baycurrent

INSIGHTS

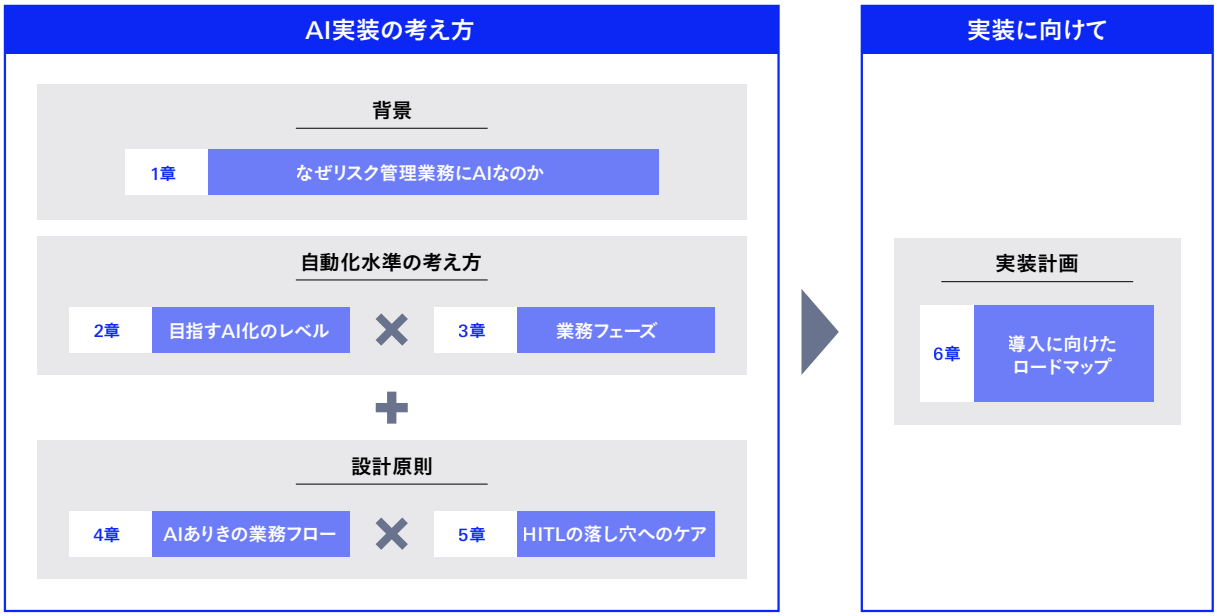


高度化する不正にAIで立ち向かう

～金融・決済業界で成果を分ける「自動化水準」と「業務の再設計」～

決済の多様化とデジタル取引の拡大により、リスク管理が向き合う取引の「量」は増え続けている。一方、生成AIを悪用したなりすましなど手口の巧妙化により、不正の「質」も高度化している。ルール追加と目視審査を積み上げる人手中心の運用では、アラートの増加が判断の遅延を招き、限界は目前である。本論考では「自動化水準の明確化」と「業務・統制の再設計」を、AI導入の成否を分ける要点として考察する。

図1　本論考の構成



はじめに

決済サービスの多様化やデジタル取引の拡大を背景に、不正取引の手口は急速に高度化している。フィッシング、アカウント乗っ取り（ATO）、SIM スワップやディープフェイクによるなりすましといった攻撃手法は機械学習や生成 AI を悪用した形にまで進化しており、従来のルールベースの検知モデルでは対応の限界が現実のものとなりつつある。

こうした環境変化の中、決済事業者のリスク管理部門には、不正検知の精度向上と業務処理の効率化という二つの命題に同時に応えることが求められている。その有力な手段として注目されるのが AI（機械学習・深層学習）の活用であるが、導入の検討が進む一方、「期待した成果が得られない」「運用に乗らない」という事例も少なくない。

本論考では、金融・決済業界におけるリスク管理業務を対象に、AI 導入の必要性を改めて整理するとともに、導入後の成果を左右する実務上の論点を、コンサルティング実務の知見をもとに提示する。リスク管理部門・IT 部門の役職者が、AI 導入の方向性と優先課題を判断するうえでの一助となれば幸いである。

要旨：AI導入の成否は「自動化水準」と「業務再設計」で決まる

リスク管理業務における AI 導入の成否は、「何をどこまで AI 化するか（自動化水準の明確化）」と、「AI を前提として業務と統制を作り替えられるか」の二点に集約される。（図 2）

