GenerativeX
生成AIによるビジネスの変革を支援する

WHITE PAPER

AIのリターンを どう測るか

自律度とリスク調整後ROIで決める、
ユースケース単位のAI投資

2026年11月26日

株式会社GenerativeX

EXECUTIVE SUMMARY

エグゼクティブサマリー

80% / 37%

AIで個人の生産性が上がったと答えた回答者と、利益（EBIT）への寄与を認めた回答者

McKinsey 2026、n=1,719

56%

AIによる大きな財務上の便益がまだ見られないと答えたCEO

PwC 2026、95の国と地域の4,454人

約40%

AIで節約した時間のうち、出力の手直しで失われた割合

Workday 2025、n=3,200

01 AIの利用はどの業界でも広がったが、利益への効果を示せている企業は少ない。多くの企業が経営に報告しているのは、利用者数や操作回数と、そこから推計した削減時間である。日本企業の半数は、生成AIの効果をそもそも測っていない。

02 利用量の推計では、経営が下すべき3つの判断ができない。どの業務に予算を集めるか、試行中の取り組みを本番に上げるか、取締役会や監査、規制当局にどう説明するか、である。推計はライセンスと利用者の単位で集計され、人がAIの出力を確認し手直しする時間も、AIの誤りによる損失も含まないからである。

03 本書は、AI投資を判断する単位を、ライセンスから業務のユースケースに移すことを提案する。ユースケースごとに、AIが人の手直しなしに仕事を終えた度合いを示す**自律度**と、正味の効果を誤りのリスクで割り引いた**リスク調整後ROI**の2つを測る。同じ業務でも、利用量の推計によるROIは、リスク調整後ROIの2倍以上になりうる。

04 2つの指標の高低で、ユースケースは「広げる」「直す」「縛る」「やめる」の4つに分かれる。やめる基準を先に決めておけば、効果の出ている業務に予算を集める判断にも根拠ができる。

05 この判断を四半期ごとに下す経営の場を、本書は**AI投資委員会**と呼ぶ。当社は、この測り方を自社の稟議書の作成業務に組み込み、既存の業務ツールだけで約1日で動かした。架空の案件2件で記録から判定までを通し、判定が揺れる場面を4つ見つけて、判定のルールを足した。

CHAPTER 01

AIの効果は、何で報告されているか

この章の要点

- 個人の生産性は上がったと多くの企業が答えるが、利益への寄与を示せる企業は少ない。
- 日本企業の半数は生成AIの効果を測っておらず、測っている企業の多くは削減時間で測っている。
- 公表される効果の数字も、多くは一定の前提を置いた推計である。

個人の実感と、会社の利益の差

AIの利用は、業種を問わず当たり前になった。しかし、利用の広がりは、まだ利益の増加と結び付いていない。McKinseyの2026年の調査では、回答者の80%がAIによって個人の生産性が上がったと答えた一方、AIが利益（EBIT）に寄与したと答えた回答者は37%で、前年とほぼ同じだった[1]。PwCが95の国と地域のCEO4,454人に行った調査では、56%が、これまでのところAIによる大きな財務上の便益は見られないと答えている[2]。

成果が出ていると答える調査もある。Google Cloudの調査では、経営幹部の74%が1年以内にROIを達成したと答えた[3]。2つの結果は矛盾しない。個人の作業時間が減ることと、それが会社の利益として現れることの間には、いくつもの段階がある。その段階を測っていない企業では、効果の数字は回答者の実感に左右される。

日本企業は、効果を何で測っているか

日本企業では、効果を測ること自体が始まったばかりである。JUASの企業IT動向調査2026によれば、生成AIの効果測定を行っていない企業は51.9%に上る[4]。測っている企業の測り方は、削減できた労働時間が39.4%で最も多く、売上金額の向上などの金額で測る企業は6.5%にとどまる。PwC Japanの調査でも、生成AIの効果を「まだ評価できていない」と答えた割合は、米国の0%に対し、日本は13%だった[5]。

規制のある業界では、説明の難しさがすでに表面化している

金融業界は、この問題がすでに表に出ている例である。日本銀行が取引先の金融機関150先に行った調査では、試行中を含めて9割強が生成AIを利用している[6]。一方、金融庁のAIディスカッションペーパーは、投資対効果の説明が難しく社内の合意形成に時間が掛かるという意見が、複数の金融機関から寄せられたと記している[7]。同じ文書は、AI官民フォーラムで、DX案件の代表的なKPIだった「削減時間」が、業務プロセスそのものを変革するAIの評価に「なじまない可能性が指摘された」とも記録している。

公表される効果の数字にも、この事情が表れている。ある大手金融グループは、統合報告書に「投資効果の適切な測定方法は確立していません」と書いた[8]。別の大手金融グループは、AIの効果を削減時間から億円単位の財務効果へと開示の単位を改めたが、その数字にも「一定の前提およびロジックに基づき試算した」「みなし効果を含む」という注記が付いている[9]。開示の単位が時間から金額に変わっても、効果が推計であることは変わっていない。規制当局への説明を求められない業界でも、取締役会と株主が同じ説明を求める日は近い。

CHAPTER 02

利用量の推計では、何が決められないか

この章の要点

- 経営がAIについて下す判断は、集中投資、本番への昇格、取締役会や監査への説明の3つである。
- 利用量の推計は、集計の単位、確認と手直しの時間、誤りの損失の3点で、判断に使えない。
- 利用量の指標は定着を見るために残し、判断のための指標を別に持つ。

経営が下すべき3つの判断

AIの効果を測る目的は、報告ではなく判断である。経営がAIについて下す判断は、少なくとも3つある。どの業務にAIの予算を集め、どの業務から引き上げるか。試行中の取り組みを、部門や全社の本番運用に上げてよいか。AIが関わる業務で人がどこまで関与し、誤りをどう抑えているかを、取締役会、監査、規制当局にどう説明するか。この3つに必要な情報と、今の指標で分かることを並べると、差は明らかである。

図表1　経営の判断と、今の指標で分かること

利用量の指標は、判断に必要な情報の手前で止まっている

判断	判断に必要な情報	今の指標で分かること	指標が足りないときに起きること
集中投資	どの業務で、AIが仕事を終えているか	アプリ別、利用者別の利用回数	効果の出る業務と出ない業務に、同じ費用をかけ続ける
本番への昇格	本番に上げてよい品質と効果に達したか	昇格の基準が無い	試行が停滞するか、統制のないまま広がる
説明責任	人の関与の範囲と、誤りの抑え方	操作回数から推計した削減時間	根拠のある効果と統制を示せない

