範囲に含まれれば、参照・処理の対象となり得る。

多くの法人向け生成 AI は、広範な公開情報に加え、あらかじめ設定された組織内のデータ領域を参照して回答を生成する検索拡張生成（RAG：Retrieval-Augmented Generation）の仕組みを組み込んでいる。従来の全文検索が情報の保存場所や関連文書を提示するにとどまっていたのに対し、生成 AI は、権限範囲内の複数の情報を読み込み、要約し、ユーザーの問いに応じた回答として再構成する。

さらに、AI エージェント型サービスでは、MCP（Model Context Protocol）等の連携方式を通じ、各種システムや外部ツールと連携し、情報の検索・抽出、登録・更新・通知等の操作を行う利用が広がっている。このように、生成 AI や AI エージェント型サービスは、情報の所在を示すだけでなく、それらの情報を回答や処理結果として再構成し、場合により操作につなげる。こうした特性により、組織内に既に存在する情報管理上の脆弱性が表面化し得る。

こうした脆弱性を具体化する要因として、まず、AI 利用を前提とした再評価・棚卸しが行われていない過去からの蓄積データがある。加えて、メール、チャットログ、会議録、作業メモ、業務システム上のログや中間ファイル、ローカル端末で作成されクラウド同期等により蓄積された文書案など、日々の業務過程で生成・保存される膨大な非構造化データもある。多くの組織では、以下のような類型のデータが、長年にわたり十分に整理されないまま蓄積され、AI の参照・処理対象となり得る状態にある（図表１）。

図表１：リスクとなり得る「未整理の蓄積・非構造化データ」の類型

区分	データの例
【保管・アクセス未整理】 保管・アクセス管理において権限が整理されていないデータ	過去のプロジェクトや組織体制の下で作成され、管理のライフサイクル（データの作成から廃棄までの管理プロセス）から外れたまま蓄積され、整理・更新されていない共有領域（共有ドライブや部門別共有ドライブ等）や文書
	全社共有（Everyone 権限）や部署内全員参照など、利便性を優先して広範な参照許可が与えられたまま、現在の実務実態と乖離している共有領域や文書
	紙媒体では廃棄されることを前提としていた一方、電子データとしては保存・利用の前提が改めて整理されないまま残存している文書
【管理主体・文脈未整理】 管理主体や利用文脈が不明確なまま蓄積されたデータ	退職者や案件の旧担当者が作成し、暫定的な引継ぎ用や個人メモとして残され、管理主体が不明確なまま残っている文書やメールアーカイブ
	会議の自動文字起こしやチャットログなどの機密性の高い発言を含み得る記録データ
	PC 端末内のローカル領域（C・D ドライブ等）で作成・保存され、クラウド同期等によりユーザーが意識しない形でクラウドストレージ上に蓄積された作業メモや文書案
	過去に CC（写し）で共有された内部議論やその背景となるクローズド情報
【利用目的固定】 情報取得時の利用目的や前提が固定されたままのデータ	AI 利用を想定せずに取得・蓄積された顧客・従業員に関する情報（問い合わせ履歴、面談記録等）
	取得時の同意内容や利用目的等の法的前提が、AI 利用という新たな文脈で精査されないまま残存している情報（例：電話応対の「応対品質向上」目的の録音データが、AI によりマーケティング分析や従業員の業務習熟度の判定（人事評価）等に準用されるなど、当初の想定範囲を超えて利用される情報）
【派生データ未管理】 AI によって要約・再構成された二次生成物	AI の生成物（要約テキスト等）であって、元の情報に付随する利用制限や廃棄ルールが当該生成物に適切に継承・適用されないまま残存している情報

注）本レポートの図表における「AI」は、法人向け生成 AI および AI エージェント型サービスを指す。

出所）筆者作成

これらはいずれも、必ずしも不適切に取得・保存された情報ではない。しかし、保管場所やアクセス権限、管理主体、利用文脈、利用目的が十分に整理されないまま残存している点に

共通の特徴がある。これまで各組織は、ストレージコストの低廉化を背景に、これらのデータを「将来の活用可能性（資産性）」を期待して広く保持してきた面がある。一方で、生成 AI や AI エージェント型サービスは、これらを横断的に検索・連結・処理することで、ガバナンス上の潜在的負債として顕在化させる可能性がある。

