AIガバナンス構築の肝となるモデルの独立検証の重要性

既存のモデル管理の枠組みにおける検証の視点を有効に活用せよ

AI（人工知能）の活用が広がると同時に、AIの管理（AIガバナンス）の重要性が高まっている。海外の金融機関では、AIガバナンスはモデル管理（Model Risk Management＝MRM）の枠組みに組み込むことが一般的である。本稿では、まずMRMの枠組みを生かしたAIガバナンスの考え方とその利点について触れつつ、モデルの独立検証の重要性を述べる。その上で、独立検証では、どのような点に留意してAI（含む生成AI）の検証を行うべきか、個別のAIを具体例に挙げながら考え方を示す。さらに、今後のAIガバナンスのポイントにも簡単に触れたい。

あずさ監査法人
金融統轄事業部
アドバイザリー統轄事業部
ディレクター
田中 康浩

シニアマネジャー
竹川 正浩

AIガバナンスの土台としてのモデル管理

金融機関においてAIの活用が広がっている。その範囲も定型的な業務から、効率化・高度化を狙った業務まで拡大し、さまざまな用途で活用されている。

金融庁も今年3月に公表した「AIディスカッションペーパー（第1・0版）」において、（AI活用で）「チャレンジしないリスクも十分に認識されるべき」として、活用を後押ししている。同時に、AIのリスクの評価・軽減に焦点を当てたAIガバナンスの議論が盛り上がっている。

海外、とりわけ米国の金融機関では、AIはモデルの一つで

あり、モデル管理（ＭＲＭ）の枠組みを活用すべきという考え方が一般的である（注1）。ＭＲＭでは、ＡＩの開発者・使用者である1線に対して、1線から独立した2線（モデル検証者）が検証を行う。

米国のＭＲＭに係る監督は厳格である。ＦＲＢ（連邦準備制度理事会）とＯＣＣ（通貨監督庁）のガイダンス（注2）が公表されて以降、金融機関は相当のリソースをかけてＭＲＭの態勢を整備・高度化してきた。ＭＲＭでは、モデル開発からモデル特定・格付け付与、検証、承認、使用中のモニタリング・再検証、使用停止に至るライフサイクルに応じて、1線と2線の役割と責任を定め、モデルの管理を求めている。

ＭＲＭが経営管理に深く浸透している米国では、ＡＩをＭＲＭで管理するという発想は合理的である。一方、本邦金融機関では、独立検証を経てＡＩを使用するという発想は乏しく、ＡＩ（モデル）をＭＲＭで管理するという考え方はあまり浸透していない。

モデル管理の利点は三つの「仕組み」

ＭＲＭ活用の利点は、以下の三つの「仕組み」がすでにＭＲＭに備わっていることにある。

一つ目は、ライフサイクル管理である。ＡＩにもモデルのライフサイクルと同じ概念が存在する。

具体的には、構想から開発の際にテストが行われ、承認を受けて使用される。使用中も、挙動などに関するモニタリングが行われる。こうしたライフサイクル管理は、ＭＲＭが得意とする分野である。すなわち、ＭＲＭを活用すれば、新たなＡＩ管理の仕組みは不要である。

二つ目は、リスク格付けとインベントリー（一覧表）である。

モデルのリスク格付けは、リスクベース・アプローチの基礎となる重要な概念である。米国Ｇ-ＳＩＢｓ（グローバルな金融システム上重要な銀行）ではモデルは数千以上あり、すべてについて厳格な管理を行うことは不可能である。管理にめりはりを付けるためにも、ＡＩにリスク格付けを付与することは重要である。

また、格付けと共にＡＩをインベントリーに登録しておけば、使用中のＡＩの「見える化」につながる。これは、ＡＩ管理のインフラを提供するものだといえる。

三つ目は、検証・承認である。あらゆるモデルは2線による独立検証・承認を経て使用される。

モデル検証では、開発者にとって都合の良いモデルになっていないか、リスクや弱点・制約、使用上の留意点は何かなど、多面的な評価がなされる。米国ではモデル検証は厳格であり、リスクの高いモデルであれば１００ページを超える評価書（検証報告書）を作成し、承認を行う（否認や条件付き承認もあり得る）。