利用量の推計が判断に使えない3つの理由

利用量の推計が判断に使えないのは、数字が誤っているからではない。数字が、判断の対象とは別のものを数えているからである。

1つめは、集計の単位である。利用者数や操作回数は、ライセンスと人の単位で集計される。経営が予算を配分する対象は、稟議書の作成、提案書の作成、問い合わせへの回答といった業務である。1人の利用者は複数の業務でAIを使うので、人の単位の数字を業務に割り振ることはできない。

2つめは、確認と手直しの時間である。多くの製品は、操作の回数に1回あたりの想定時間を掛けて、削減時間を推計する。この推計には、人がAIの出力を読み、誤りを探し、書き直す時間が入らない。Workdayが3,200人に行った調査では、AIで節約した時間の40%近くが、出力の手直しで失われていた[10]。顧客に影響しうる文書では、人の確認を必須とする運用が大半である[11]。確認の時間が大きい業務ほど、推計の削減時間は実態から離れる。

3つめは、誤りの損失である。AIの出力に誤りがあれば、差し戻し、訂正、顧客への説明といった費用が生じ、誤りが顧客に届けば損失は時間では測れなくなる。カナダでは、航空会社のチャットボットが誤った運賃規定を案内した件で、会社が損害の支払いを命じられた[12]。米国では、AIの審査モデルを使った貸付業者が、州の司法長官と250万ドルで和解している[13]。推計の削減時間は、こうした損失の見込みを差し引いていない。

利用量の推計は、AIがどこで使われ、どこで定着していないかを知るには役に立つ。問題は、この数字を投資判断に使うことにある。必要なのは、利用量の指標を残したまま、判断のための指標を別に持つことである。

CHAPTER 03

判断の単位を、ライセンスからユースケースへ

この章の要点

- 経営が予算を配分する対象は業務であり、ライセンスではない。
- ユースケースの単位で測るには、1件ごとの仕事を記録する必要がある。
- 3つの測り方のうち、判断に使えるのはタスク単位の計測である。利用ログの推計とは組み合わせて使う。

ユースケースとタスク

本書でいう**ユースケース**は、AIを使って行う1つの業務を指す。社内稟議書の起案、提案書の作成、会議記録の作成、問い合わせへの一次回答などが、それぞれ1つのユースケースである。ユースケースの単位で測れば、どの業務に予算を集めるかを、そのまま比べられる。

ユースケースの単位で測るには、1件ごとの仕事（タスク）を記録する必要がある。1件の稟議書について、いつ作業を始めて、いつ終えたか、AIの下書きがどれだけ最終版に残ったか、決裁で差し戻されたかを記録する。これを業務ごとに集計したものが、ユースケースの指標になる。

3つの測り方の比較

当社は、AIの効果を測る方法を3つに分けて比べた。利用ログからの推計は導入が速く、利用の定着を見るには十分である。利用者へのアンケートは業務を選べるが、効果の根拠が本人の申告に依存する。タスク単位の計測は立ち上げに手間がかかるが、3つの判断に必要な情報を出せるのはこの方法だけである。

図表2　AIの効果の3つの測り方

判断に使えるのはタスク単位の計測。利用ログの推計は置き換えずに併用する

測り方	業務単位で測れるか	自律度が分かるか	効果の根拠	立ち上げ
利用ログからの推計 製品の標準のダッシュボード	測れない。アプリ別、利用者別まで	分からない	操作回数に想定時間を掛けた推計	設定のみ
利用者へのアンケート	回答者が業務を選ぶ	本人の印象	自己申告の時間	数週間
タスク単位の計測	**測れる。業務の区分を記録する**	**分かる。成果物と後工程から判定する**	**記録の時刻と、AIを使わない場合の基準時間の差**	1業務ずつ数週間

出典: 当社の分析

3つは置き換えの関係にない。利用ログの推計で全社の利用状況を見ながら、予算の大きい業務から順にタスク単位で測るのが現実的である。第4章と第5章で2つの指標を定義し、第6章でその測り方を、第7章で当社の実証を示す。

CHAPTER 04

自律度：AIはどこまで仕事を終えたか

この章の要点

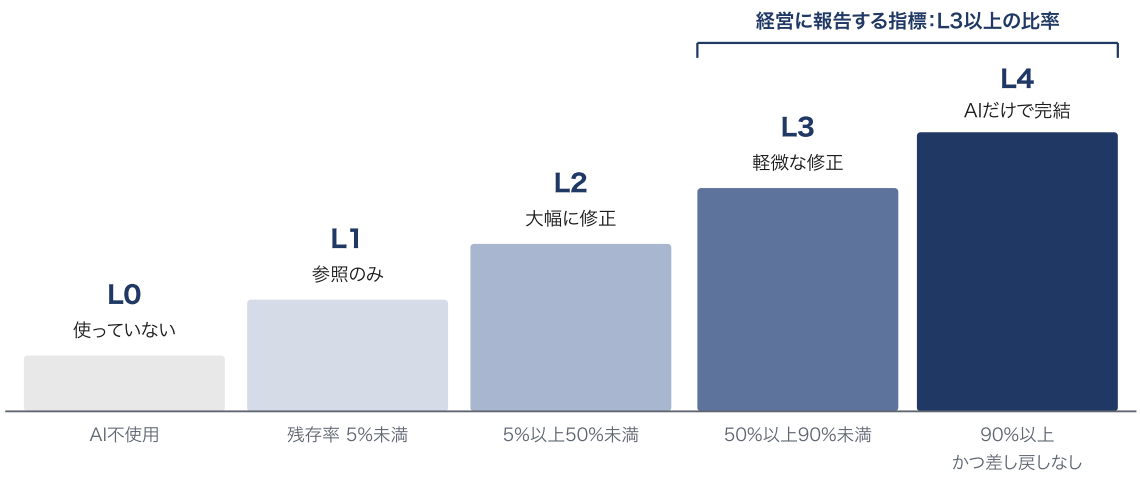
- 最終版の成果物に残ったAIの出力の割合（残存率）と、後工程の結果から、1件ごとの仕事を5段階に分ける。
- 経営に報告するのは、ユースケースごとの「L3以上の比率」である。
- 本書の自律度は、AIにどこまで任せるかという設計ではなく、実際にどこまで終わったかという結果を測る。