人間前提の管理方法の限界

前節で述べたとおり、生成 AI や AI エージェント型サービスは、ユーザーに付与されたアクセス権限や、組織があらかじめ連携対象として設定した範囲に基づいて情報を参照・処理する。しかし、人間が個別に探し、読み、判断することを前提に構築されてきた管理方法が、AI の参照・処理を技術的に制御する仕組みとしては十分でない場合がある。

例えば、文書名に「機密」や「AI 参照不可」と記載しても、生成 AI がそれを技術的な読み取り禁止命令として解釈するわけではない。パスワード保護されたファイルも、運用上パスワードを解除して共有領域に保存し直していれば参照対象となり得る。また、一見したところ安全に思える情報の分散保管であっても、AI によりそれらが検索・連結・再構成され、意図しない内部露出が生じる可能性もある（図表２）。

図表２：人間前提の管理方法の例と AI 利用環境下での限界（イメージ）

従来の管理方法例	人間前提の環境での効果 （なぜ通用していたか）	AI 利用環境下での限界 （なぜ脆弱性となり得るか）
ファイル名による注記（「機密」、「AI 参照不可」等）	注意喚起や、人間に対する心理的な抑止力としての効果	AI はファイル名の注記を技術的な読み取り禁止命令として解釈せず、参照は自動的には停止されない。プロンプトによる除外指示もシステム的な強制力を持たないため、参照・処理対象となり得る。
パスワード保護	ファイルへのアクセスを技術的・個別的に制限する効果	チーム内での検索性を優先してパスワードを解除して共有領域に保存し直している場合や、閲覧権限のみで制御しファイル自体の暗号化を行っていない場合には、参照・処理対象となり得る。
PDF 化や画像データ化	改ざんや、容易なコピー・再利用を防ぐ効果	改ざん防止には有効なものの、テキストが埋め込まれた PDF だけではなく、スキャン画像の PDF も OCR 等の設定次第で参照・処理対象となり得る。
深いフォルダ階層への格納	「階層の奥深くに隠す」ことで、関係者以外のアクセスを実質的に防ぐ効果	AI はフォルダ階層にかかわらずアクセス権限の範囲内の情報を参照するため、「階層の奥深くに隠す」ことによる秘匿効果が薄れ、参照・処理対象となり得る。
PPAP・情報の分散保管	情報を分散させることで突き合わせの手間を生み、結果として実質的な秘匿性を担保していた側面	別々のメールや文書に分散して存在する関連情報が連結され、ユーザーが明示的に突き合わせていない情報まで、参照・処理対象となり得る。 ※情報セキュリティの観点から、PPAP の見直しが進む組織であっても、情報が分散保管されている状況は同様であり、アクセス権管理の再整理が求められる。

注）PPAP は、パスワード付きファイルをメールで送付し、パスワードを別送する運用を指す。

出所）筆者作成

これらの例はいずれも、従来、人間が手間をかけて情報を探し、読み、突き合わせるという物理的障壁、いわば「手間の壁」が、実質的な秘匿性を支えていたことを示している。生成AIやAIエージェント型サービスは、この手間の壁を低くし、業務の効率化をもたらした。一方で、従来の管理方法が、AIによる高速な参照・連結・処理を前提とした環境でも同じように機能するとは限らない。今後は、アクセス権限やデータ分類といった技術的対策の有効性に加え、運用面・ガバナンス面を含めて、設計段階に遡った検討が必要となる。

「暗黙知フィルタ」が効かない

人間前提の業務環境では、形式的にはアクセス可能な情報であっても、情報の鮮度や確度、重要度、自身の関与度、参照頻度、さらに当該情報がどのような信頼関係や文脈（作成者の意図や前提条件を含む、情報の背景事情）の下で共有されたかを踏まえ、実務上の判断材料として用いるか、また他者と共有してよい情報かが、暗黙に選別されてきた。日本の組織においては、「念のため関係者全員をCCに入れる」、「とりあえず共有フォルダに置く」といった広範な共有慣行も少なくない。こうした慣行には相応の合理性があるものの、時点、確度、文脈の異なる情報が大量に蓄積され、実務上の「情報のノイズ」も生じやすい。もっとも、人間は文脈や空気を読んで不必要な情報を除外し、あるいは「これはあくまで参考情報にとどまる」と割り引くことで、判断材料を選別してきた。本レポートでは、こうした暗黙の選別を「暗黙知フィルタ」と呼ぶ。