既存フローの一部を AI に置き換えるだけでは、業務の複雑性・アラート過多・属人化が温存され、導入後に期待される効果が得られないケースが多い。

AI 導入の要諦は、人・AI・統制の分業を再設計し、HITL（Human in the Loop）を含む新しい業務設計を行うことにある。HITL とは、AI や自動化システムのプロセスに人が介在し、監督・承認・修正を行う仕組みであり、単なるツールの追加ではなく、意思決定と統制の構造そのものを問い直す取り組みとして位置づけることが出発点となる。

図2　自動化水準と業務再設計

何をどこまでAI化するか（自動化水準の明確化）

- Lv1：人の判断の補助にAIを活用する
- Lv2：業務のうちの、特定のタスクをAIが処理する
- Lv3：AIによって判断とアクションを実行する
- Lv4：関連業務を含む、入口から事後・改善までをAIで最適化する

AIを前提として業務と統制を作り替える

- AI導入に合わせ、入口・取引時・事後・改善フェーズの強度のバランスを見直す
- 参照元や判断根拠、記録様式等を標準化し、基盤を整えたうえでAIを導入する
- 審査の一連のフローはAIに置換し、人はAIのチューニングや監視、KPI管理や事故対応に集中する 等

なぜ今、リスク管理業務にAIが必要なのか

リスク管理業務における従来型のルール運用や人手中心の目視審査は、複数の構造的な限界に直面している。

第一に、決済チャネルの拡大やサービス形態の多様化を背景に、監視対象となる取引量は増加の一途をたどっている。こうした取引の増勢と並行して、不正利用による被害も増加傾向にある。その結果、リスク管理業務において判断すべき変数の種類と量は拡大しており、人手のみで取引を網羅的にモニタリングすることは実質的に困難になっている。（図 3）

第二に、攻撃側でも AI の活用が進み、不正手口の探索・試行・検知回避のサイクルが高速化している。特定の手口に対してルールを追加しても、攻撃者はそのルールをすり抜ける新たなパターンを短期間で生み出している。防御側がルール追加で後追いし、目視審査を重ねる従来の構造は、業務の属人化を招くだけでなく、不正被害に対して構造的に後手に回りやすい。

＜例＞

- ◆ 攻撃者は対策の抜け穴を見つけた場合、早ければ数分 ~ 数時間の間に複数の決済をおこない資金流出を図る。
- ◆ 攻撃者は対策の抜け穴が塞がった後、数時間以内には攻撃手法を変え別の抜け穴を探す動きをとる。

第三に、堅牢性を高めようとした際のトレードオフが深刻化している。過剰なルールを設定すれば、誤検知の増加によるアラート過多と顧客体験の毀損に直結する。逆に審査を緩めれば不正被害が拡大する。このトレードオフを人手とルールだけで管理し続けることには、もはや限界がある。

第四に、業務の分断がリスク管理の品質を損なっている。KYC・法人審査・与信調査・取引調査・補償判断といった業務が組織的・システム的に分断されている場合、同じ顧客・同じ事象に対して判断基準が揺れ、説明責任の観点でも弱点となる。