残存率と後工程の結果

AIが仕事をどこまで終えたかを知るには、成果物を見るのが最も直接的である。本書は、最終版の成果物に含まれるAIの出力の割合を**残存率**と呼ぶ。AIの下書きを人がほとんど直さずに使えば残存率は高く、大きく書き直せば低くなる。ただし、残存率だけでは足りない。人がほとんど直さずに出した稟議書が、決裁で差し戻されることがあるからである。そこで、成果物の後の工程（決裁、レビュー、顧客への送付）で差し戻しが無かったかを合わせて見る。

図表3　自律度の5段階

残存率と後工程の結果で、1件ごとの仕事を判定する。しきい値は初期値で、業務ごとに較正する



残存率：最終版の成果物に含まれるAIの出力の割合。L4は後工程（決裁、レビュー）の結果を待って確定する

経営に報告するのは、ユースケースごとの「L3以上の比率」である。この比率は、そのユースケースで、AIの下書きの大半を変えずに使った仕事の割合を表す。

先行する分類との違い

AIエージェントの自律性を段階で表す考え方には先行例がある。ワシントン大学の研究者は、利用者がAIエージェントに対して担う役割から、自律性を5段階に分けた[14]。シンガポールの情報通信メディア開発庁（IMDA）は、エージェント

型AIのガバナンスの枠組みで、人とエージェントの分担を4段階に分けている[15]。金融安定理事会（FSB）は、自律性の高いAIが人の介入を減らし、市場での相関を高めるおそれを指摘した[16]。これらの分類は、AIにどこまで任せるかという設計の段階を表す。本書の自律度は、実際に処理した1件ごとの仕事について、結果としてどこまでAIで終わったかを測る点が異なる。

CHAPTER 05

リスク調整後ROI：効果からリスクを差し引く

この章の要点

- 金融機関が収益から予想損失を差し引いてきたのと同じ考え方を、AIの効果に当てはめる。
- 実測の時間には確認と手直しの時間を含め、リスク割引は精度、誤りの影響、統制の3つから決める。
- 同じ業務でも、利用量の推計によるROIは、リスク調整後ROIの2倍以上になりうる。

収益からリスクを引くという考え方

金融機関は、収益を評価するときに、予想される損失を差し引いてきた。日本銀行の2001年の資料は、リスク調整後収益を「業務純益 − 予想損失」と定義し、当時すでにほとんどの大手行が何らかのリスク調整後収益指標を算出していたと記している[17]。金融庁の「モデル・リスク管理に関する原則」は、モデルに内在するリスクを評価し、その結果に基づいて管理するリスクベース・アプローチを求めている[18]。業種を問わない投資評価の実務にも、便益をリスクで割り引く方法がある。Forresterの投資効果の算定手法（TEI）は、見積もりの不確実性を反映して、便益を10%や15%下方修正する[19]。

AIの効果の評価には、この考え方がまだ持ち込まれていない。AIの出力には誤りが含まれうるので、効果からリスクを差し引かなければ、AIの効果は過大に見える。米国の銀行監督当局が2026年4月に改定したモデルリスク管理の指針は、生成AIとエージェント型AIを対象から外し、各組織のリスク管理とガバナンスで適切な統制を決めるよう求めた[20]。生成AIの効果とリスクをどう測るかは、規制の厳しい業界でさえ、各社が自ら決めるしかない状況にある。

リスク調整後ROIの定義

リスク調整後ROI ＝（リターン − インベストメント）×（1 − リスク割引）÷ インベストメント

リターンとインベストメントはユースケースごとに月次で集計する

図表4　リスク調整後ROIの3つの項と、その測り方

項	中身	測り方
リターン	削減時間の金額（基準時間 − 実測時間）× 件数 × 時間単価。成約率や処理件数の増加などの収益への寄与	タスクの記録の時刻、業務フローの時刻、業務のKPI
インベストメント	ライセンスの按分、推論と基盤の利用料、構築と運用改善の工数	利用料の請求、工数の記録
リスク割引	精度、誤りの影響、統制の3つから決める率	後工程の差し戻し率、人の採点との不一致、扱う文書の機密度、顧客に出すかどうか、権限と参照範囲

実測時間には、人がAIの出力を確認し手直しした時間を含める。こうすると、確認と手直しの費用はリターンの側で差し引かれ、二重に数えることがない。

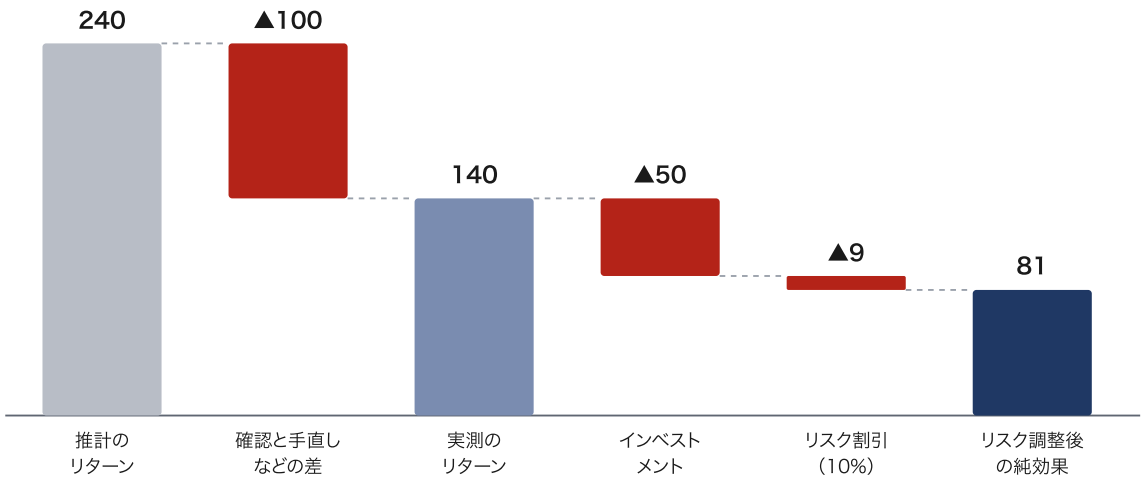
リスク割引は、精度などのリスクに応じて純効果を減額する率であり、3つの要素から決める。**精度**は、後工程の差し戻し率と人の採点との不一致で表す、誤りが起きる頻度である。**誤りの影響**は、参照した文書の機密度と、成果物を顧客に出すかどうかで表す、誤りが起きたときの損害の大きさである。**統制**は、AIに与えた権限と参照範囲で表す、誤りが広がる範囲である。初期値は10%とし、社内向けで差し戻しの少ない業務は10%前後に、顧客に出す文書や機密度の高い情報を扱う業務はより高い率に較正する。