ＭＲＭで最も重要かつ強固な仕組みが、独立検証である。これによって安心・安全なＡＩ活用を担保できる。

ＭＲＭの独立検証では、あらゆるモデルに対して、データ、ロジック等の手法（Conceptual Soundness）、テスト（Outcome Analysis）、実装、ガバナンス等の視点から評価を行う（注3）。この検証の視点は、伝統的な統計モデルであれＡＩであれ、基本的には同じであるが、ＡＩでは特有のリスクを考慮することが重要になる。具体的には、説明可能性や公平性等の観点、ＥＵ（欧州連合）のＡＩ法などの法令リスク、プライバシー保護の観点等である（注4）。

従来型ＡＩで論点となる説明可能性や公平性の担保

本邦ではＭＲＭでＡＩの検証を行う事例はまだ少数である。そこで次に、どのような点に留意してＡＩの検証を実施すべきか、従来型ＡＩと生成ＡＩのそれぞれについて、イメージしやすいように架空の例を交えて考え方を示す（注5）。

まず図表1のような従来型ＡＩを取り上げる。

このタイプのＡＩが採用する機械学習アルゴリズムのランダムフォレストは、複雑な構造を持つため、伝統的な統計モデル

のように特定の出力に至った理由を直接的に説明することはできない。しかしSHAP（SHapley Additive exPlanations）値やLIME（Local Interpretable Model-agnostic Explanations）値から出力に対する特徴量の貢献度を知ることで、ランダムフォレストの説明可能性を一定程度、評価することができる。

また、伝統的な統計モデルやほかのAIを用いて、モデル構造の理解のしやすさと精度の両面から比較評価を行うことも有用である（ベンチマーキング）。モデル精度が大きく変わらなければ、説明可能性を優先して伝統的な統計モデルの利用を促す結論に至ることもある（この場合、伝統的な統計モデルがチャンピオンモデルになり、AIがチャレンジャーモデルになる）。

公平性の観点では、人種・国籍・性別等の情報を特徴量として採用すると、不当な差別や不利益が生じる場合がある。こうした点を防ぐために、属性グループ別にスコアの分布を確認し、特定のグループが不当に偏った評価になっていないかを分析することが必要になる。AIの出力についても、公平性が保たれているかを検証することが重要である。

以上のような視点で評価を実施し、その結果を検証報告書に記載の上、承認プロセスに入る。ここで挙げた検証の視点はロジック等の手法やテストの一部ではあるが、リスク格付けが「高」であることを踏まえると、ほかにもさまざまな視点から検証が必要になる。

〔図表1〕　**従来型AI（例）の概要**

名称	取引データに基づく個人の信用スコアリングAI
リスク格付け	高（AIの手法を用いており複雑であり、審査という重要な用途で活用）
使用部署	個人ローン審査部門
入力データ	預金口座の動きや属性情報、給与振り込み、公共料金支払い等
出力	信用力スコア
AIの特徴	決定木を発展させたランダムフォレスト等のAIを採用するケースが多い。ランダムフォレストは多数の決定木から構成され、個々の決定木のスコアを集約することで最終的な出力を得る。特徴量と出力の非線形な関係を表現することができる。また、オーバーフィッティングしにくく、外れ値にも強い

（出所）　KPMGジャパン作成（図表2も同じ）

〔図表2〕　**生成AI（例）の概要**

名称	融資業務における財務分析AI
リスク格付け	中（AIの手法を用いており複雑であるが、分析の参考程度の活用）
使用部署	融資部門
入力データ	分析対象企業名、決算年度、過去の財務分析結果、同業他社の財務分析結果等
出力	当該年度における財務分析結果
AIの特徴	ドラフトエージェントが入力データに基づき財務分析を実施し、ドラフトを作成。評価エージェントがレビューを行い、修正を指示。修正不要と判定されると、ドラフトエージェントが全体を統合した財務分析リポートを作成