しかし、生成AIは、人間が行ってきたこうした暗黙の選別を、そのまま再現するわけではない。参照範囲の日付指定や対象フォルダの限定など、AIの挙動を制約するハーネス技術により対応可能な部分もあるが、システム上の更新日と内容の鮮度が一致しない場合や、日付や名称だけでは決定事項か検討段階かを判別できない場合には、情報の時点、確度、文脈を十分に切り分けられないことがある。

その結果、異なる時点、確度、文脈の情報が、ローコンテクストかつフラットに扱われ、一つの回答の中で連結・再構成されることがある。これがガバナンスの観点では、情報の予期せぬ連結や、当初の文脈を離れた再構成として現れる。もちろん、RAGの技術的制約（例えば、読み込み文字数の限界や検索精度）により全ての情報が常に連結されるわけではないが、これまで人間の目には触れにくかった情報が回答の文脈に入り込む可能性が高まり得る。この点は、複数の情報源の検索・参照から業務上の操作までを一連の流れとして行うAIエージェント型サービスでは、従来以上にガバナンス上の論点となる（図表３）。

図表３：情報処理のプロセス比較（イメージ）

プロセス	人間による情報処理	AI による情報処理
情報の取得	「ユーザーの意図」に基づき、特定の文書・データの所在を検索する	「ユーザーの権限」に基づき、権限範囲内の文書・データを横断的に検索・参照する
情報の選別・処理	【暗黙知フィルタ】 鮮度、確度、重要度等を踏まえ、情報を暗黙のうちに選別する	【機械的な情報連結】 情報をローコンテクストかつフラットに扱い、技術的に参照可能な情報を連結する
情報の提示	文書を個別に開き、必要な箇所を参照・転記する	読み込んだ複数の情報を回答や処理結果として再構成し、場合によっては操作につなげる
ガバナンス	「探す・読む手間（物理的障壁）」が実質的な防壁になる	「情報の予期せぬ連結」や「当初の文脈を離れた再構成」のリスクが顕在化しやすい

出所）筆者作成

例えば、正規のアクセス権限を持つユーザーが「次年度のオフィス戦略をまとめて」と入力した際、生成 AI が、古い備品リスト、新しい拠点契約書、公開前の社内原稿、ローカル端末からクラウド同期された検討メモ、CC で形式的に共有された過去のメールを一つの回答の中で結び付ける可能性がある。その結果、公式発表前の拠点閉鎖や縮小移転、企業買収に伴う拠点再編の計画が、確定・未確定の区別が十分に示されないまま提示されるリスクがある。

このとき問題となるのは、個々の情報の正確性だけではない。どの情報が、どの時点の、どの確度・文脈の情報として用いられたのかが、回答の中で見えにくくなる点である。一部のシステムでは参照元を提示する機能もあるが、複数の情報がどのように選択され、どのような重み付けで回答に反映されたのかを完全に可視化することには限界がある。人間であれば見過ごすような、あるいは物理的に探し出すことが困難だった「点と点」の情報が AI によって再構成されることで、本来その情報を知る必要のない立場の役職員にまで、意図しない機密の露出を招く可能性がある。

特に、厳格な情報隔離が求められる組織では、AI による情報の横断的な連結はより深刻な意味を持つ。銀行・証券・信託が一体となった金融グループにおけるチャイニーズウォール（情報隔壁）に限らず、部門間・グループ会社間で利益相反管理や情報隔離が求められる組織においても、AI による横断的な参照はアクセス分離や情報管理の実効性を問う契機となり得る。こうした管理上の課題は、生成 AI 導入前から存在していたものの、人間前提の業務環境や既存の管理体制の下では見えにくかった面もある。

内部露出リスクの具体的シナリオ

ここまで整理してきた内部露出リスクが具体的にどのように生じ得るか、人事領域と顧客データの利用を例に確認する。

例えば、人事領域では、過去の制度検討資料やそのドラフト、バージョン違いの資料、各種面談の要点整理メモなどが、当時の意思決定や業務遂行を前提として保管されていることがある。こうした情報の中には、一定期間後も削除されず、あるいは削除対象とは必ずしも位置付けられないまま、蓄積されているものも少なくない。この状態で生成 AI に「過去の人事制度見直しの経緯を要約して」と入力すると、人間の業務では既に利用されない前提であった資料や未確定の検討メモが、現在の有効な判断材料であるかのように再構成され、回答として提示される可能性がある。