推計と実測で、ROIはどれだけ違うか

社内稟議書の作成を例に、利用量の推計とリスク調整後ROIを比べる。月300件、時間単価8,000円とする。利用量の推計では1件あたり60分の短縮と見なされ、リターンは月240万円になる。実測では、AIを使わない場合の基準時間120分に対し、確認と手直しを含めた実測時間が85分で、短縮は35分、リターンは月140万円である。

図表5　社内稟議書の月次の効果：推計から、リスク調整後の純効果まで

単位は万円。数値はすべて説明のための架空の値。ROIは推計で3.8倍、リスク調整後で1.6倍



インベストメントの内訳：ライセンス按分25、推論と基盤5、構築と運用改善20。推計のROI＝(240−50)÷50＝3.8倍。リスク調整後ROI＝(140−50)×0.9÷50＝1.6倍

差の大部分は、確認と手直しの時間である。経営が推計の3.8倍を前提にこの業務へ予算を集め、翌年に実際の効果を測れば、期待との差を説明できなくなる。

自律度とリスク調整後ROIの関係

2つの指標は独立していない。人の手直しが減って自律度が上がると、実測時間が短くなり、リターンが増える。同時に、人の確認が減って誤りを見逃し、差し戻しが増えれば、リスク割引が上がる。自律度を上げることは、ROIを上げる方向

と下げる方向の両方に働く。2つを同時に測っていれば、自律度を上げた結果ROIが上がったのか下がったのかを、ユースケースごとに確かめられる。

CHAPTER 06

測る仕組み：1件ごとの仕事を、どう記録し判定するか

この章の要点

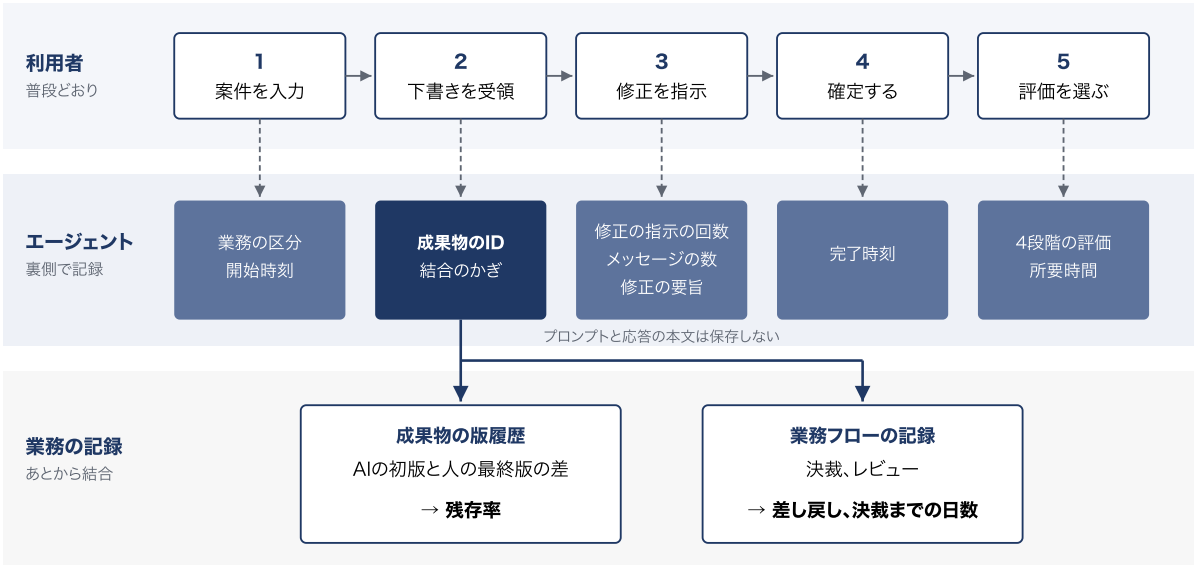
- 計測は業務で使うAIエージェントに組み込み、記録を業務の副産物として得る。利用者に新たな入力は求めない。
- 判定の中心は成果物の残存率に置く。会話の記録は補助に、後工程の結果は品質の確認に使う。
- 判定の信頼性は人の採点で、従業員の信頼はデータの扱いの原則で担保する。

記録を、業務の副産物として得る

タスク単位の計測で最初に問題になるのは、記録の負担である。利用者に作業時間や成果を毎回入力させれば、記録は長く続かず、続いたとしても自己申告の域を出ない。本書の測り方は、この負担を利用者に負わせない。計測を業務で使うAIエージェントの中に組み込み、利用者がエージェントと普段どおりにやり取りする間に、エージェント自身が1件ごとの記録を残す。記録には成果物のIDを付け、あとから成果物の版履歴と業務フローの記録に結び付ける（図表6）。こうして記録は、追加の手間なしに、業務の副産物として蓄積される。

図表6　利用者の操作と、エージェントが残す記録

利用者の画面は普段どおりの会話のまま、記録は裏側で残り、成果物のIDで後工程の記録と結び付く



業務エージェントを新たに作る段階で計測を組み込んでおけば、計測のための仕組みを後から別に作る必要はない。計測を前提にエージェントを設計することが、投資判断の材料を最も少ない費用で得る方法である。