（注）　本稿ではエージェント機能を持つ生成AIをイメージして例示している。

ブラックボックス性が一層高い生成AIの検証

一方の生成AIは、従来型AIよりさらに複雑かつブラックボックス性が高く、確立された生成AIの検証は試行が続いている。本稿では一つの考え方として、生成AIをモデルレベル（基盤モデル単体）、タスクレベル（モデルを組み込んだ特定のタスク）、システムレベル（複数のタスクを組み合わせたAIシステム）で階層的に検証するアプローチを紹介する。

従来型AIと同じく架空の例として、図表2のような生成AIをイメージする。

まずモデルレベルの検証では、ドラフトエージェントおよび評価エージェントの中核となる生成AIモデル単体の性能を評価

する。性能評価指標は、さまざまなベンチマークが存在する。例えば、幅広い学問分野や一般的な知識の理解の評価、数学の問題を解く能力の評価、コード生成能力の評価、金融データに対する回答能力の評価等である（注6）。今回は財務分析での利用が想定されるため、数学の問題を解く能力や、金融データに対する回答能力の評価に係る指標に重点を置くべきだろう。

また、リスクが高くないモデルについては、検証の負荷を減らす目的で、第三者が公開しているモデルの順位表を参照して、簡易的に検証を行うことも一案である。なお、基盤モデルの多くがベンダーモデルであるという現状を踏まえると、「使用目的に適合したベンダーモデルが選定されているか」もポイントになる。ベンダーが公表している検証報告書の確認が中心になるが、公表内容が限定的である点に留意が必要である。

以上のようなモデル自体の性能のほか、契約や利用条件（入力したデータが再学習に利用されない等）に関する評価も重要だろう。

リスクに備えて人による判断が不可欠

続いてタスクレベルの検証では、タスクを担うエージェント単位で評価を行う。モデルに加えＲＡＧ（検索拡張生成）やガードレール等の要素が適切に組み合わされて、自律的にタスクを遂行できるかを検証する。適合率、再現率、Ｆ１スコア（注7）といった従来型ＡＩでも使用されている指標に加えて、タスク固有の指標も加味して評価を行う。

今回は、過去に人が行った財務分析結果を正として、どれだけ近い出力が得られたかを意味的類似度指標で確認することが一案だろう。また、ハルシネーションのほか、タスクの内容次第では著作権やプライバシーの侵害等についても評価が必要となる。

システムレベルの検証では、ユースケース単位で評価を行う。財務分析レポートの作成というユースケースの場合、財務分析やドラフト作成、ドラフトのレビューといった個々のタスクを適切に定義できているか、それらのタスクを実行するエージェントを適切に組み合わせているかを検証する。

「期待どおりの財務分析リポートが生成されたか」といった、入力データから出力まで（End-to-End）の成功率の評価も重要である。財務分析リポートが期待外れであった場合、その原因究明のためシステムの動きを詳細に把握する必要があり、監視やログ収集機能等を有するツールを活用することは有用だろう。こうしたツールは、使用中のＡＩのモニタリングにも有効である。

以上のような階層的な検証によってリスクの軽減を図ることができる。なかでも生成ＡＩは、ユースケース単位での評価がポイントになる。だが、ＡＩの潜在的なリスクを完全になくすことは非現実的である。財務分析結果の確定や結果を踏まえた融資判断といった重要なステップには、人による判断（Human in the Loop）が欠かせない。

また、ベンダーモデルはすべての情報が開示されるわけではないため、モデルレベルの検証が限定的にならざるを得ない。モデルが頻繁にアップデートされたり、サイバー攻撃にさらされたりする可能性があることも踏まえると、使用中のＡＩモニタリングの重要性は高くなる。

さらに生成ＡＩでは、ＩＴシステム、コンプライアンス、リーガル、データ等の専門知識がより広く深くなる。ＭＲＭ部署が中心となり、これらの関連部署と連携を取りながら検証を進めることが重要である。

過度なブレーキをかけない発想を

米国では、トランプ政権の発足を受けて、規制・監督の強度が弱まる方向である。一方でＭＲＭについては、頑健なＭＲＭや独立検証の仕組みが、安心・安全なモデルやＡＩの活用に寄与しているとの声が聞かれている。