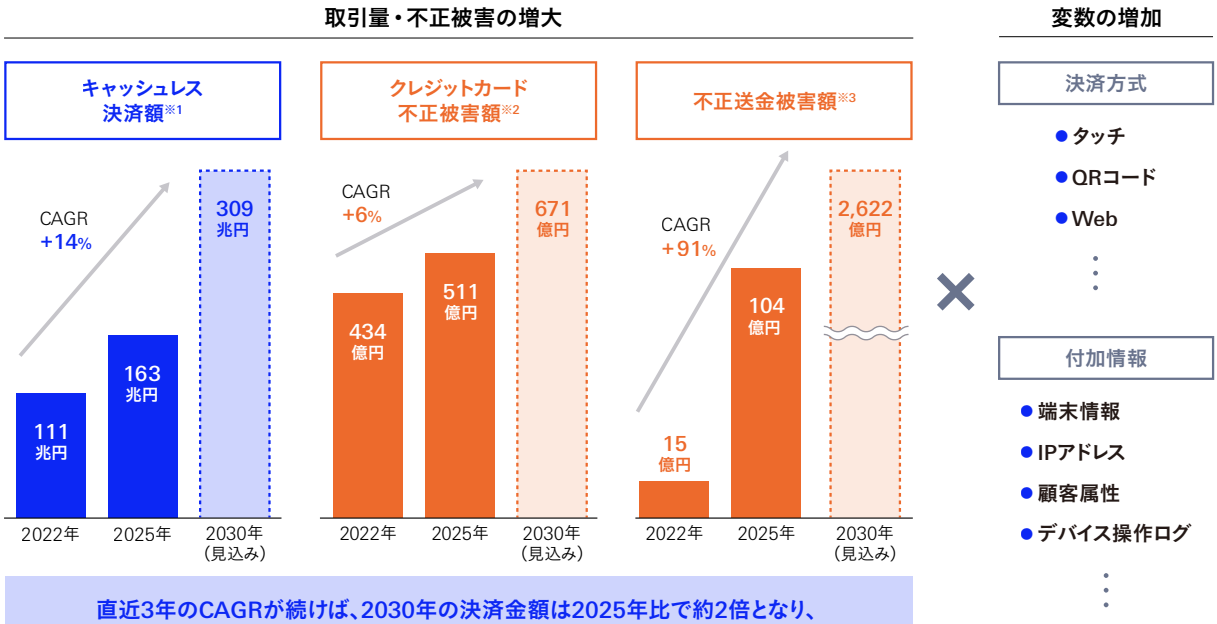
規制や社会要請が強まる中、再現性のある判断・即時性の確保・誤検知の抑制・監査に耐える証跡の整備が同時に求められている。AI はこれら複数の命題に対応するための有力な手段であり、もはや「検討すべき選択肢の一つ」ではなく、「早期に実装を判断すべき経営課題」と位置づけるべき段階にある。

「どのレベルのAI化を目指すか」で設計が根本から変わる

AI 導入の議論が停滞する典型的なパターンは、「とりあえず生成 AI を試してみる」ことが目的化し、どこまで自動化するか、自動化水準が曖昧なまま PoC が積み上がるケースである。

リスク管理業務は意思決定の連続であり、判断には責任が伴う。したがって、AI 化の範囲・深度によって、得られる効果、必要なデータ、求められる統制・運用体制、設定すべき評価指標が根本から変わる。目指すべき水準を最初に整理しない限り、導入は部分最適の集合となり、運用負荷と監査負荷が増大し、想定した

図3　リスク管理業務を取り巻く外部環境



出所：※1 METI/経済産業省｜2025年のキャッシュレス決済比率を算出しました　※2 日本クレジット協会｜日本のクレジット統計2025年版
※3 警察庁｜令和7年における サイバー空間をめぐる脅威の情勢等について

効果は発現しない。

以下では、AI 化を四つのレベルに分類し、各レベルにおける導入効果と実務上の論点を考察する。（図 4）

（ア）レベル 1：支援

AI が人の判断を補助するレベル。生成 AI の業務利用が代表例となる。

＜例＞

◆ 審査担当者が記録するコメントの下書きを生成

◆ 過去事例や規程を横断検索し、要点を提示

◆ 問い合わせに対する一次回答案を作成

これらは、意思決定そのものを AI に委ねるのではなく、判断の前後にある「調べる」「書く」「整理する」作業を圧縮する取り組みである。活用範囲は限定的である一方、導入難易度が低く、効果が早期に得られる点に価値がある。

ただし、生成 AI は誤った内容をもっともらしく生成するリスクを内包しており、根拠が曖昧なまま業務に紛れ込みやすい。レベル 1 であっても、機密情報の取り扱い、ログ管理、参照元の制御、出力の利用ルールを整備しなければ、情報漏えいや誤判断の温床となる。支援レベルの価値は即効性にあるが、「安全に使うための基盤づくり」を先送りしないことが、次のレベルへの布石でもある。