判定の中心を、成果物に置く理由

記録は3つの層から成り、層ごとに判定での役割が異なる（図表7）。

図表7　3つの層の記録と、判定での役割

補助	中心	品質の確認
会話の進め方	**成果物**	**後工程**
往復の回数、修正の指示	最終版に残ったAIの出力の割合（残存率）	差し戻し、上長による修正

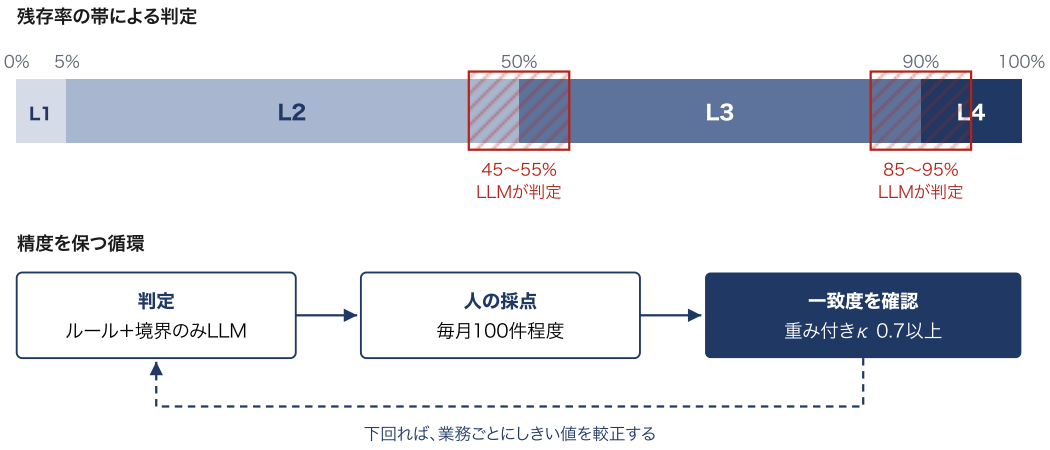
判定の中心を成果物に置くのは、会話の記録と本人の評価が、いずれも手直しの実態を映さないことがあるからである。会話の回数は、利用者の使い方の癖によって増減する。本人の評価は作業を終えた直後の印象であり、その後に加わった手直しを含まない。これに対し成果物は、人がAIの出力をどれだけ受け入れたかを、結果として記録している。後工程の結果を加えるのは、受け入れた出力が業務の品質を満たしていたかを確かめるためである。

判定の信頼性をどう担保するか

判定は、残存率の帯にもとづくルールを基本とする。ただし、しきい値の近く（45%から55%、85%から95%）では、数ポイントの差で段階が入れ替わり、ルールだけでは判定が安定しない。この区間の件に限り、大規模言語モデルがAIの初版と人の最終版の差分を読み、修正が判断の中身に及んでいるかを見て判定する（図表8）。

図表8　判定の仕組みと、精度を保つ循環

ルールで判定し、境界の近くだけを大規模言語モデルが判定する。人の採点で毎月精度を確かめ、外れればしきい値を較正する



判定そのものの精度は、人の採点で検証する。業務部門の担当者が毎月100件程度を採点し、判定との一致度（重み付きκ）が0.7を下回らないことを確かめる。下回った場合は、その業務のしきい値を較正し直す。経営に報告する数字を、機械の判定だけに依存させない。この構造が、自律度を投資判断の根拠として使い続けるための条件になる。

従業員の信頼をどう確保するか

計測は、従業員の仕事の中身に踏み込む。従業員がこれを監視と受け取れば、AIの利用そのものが減り、測るべき対象が失われる。データの扱いは、計測を始める前に原則として定め、従業員に開示しておく必要がある（図表9）。

図表9　計測のデータの扱いに関する4つの原則

01
自社の環境の中で完結させる
計測の処理はすべて自社の環境の中で行い、外部の事業者の環境にはデータを送らない

02
本文を残さない
プロンプトと応答の本文は評価したあと破棄し、スコアと根拠のハッシュだけを保存する

03
個人を判定しない
判定の単位は業務のタスク。集計はチーム単位、担当者は仮名化、最小集計人数は10人

04
小さく始められる
本文を参照する権限が得られなくても、メタデータだけで始められる。その場合、判定は2区分程度に粗くなる

計測の目的が業務への投資判断であり、個人の評価ではないことを明文化しておけば、この原則は従業員への説明にもそのまま使える。

経営に何を報告するか

計測の結果は、経営会議では1枚に集約する（図表10）。上段には、L3以上の比率、実測の削減時間、純効果、リスク調整後ROIの4つを置き、実測の削減時間の横には、利用ログから推計した削減時間を並べる。推計と実測の差を毎回示し続けることで、経営の議論は推計の数字から離れていく。中段にはAI投資マップ（第8章）を、下段にはユースケースごとの純効果と次の四半期の判断を置く。

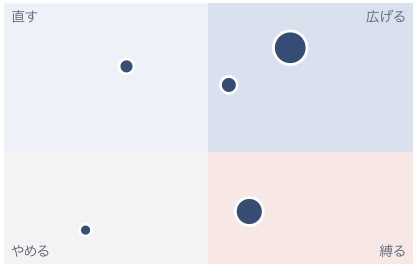
図表10　経営会議に報告する1枚の構成

四半期ごとに、この1枚でユースケースごとの投資の継続を判断する。数値の欄は構成を示すための伏せ字

経営サマリー（四半期）

L3以上の比率	実測の削減時間	純効果	リスク調整後ROI
○○%	○○時間　推計 ○○時間（推計と並べて表示）	○○万円	○.○倍

AI投資マップ



ユースケースごとの純効果と判断

ユースケース	件数、L3以上	純効果	判断
会議と面談の記録	○○件、○○%	○○	広げる
社内稟議書	○○件、○○%	○○	広げる
顧客宛てメールの下書き	○○件、○○%	○○	縛る
調査レポート	○○件、○○%	○○	直す
定型データの集計	○○件、○○%	○○	やめる

判断に注意が必要な項目（統制の強化、見直し）は色で示す

部門長向けには業務ごとの1枚を別に作り、起案から決裁までの日数と差し戻し率を組で示す

部門長には、業務ごとの1枚を別に用意する。この1枚では、起案から決裁までの日数と差し戻し率を組にして示し、判定に注意が必要な件（差し戻された件、評価と残存率が食い違った件）を一覧にする。速さと品質を同じ紙面で見せることで、速さだけを追う改善を防げる。

CHAPTER 07

当社での実証：自社の稟議書業務で何が分かったか

この章の要点

- 当社は自社の社内稟議書の作成業務に第6章の計測を組み込み、既存の業務ツールだけで約1日で稼働させた。
- 架空の案件2件で、記録から自律度の判定までを通し、L4とL3の判定を得た。
- 判定の流れを確かめる中で、判定が揺れる場面が4つ見つかり、判定のルールを足した。