また、顧客データについても同様の論点がある。顧客データを収集した時点では、生成 AI や AI エージェント型サービスによる分析、自動要約、営業・マーケティング上の示唆抽出などを利用目的として想定していなかった場合がある。その後、これらのサービスを通じて当該データが分析結果や示唆として再構成・提示されると、技術的には正当なアクセス範囲内であっても、当該データが取得時とは異なる利用文脈に置かれることで、法的整理や説明責任の観点から論点が生じ得る。

上記の例に限らず、情報セキュリティやアクセス管理が相当程度整備されている組織であっても、AI が蓄積された情報を横断的に参照・処理する場合には、既存の管理前提が現在の利用文脈に照らしてなお妥当かを再確認する必要がある。単なる業務フローの見直しや従来型の閲覧範囲の限定にとどまり、AI による連結・再構成リスクに対応したデータ統治が整理されていない場合、形式的には正当なアクセスの範囲内であっても、管理上の前提が曖昧な情報を想定外の形で可視化する可能性がある。

責任の重心は「設計」へ

このような状況では、「AI の回答や操作の結果に誰が責任を持つのか」という問題が生じる。このとき問われるのは、ユーザーが個々の場面でどのようなプロンプトを入力したかだけではない。むしろ、どのデータを AI の参照・処理対象とし、どの範囲でアクセスを認め、回答・操作結果や二次生成物をどう管理するかといった、事前の設計そのものである。

すなわち、責任の重心は、個々の操作から、アクセス権限、データ配置、検索対象、利用目的を含むガバナンス設計へと移る。データの利活用範囲をどこまで広げ、どのような統制を置くのかというトレードオフは、単なる IT 部門の技術課題ではなく、組織としてのリスク許容度（リスク・アペタイト）の問題である。そのため、経営層、業務部門、法務・コンプライアンス、リスク管理、情報システム、内部監査などが関与して組織横断的に判断し、必要に応じてステークホルダーに説明できる体制を整えることが求められる。

こうした整理は、内部監査の国際的実務指針にも反映されている。内部監査人協会（The Institute of Internal Auditors）の AI 監査フレームワークは、AI 利用を個々の現場行為として捉えるのではなく、ガバナンス、責任分担（業務、リスク・コンプライアンス、内部監査の各

機能による、いわゆる三線モデル)、および統制の枠組みが適切に整備・運用されているかを評価の対象として位置付けている。したがって、生成 AI や AI エージェント型サービスの導入に伴って問われるべきは、ユーザーの操作の是非だけではなく、どのような統治と設計の下で AI を利用しているかである。こうした認識の下、国内外の公的機関においても、各国・地域でアプローチは異なるものの、AI の利活用に関する枠組みの整備が進められている（図表４）。

図表４： 公的機関による AI の利活用に関する主な枠組み

国・地域	主な枠組み	概要
日本	内閣府 AI 戦略会議資料「AI 制度に関する考え方」（2024 年 5 月）、総務省および経済産業省の「AI 事業者ガイドライン」等	イノベーションの促進とリスク対応を両立させる制度・ガイドライン
	独立行政法人情報処理推進機構の「テキスト生成 AI の導入・運用ガイドライン」、日本銀行の「金融システムレポート別冊：金融機関における生成 AI の利用状況とリスク管理」、金融庁の「AI ディスカッションペーパー：金融分野における AI の健全な利活用の促進に向けた初期的な論点整理」等	現場向けの手引きや業種固有の実務整理
米国	大統領令「人工知能に関する国家政策の枠組みの確立（*Ensuring a National Policy Framework for Artificial Intelligence*）」（2025 年 12 月）	連邦レベルでの統一的な政策方針
	米国立標準技術研究所（National Institute of Standards and Technology：NIST）の「AI リスクマネジメントフレームワーク（*AI Risk Management Framework*）」、同「生成 AI プロファイル（*Generative AI Profile*）」	AI リスク管理に関する技術的ガイダンス
欧州	「欧州 AI 法（*EU AI Act*）」（2024 年 8 月発効）	リスクの影響度に応じた義務付け（段階的に適用）
（参考）国際	ISO/IEC 42001	AI 管理体制の整備･運用に関する国際規格