ＦＲＢ／ＯＣＣのガイダンスを土台とするＭＲＭ自体は、Ａ

Ｉ（特に生成ＡＩ）ガバナンスの最適解になるわけではない。理想は、本稿で示したような検証の視点等をＭＲＭに取り込んでいくかたちである。ＭＲＭの重要性は、ＡＩ時代においてむしろ高まると考えられる（注８）。

半面、ＡＩをＭＲＭに組み込むことのデメリットにも留意が必要である。ＭＲＭは厳格な管理を要求するために、ＡＩの活用にブレーキがかかる可能性がある。

活用とそれに伴うリスク管理のバランスを取るために、リスクの高いＡＩは厳格な独立検証を行う一方、リスクの低いＡＩは管理の強度を緩める、または管理から外すという考え方も検討に値する。生成ＡＩについては、個人業務や分析のアシスタント等に使用が限定されるのであればなおさらである。ＡＩの活用に過度なブレーキを掛けない範囲でＭＲＭに組み込む発想が求められる。

＊　＊　＊

ＡＩの進化はとどまるところを知らない。ＡＩガバナンスの観点では、伝統的なＭＲＭだけではＡＩ（特に生成ＡＩ）に対応し切れないことも事実である。関連部署との連携に加え、ＡＩの進化に合わせてＭＲＭも進化が必要だろう（将来、ＭＲＭエージェントのような概念が出現するかもしれない）。

ＡＩは金融以外でも、あらゆる業態・ビジネスで活用されている。金融やＭＲＭを超えて、さまざまな業態・ビジネスのＡＩ活用事例や、それを受けたリスク管理の理解も重要になる。本稿で紹介したＭＲＭや独立検証の視点が、金融機関のみならずさまざまな業態におけるＡＩガバナンスの進化および安心・安全なＡＩ活用の一助になれば幸いである。

（注）１　ＫＰＭＧジャパン「本邦金融機関のＡＩガバナンスの在り方」を参照。なお、ＭＲＭ全般については、田中康浩・曽我部淳編著『詳解 金融機関のためのモデル・リスク管理』（中央経済社、２０２４年）が詳しい。

２　「SUPERVISORY GUIDANCE ON MODEL RISK MANAGEMENT」（２０１１年、SR11－7）

３　Conceptual Soundness とOutcome Analysisはモデル検証の中で中核を成す概念であり、前掲注２（SR11－7）で詳しく述べられている。車に例えると、前者は車種に見合った「エンジンの設計」について評価し、後者は「車が適切に動くか」の評価といえる。

４　ＡＩ特有のリスク等に関する詳細は、「ＫＰＭＧ Ｔｒｕｓｔｅｄ ＡＩ」フレームワークを参照。

５　従来型ＡＩと生成ＡＩについては、金融庁「ＡＩディスカッションペーパー（第１・０版）」の６～７ページの記載が参考になる。

６　生成ＡＩの進化のスピードはすさまじく、それに合わせて性能評価指標も日々さまざまなものが出現している。ＡＩ開発者・検証者等の実務家は、評価に際して最適な性能指標等の最新の動向をフォローすることが重要になる。

７　適合率と再現率のバランスを示す指標。

８　最新のＦＳＢのペーパー（Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector）でもＭＲＭの重要性について触れている。

たなか　やすひろ
京都大学経済学部卒。日本銀行やボストンコンサルティンググループを経て、17年ＫＰＭＧ／あずさ監査法人。モデル管理・検証やＡＩガバナンスに関するアドバイザリー業務を統括。金融庁監督局に出向し、「モデル・リスク管理に関する原則」を策定。著書に『詳解 金融機関のためのモデル・リスク管理』（編著）や『マネロン対策等の新論点』（共著）。

たけかわ　まさひろ
大阪大学大学院工学研究科修了。信用リスク関連のデータベース機関等を経て、19年ＫＰＭＧ／あずさ監査法人。信用リスク管理領域を中心にＡＩ・機械学習等を踏まえたアドバイザリー業務を提供。本邦大手金融機関に出向し、リスク管理業務に従事。KPMG Japan Trusted AIメンバー。