実証の設計

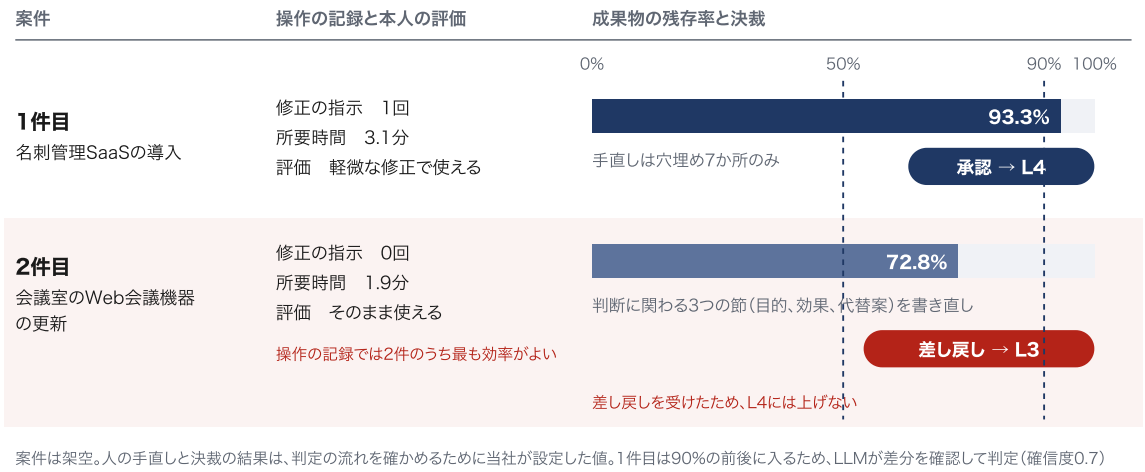
当社は2026年11月、第6章の測り方を、自社の社内稟議書の作成業務で検証した。検証の目的は2つある。1つは、計測の仕組みが既存の業務ツールの範囲で動くかを確かめることである。もう1つは、自律度の判定の考え方に、机上では見えない弱点がないかを洗い出すことである。

稟議書の下書きを作るエージェントはMicrosoft Copilot Studioで構築し、成果物は版履歴の残るSharePointの文書ライブラリに、1件ごとの記録はSharePointのリストに保存した。データ基盤の追加も個別の開発も行わず、構築に要した期間は約1日である。案件は架空とし、人の手直しと決裁の結果は、判定の流れを確かめるために当社が設定した。したがって本章の数字は、仕組みの動作と判定の妥当性を示すものであり、効果の大きさを示すものではない。

2件の判定結果

図表11　2件の稟議書の記録と判定

操作の記録で見ると2件目のほうが効率がよい。成果物と決裁の結果では、2件目に大きな手直しと差し戻しがある（案件と結果は当社が設定した値）



案件は架空。人の手直しと決裁の結果は、判定の流れを確かめるために当社が設定した値。1件目は90%の前後に入るため、LLMが差分を確認して判定（確信度0.7）

1件目は、想定どおりの判定である。人の手直しは起案者名や日付の穴埋めにとどまり、残存率は93.3%、決裁も承認された。残存率がしきい値の90%に近いため、大規模言語モデルが差分を確認し、修正が判断の中身に及んでいないこと

からL4と判定した。

2件目は、測り方の違いが、そのまま判定の違いになった例である。操作の記録では、修正の指示は0回、所要時間は1.9分と、2件のうちで最も効率がよい。本人の評価も「そのまま使える」である。利用ログや利用者アンケートで測れば、この件はAIが最もうまく仕事を終えた件として計上される。しかし成果物を見ると、稟議の判断に関わる3つの節（目的、効果、代替案）が書き直されており、決裁では差し戻された。残存率は72.8%でL3にとどまり、差し戻しを受けたためL4には上げない。成果物と後工程を見て初めて、この件を正しく位置付けられる。

実証で見つかった判定の揺れ

実証では、判定が揺れる場面が4つ見つかった。いずれも、机上で設計した判定のルールでは扱いが決まっていなかった場面であり、それぞれにルールを追加した（図表12）。

図表12　実証で見つかった判定の揺れと、追加したルール

判定の段	見つかったこと	追加したルール
残存率	判断に関わる節を書き直しても、文書全体の残存率は70%台に留まった	判断に関わる節の残存率を、全体とは別に出す
後工程	決裁で差し戻された件が、残存率の帯ではL3と判定された	差し戻しの件は1段下げる案を、較正の中で検討する
自律度	起案者名や日付を埋めただけの件が、L3とL4のどちらにも読めた	穴埋めだけの修正は、L4の妨げにしない
人の採点	利用者が「そのまま使える」と評価した件を、あとで大きく直していた	自己評価と残存率が食い違う件は、人が採点する

4つのうち、判定の考え方に最も関わるのは1つめである。稟議書では、判断の根拠となる節が数行であっても、そこを書き直せば文書の結論が変わる。しかし文書全体の残存率だけを見ると、この書き直しは数字にほとんど表れない。判断に関わる節を別に測ることで、手直しの量だけでなく、手直しの重さを判定に反映できる。

実証から導かれる3つの示唆

01　計測に大がかりな基盤は要らない

既存の業務ツールの範囲で、1件ごとの記録から自律度の判定までを動かせた。計測を始めるうえで先に決めるべきは、技術の選定よりも、どの業務から測るかである。

02　操作の記録と本人の評価だけでは、手直しの量と品質を見誤るおそれがある

2件目のように、操作の記録では効率がよく見える件でも、成果物には大きな手直しが入っていることがある。投資判断の根拠には、成果物と後工程の結果を使う必要がある。

03 判定のルールは、実際の業務で較正して初めて定まる

4つのルールは、いずれも実際に処理した件から見つかった。自社に導入する企業も、最初の1業務でルールを較正してから対象を広げれば、判定の揺れを全社に持ち込まずに済む。