出所）各機関公表資料を基に筆者作成

これらの枠組みは、AI 利活用における原則や統治の方向性を示すものである。本レポートが論じているのは、そうした枠組みの下でもなお「具体的に検討が必要」となる、法人向け生成 AI や AI エージェント型サービスに特有の情報管理上の論点である。

「事故」を伴わない情報漏洩

一般的に「情報漏洩」や「データ侵害」と言えば、不正アクセス、サイバー攻撃、誤送信、外部への持ち出しといった「事故」を想起しやすい。しかし、本レポートで問題としている内部露出リスクは、必ずしもこうした従来型の事故を伴うものではない。むしろ、次の掛け合わせとして整理できる。

正当な権限 × 通常の操作 × 意図しない連結・再構成

いずれも正規の業務範囲内で生じるため、こうした事故を伴わない形での内部露出は、従来のセキュリティ対策では異常として検知されにくい。問題を「事故が起きたかどうか」だけ

で捉えるのではなく、AI がどの情報を参照し、どのように結び付け、どの文脈で提示したのかを、データ統治の観点から再整理する必要がある。

新しい論点か、既存課題の可視化か

本レポートがここまで整理してきた情報管理上のリスクは、ハルシネーションや著作権侵害のような生成物の品質や外部利用に関わる論点とは射程を異にする。むしろ、既存課題が AI の導入により増幅・顕在化され得るとい__う認識は、公的機関やプラットフォーマー、業界団体においても示されている。

（1）公的機関：既存論点の増幅・顕在化リスク

米国立標準技術研究所（NIST）は、前述の「生成 AI プロファイル」において、生成 AI に固有のリスクだけではなく、生成 AI によって悪化・増幅（exacerbate）され得る既存リスクを整理している。前者には、ハルシネーションや不適切なコンテンツ生成などが含まれ、後者には、情報セキュリティ、プライバシー、ガバナンスなど、既存の情報管理上の論点が含まれる。この整理は、組織が従来から抱えていた管理上の課題が、生成 AI の利用環境において予期せぬ形で顕在化し得るという本レポートの問題意識とも整合する。同様の視点は国内の金融分野における整理とも接続しており、前述の日本銀行や金融庁の公表資料においても、生成 AI の利活用に伴うリスク管理やガバナンス上の論点が整理されている。

（2）プラットフォーマー：意図せぬ過剰共有リスク

AI・クラウドサービスを提供するプラットフォーマーも、利用者側の重要な対応として、組織内での「意図しない過剰共有（accidental oversharing）」の削減や是正を推奨している。これは、AI モデルそのものの問題に限らず、人間による業務において許容されてきた既存の広範な共有設定というデータ設計が、AI 導入後に顕在化するリスクの源泉であることを示している。

（3）業界団体：入力文脈に入り込むデータのリスク

業界団体においても、データセキュリティに着目した整理が進んでいる。アプリケーションセキュリティの標準策定を主導する非営利団体の OWASP（Open Worldwide Application Security Project）の生成 AI セキュリティ・プロジェクトによる「生成 AI データセキュリティのリスクと緩和策（*OWASP GenAI Data Security Risks & Mitigations 2026*）」（2026 年 3 月）では、主なリスクの一つとして、過度に広いコンテキストウィンドウやプロンプトへの

過剰な情報付与（over-broad context windows & prompt over-sharing）が挙げられている。これは、ユーザーが明示的に情報を入力する場合だけでなく、RAG やツール連携、AI エージェント型サービス等を通じて、回答や処理に必要な範囲を超える情報が入力文脈（コンテキスト）に含まれ得るという問題である。

前述のプラットフォーマーが指摘する「意図しない過剰共有」が権限設定や共有慣行に起因するリスクであるのに対し、OWASP の整理は、入力コンテキストに含まれた情報の確度や利用の適否を、AI が人間と同じように選別するわけではないという、設計・運用上のリスクに着目している。

（4）現実：リスク認識に対して対策が追いつかない

生成 AI や AI エージェント型サービスの利用が進展する中で、既存データの整理や統制はなお実務上の課題として残っている。米国大企業を対象とした調査（Utimaco、2026 年 3 月）では、生成 AI に伴うデータ保護上の懸念が高いにもかかわらず、過半数の企業で具体的な対策が導入されていないことが示されている。国内でも、経済産業省や独立行政法人情報処理推進機構の整理・調査から、AI やデータの利活用の重要性が政策的・戦略的に位置付けられる一方、データ統治や推進体制の整備にはなお課題が残ることがうかがえる。