CHAPTER 08

AI投資マップ：広げる、直す、縛る、やめる

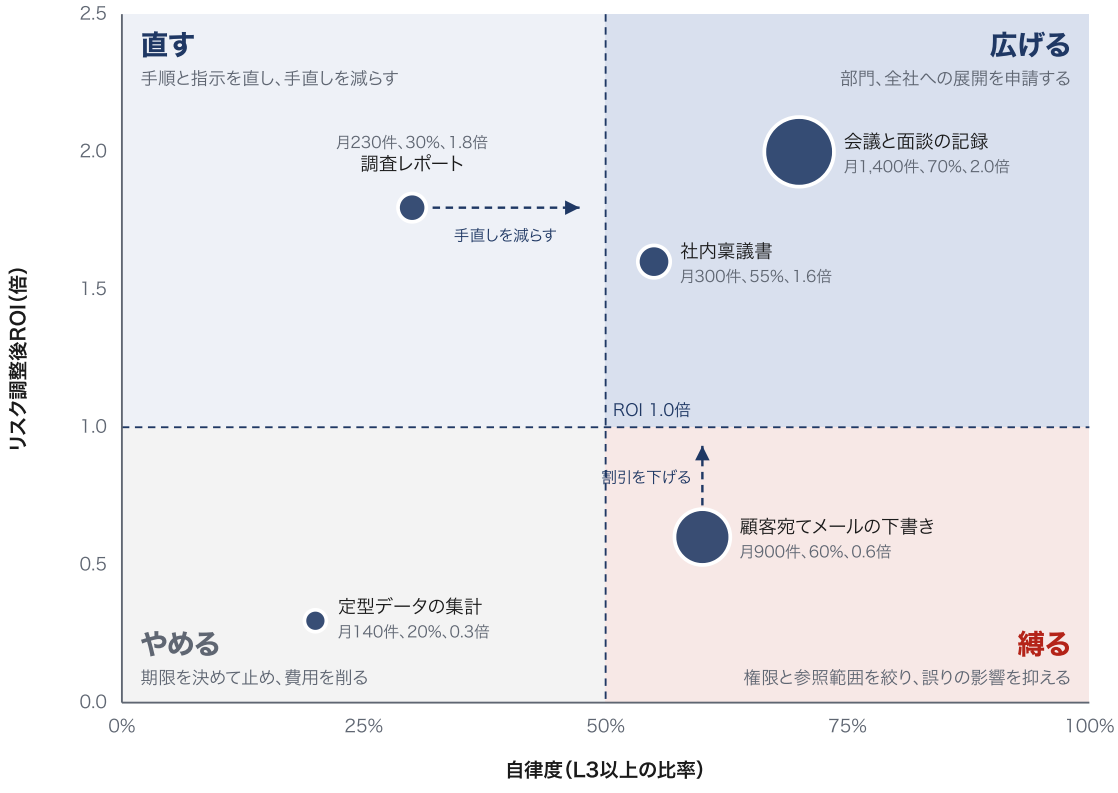
この章の要点

- 自律度とリスク調整後ROIの2軸で、ユースケースを4つの区分に分ける。
- 区分ごとに、次の四半期にすることが決まる。
- やめる基準を先に決めておけば、広げる判断にも根拠ができる。

ユースケースを、横軸に自律度（L3以上の比率）、縦軸にリスク調整後ROIをとった平面に置く。本書はこの図を**AI投資マップ**と呼ぶ。区分の切り目は、自律度50%、ROI1.0倍を初期値とし、四半期ごとに見直す。

図表13　AI投資マップ

あるコーポレート部門の5つのユースケースを置いた例。円の面積は月間の件数。点線の矢印は次の四半期に目指す動き。数値はすべて架空



4つの区分と、次の四半期にすること

広げるは、AIが仕事を終えていて、リスクを差し引いても効果が出ている業務である。昇格の条件を満たせば、部門や全社への展開を申請する。**直す**は、効果は出ているが人の手直しが多い業務である。AIの下書きが業務の型に合ってい

ないことが多く、指示や参照する資料を直せば自律度が上がる。図表13の調査レポートがこれに当たる。

縛るは、AIだけで仕事が終わっているのに、誤りの影響が大きいためにリスク割引が高い業務である。顧客に出す文書や機密度の高い情報を扱う業務が入りやすい。図表13の顧客宛てメールの下書きは、L3以上の比率が60%と高い一方、顧客に出る文書であるためリスク割引を30%と置き、ROIが0.6倍にとどまった。利用量の推計だけを見れば「全社に広げる」と判断しかねない業務である。この区分の業務は止めるのではなく、AIが参照する範囲や送付前の確認を設計し直して、リスク割引を下げる。**やめる**は、どちらの指標も低い業務である。期限を決めて止め、費用を効果の出ている業務に回す。

やめる基準が、広げる判断を支える

AIの取り組みをやめること自体は、多くの企業がすでに行っている。S&P Global（451 Research）の調査では、AIの取り組みの大半を本番前にやめた企業の割合が、1年で17%から42%に上がった[21]。しかし、やめる基準を事前に決めている企業は少ないとみられる。基準が無ければ、やめる判断は予算の締め付けや担当者の異動のたびに、場当たりに下される。

やめる基準があれば、広げる判断も通しやすくなる。経営は、効果の出ない業務を止められると分かっていれば、効果の出ている業務に予算を集めることをためらわない。金融庁のAIディスカッションペーパーも、将来収益が得られる案件ではROI等の指標で評価し、進捗に応じて資源の増強やサービスの縮小と撤退を判断するPDCAを回す必要があると述べている[7]。

CHAPTER 09

AI投資委員会：誰が、いつ、何を決めるか

この章の要点

- 月次の採点、四半期の投資判断、都度の昇格審査の3つの場で決める。
- 効果の責任は業務部門、測り方はDX推進部門、リスク割引の基準はリスク管理部門が持つ。
- 導入時に想定される論点には、計測を始める前に対応を決めておく。

3つの場

AI投資マップにもとづく判断は、頻度の異なる3つの場で行う。中心は四半期の投資判断であり、月次の採点がその数字の信頼を支え、都度の昇格審査が展開の速さを保つ。

図表14　AI投資委員会の3つの場

中心は四半期の投資判断。昇格の条件の数値は初期値で、最初の業務で較正する

判定の採点 月次	投資判断 四半期	昇格審査 都度
誰が 業務部門の担当者	**誰が** DX推進部門が起案し、経営会議が決める	**誰が** 業務部門が申請し、リスク管理、情報セキュリティ、人事が承認する
何を決めるか 判定の精度が保たれているか。しきい値の較正	**何を決めるか** ユースケースごとに、広げる、直す、縛る、やめるのどれか	**何を決めるか** 試行から部門展開へ、部門展開から全社展開への昇格
使う情報 人の採点と、判定との一致度	**使う情報** AI投資マップ	**昇格の条件の例** L3以上の比率50%以上、差し戻し率が導入前以下、リスク調整後ROI1.5倍以上、判定と採点の一致度が基準以上

役割の分担

効果の責任は、業務部門が持つ。AIを使ってその業務の成果を出すのは業務部門であり、AI推進の部門ではないからである。DX推進の部門は、測り方と道具を用意し、四半期の投資判断を起案する。リスク管理の部門は、リスク割引の基準を決め、昇格を承認する。この分担によって、リスク管理の部門は止める役から、どの条件なら広げてよいかを決める役に変わる。

導入時に想定すべき論点

タスク単位の計測は、導入の段階で4つの方面から異論や障害に直面しうる。データの取得、現場の受け止め、数字の信頼性、既存の製品との関係である。いずれも、計測を始める前に対応を決めておけば、多くは避けられる（図表15）。

図表15　導入時に想定すべき論点と、その対応

論点	想定される事態	対応
データの取得	本文を参照する権限が得られない	メタデータだけで開始する。本文を扱う場合も、自社の環境の中で評価して破棄する
現場の受け止め	従業員が監視と受け取る	チーム単位で集計し、担当者を仮名化し、最小集計人数を10人とする。計測の目的を明文化して開示する
数字の信頼性	指標を良く見せるための操作が起きる	速さの指標と品質の指標（差し戻し率）を必ず組で評価する
	判定の精度が不足する	毎月の人の採点、一致度の開示、業務ごとの較正を続ける
既存の製品との関係	製品の標準の指標と重複する	標準の指標は利用状況の把握に使い続け、自律度と金額への換算に役割を絞る
	業務エージェントの利用が、標準の指標に表れない	業務エージェントを、利用状況が集計される環境に公開する

中でも数字の操作は、指標を予算配分に結び付けた時点で起きやすくなる。速さの指標だけを見ていれば、確認を省いて早く出すことが部門にとって得になるからである。速さの指標を差し戻し率と組にしておけば、確認を省いた分は差し戻しの増加として表れ、操作の余地は小さくなる。

指標を使う範囲

リスク調整後ROIは、最初から業績評価に結び付けないほうがよい。日本銀行の2001年の資料も、大手行がリスク調整後収益指標を算出しながら、業績評価への反映は試行的な段階にあると記している[17]。新しい指標は、まず予算配分の参考として使い、数四半期の記録で信頼を得てから、評価への反映を検討する。金額に換算しにくい効果（顧客の満足、従業員の負担の軽減、新しい業務への挑戦）は、ROIに無理に含めない。こうした効果を持つユースケースや、戦略上続けると決めた取り組みは、マップの外で経営が判断し、その理由を記録して次の四半期に見直す。

1つの業務から始める

AI投資委員会は、全社の仕組みを整えてから始める必要はない。件数が多く、決裁やレビューといった後工程の記録が残る業務を1つ選び、自律度とリスク調整後ROIを出す。第7章の実証でも、架空の案件で判定の流れを確かめた段階で、ルールを4つ足す必要があった。判定のルールは、実際の業務で較正しながら定めていくものである。1つの業務で結果が出れば、経営は次の四半期から、その数字で判断を始められる。測る業務を増やすほど、業務の種類ごとの基準値がたまり、次の業務は速く測れるようになる。

最初の1つの業務を測る前に決める3つのこと

01 どの業務から測るか。件数が多く、決裁やレビューの記録が残る業務が向いている（例：提案書の作成、社内稟議書）。

02 後工程の記録を参照する権限。決裁やレビューの結果が見られなければ、L4の判定は確定しない。

03 業務エージェントをどこに公開するか。利用状況が集計される環境に置けば、利用量の指標とタスク単位の記録を並べて見られる。

おわりに

AIのリターンをどう測るか。本書の答えは、利用量ではなく業務を単位とし、AIが仕事を終えた度合いと、リスクを差し引いた正味の効果を測ることである。この2つがそろえば、経営はAIの予算を、広げる業務、直す業務、縛る業務、やめる業務に振り分けられる。推計の数字で予算を配り続けるかぎり、効果の出ている業務とそうでない業務は見分けられない。まず1つの業務で測り、次の四半期の判断をその数字で下すことが、AI投資を判断できる投資に変える最初の一歩になる。

GenerativeXのご支援

GenerativeXは、本書で示した測り方を企業の業務に組み込み、AI投資委員会を運営できる状態にするまでを支援しています。

判断の基準
自社の業務に合わせた自律度の判定基準と、リスク調整後ROIの算定方法

計測の仕組み
業務で使うAIエージェントへの計測の組み込みと、経営向けの報告

部門への定着
業務部門での運用と、四半期ごとの投資判断の支援

最初の1つの業務から始められます。自社のAI投資をユースケースの単位で見直したいとお考えの方は、お気軽にご相談ください。

ABOUT

GenerativeXについて

株式会社GenerativeXは、生成AIのコンサルティングと開発を行う会社です。金融、保険、製薬、製造、商社などの大手企業とともに、AIエージェントの設計から本番の運用、社内での定着までを手がけています。本書で示した測り方は、当社が自社の業務で検証を進めているものです。

お問い合わせ: https://gen-x.co.jp

出典

1. McKinsey & Company, "The state of AI in 2026: On the road to ROI"（2026年8月25日）。97か国1,719人、調査期間2026年5月4日から6月8日。
2. PwC「第29回世界CEO意識調査」（2026年1月19日）。95の国と地域のCEO4,454人、調査期間2025年9月30日から11月10日。
3. Google Cloud, "The ROI of AI 2025"（2025年9月4日）。24か国の経営幹部3,466人。
4. 日本情報システム・ユーザー協会「企業IT動向調査2026」p.42。調査期間2025年9月5日から10月24日。
5. PwC Japan「生成AIに関する実態調査2026春」（2026年6月10日）。日本の回答者932人。
6. 日本銀行「金融機関における生成AIの利用状況」（2026年8月）。取引先金融機関150先。利用中79.3%、試行中11.3%。
7. 金融庁「AIディスカッションペーパー（第1.1版）」（2026年3月3日公表）p.28から29。
8. ある大手金融グループの統合報告書2026、p.53。
9. ある大手金融グループのIR資料（2025年3月、2026年9月）。
10. Workday（Hanover Research実施、2025年11月）。売上1億ドル以上の組織でAIを使う従業員3,200人。
11. 金融庁、前掲（注7）p.18。稟議書のドラフトなど、顧客に影響しうる文書の作成業務について。
12. Moffatt v. Air Canada, 2024 BCCRT 149（2024年2月14日）。
13. マサチューセッツ州司法長官室の発表（2025年7月10日）。和解の対象には、AIモデル以外の違反も含まれる。
14. K. J. Kevin Feng, David W. McDonald, Amy X. Zhang, "Levels of Autonomy for AI Agents"（2025年）。
15. Infocomm Media Development Authority, "Model AI Governance Framework for Agentic AI", Version 1.5（2026年5月20日公表、6月5日更新）。
16. Financial Stability Board, "Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector"（2025年10月10日）p.16から17。
17. 日本銀行「金融機関における統合的なリスク管理」（2001年6月8日）p.24から25。
18. 金融庁「モデル・リスク管理に関する原則」（2021年11月12日）p.5、p.7。
19. Forrester, "The Total Economic Impact Of TEI"（2022年7月）。
20. Board of Governors of the Federal Reserve System, OCC, FDIC, SR 26-2 "Revised Guidance on Model Risk Management"（2026年4月17日）添付 脚注3。
21. S&P Global Market Intelligence（451 Research）。2024年10月21日から11月25日、n=1,006。