結論：AI は統治課題を露わにする「触媒」

ここまでの議論を踏まえると、生成 AI や AI エージェント型サービスは、単独で内部露出リスクを生み出す「主因」になっているわけではなく、従来から存在していた情報管理やデータ統治の曖昧さを顕在化させる「触媒」として整理できる。生成 AI は、ユーザーに付与されたアクセス権限を超えて情報を取得するものではない。しかし、権限範囲内にある情報を連結・再構成し、さらには AI エージェント型サービスでは業務上の操作にまでつなげることで、これまで人間の手間や文脈判断によって見えにくかった管理上の課題を表面化させる。

各組織に求められるのは、AI そのものを危険視して利用を抑制することでも、アルゴリズムの理解だけで事足りるとすることでもない。問題の所在は、情報の置き方、権限の与え方、データライフサイクル管理、二次生成物の扱い、利用目的の管理、そして組織横断的な統治の在り方にある。これらを再点検し、その妥当性を説明できるデータ統治体制を整えること――すなわち、生成 AI や AI エージェント型サービスの導入を、既存のデータ管理を現代的なリスク環境に合わせて再構築する契機として捉えることが重要である（図表 5）。

図表５：内部露出リスクに対応する実務アプローチ

対策	具体的な対策内容・アプローチ	対応する主な区分（図表１で提示）
アクセス・権限管理の再設計	AI による参照・処理・操作を前提とした権限設計の見直し（共有フォルダや Everyone 権限の棚卸しを含む）	保管・アクセス未整理
	部門間・グループ会社間における情報隔離の実効性の確認（AI 利用を踏まえた再点検を含む）	
	「秘密度ラベル」や RAG のホワイトリスト管理等を活用した AI の参照・処理範囲の限定（データ単位やデータ保存領域単位で、AI による参照・処理を制御する設計への移行）	
	AI エージェント型サービスによる操作に対する承認・制御の設計	
データ分類・ライフサイクル管理の再構築	共有領域や保存された文書の定期的な整理（例：一定期間が過ぎた古いファイルの自動削除や移動の仕組み化）	保管・アクセス未整理
	ユーザーが意識しない形でクラウドに蓄積されるデータも含め、AI 利用を踏まえた管理対象の再定義	管理主体・文脈未整理
	AI 利用を前提としたデータ分類、ラベリング、検索インデックス等による検索対象の管理	
	二次生成物の保存・廃棄管理	派生データ未管理
利用目的・説明責任への対応	AI 利用を前提とした同意・説明の再設計	利用目的固定
組織横断的な統治体制の確立	経営層、業務部門、法務・コンプライアンスなどが参画する AI ガバナンス機能の確立（リスク許容度の設定、ルール策定、共通基盤整備を含む）	各区分を横断
	AI 利用に関する従業員向けの教育・啓発	
	AI の利用範囲や対象データに関する例外審査の仕組みの整備	
	対象データの利用根拠と、AI による参照・処理・操作の履歴を説明できるログ管理・モニタリング体制の整備	

注）上記に示した対策は、一度の整備にとどまらず、定期的な見直しが求められる。特にクラウド型サービスでは、事業者側の仕様変更に伴い、AI の参照・処理・操作範囲が意図せず変わる可能性があり、こうした変化への継続的な確認も必要となる。

出所）筆者作成

これらの対策は、データ統治の再構築として位置付けられる。その実施を個別部署任せにせず、組織横断的な統治機能を明確にしたうえで、リスク許容度の設定、ルール策定、共通基盤整備、教育支援、例外審査、モニタリングに取り組むことが重要である。そのうえで、「止めるか許可するか」の二分法ではなく、現場での利用と全体の統制を両立させる設計として再構築していくことが実務上の要点となる。

生成 AI や AI エージェント型サービスの導入は、単なる効率化の手段にとどまらず、各組織が長年蓄積してきたデータの管理状態を改めて点検する契機でもある。今後、AI が業務システムや組織内データとより深く結び付くほど、データ統治の在り方は、個別の IT 管理を超えた経営課題として重要性を増していく。こうした導入・利用を、組織全体のデータ統治を現代的なリスク環境に適応させる機会とできるかが、顧客・株主・取引先を含むステークホルダーからの信頼を支える一要素となる。