（イ）レベル 2：部分自動化

業務の特定タスクを AI が処理するレベル。

＜例＞

◆ 書類不備の検知

◆ 取引スコアリング

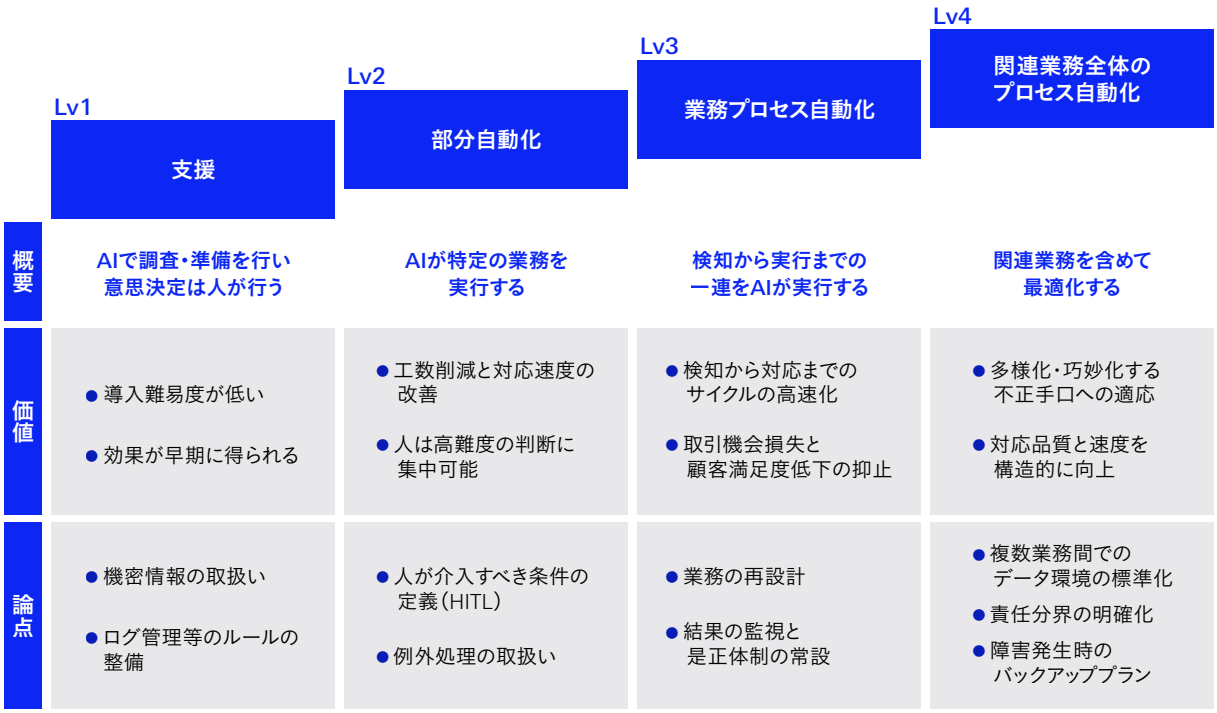
◆ アラートの優先度付け

◆ 他部署照会への自動回答

レベル 2 の価値は、工数削減と対応速度の改善にある。AI が業務処理の一部を担うことで、人が高難度の判断に集中できる環境が生まれる。

ただし、効果の最大化と堅牢性・説明責任の担保には、AI の処理結果を人がどう扱うかという設計、すなわち HITL（Human in the Loop）の設計と例外処

図4　AI化の4レベル



理の取り扱いが鍵となる。AIのスコアが高いものだけを人が確認する運用にすれば工数は削減できるが、例外や境界領域が増えるほど、担当者は「結局、全件確認したほうが安心」という方向に引き戻されやすい。レベル2では、人が介入すべき条件を明確に定義し、誰がどの閾値で何を判断するか、例外時のエスカレーション経路を規定することが不可欠である。

（ウ）レベル3：業務プロセス自動化

導入効果が質的に変わるレベル。AIは単なるスコア算出にとどまらず、判断とアクションの連鎖に組み込まれる。

＜例＞

◆ 検知・審査・対応（取引の制限・解除など）の一連の処理を自動的に実行

◆ 対象顧客への通知文を生成し、通知を実行

レベル3の価値は、損失抑制と顧客体験の両立にある。検知から対応までのサイクルが高速化することで、取引機会の損失・不正被害・顧客満足度の低下を同時に抑制できる。これは単なる工数削減を超えた、事業上の競争力となる。

一方で、レベル3は現行手順の単純な置換では成立しない。既存の業務フローをそのまま自動化しようとすると、人の判断・例外処理・証跡の作り方といった暗黙の前提とAIの特性が噛み合わず、精度・統制・運用の各面で不具合が生じやすい。また、責任分界を曖昧にしたまま自動化を進めれば、事故発生時に「誰が何を承認したのか」が追えず、監査・当局対応上の重大なリスクとなり得る。判断・統制・運用・データの各観点から業務を再設計し、モデル責任者・業務責任者・承認者・監査担当者の役割と権限を明確にしたうえで、結果の監視と是正の体制を常設することが前提となる。

（エ）レベル4：関連業務全体のプロセス自動化

複数のレベル3プロセスを接続し、関連業務を含む入口から事後・改善までを線・面として最適化するレベル。

＜例＞

◆ 審査業務（検知・審査・対応）に続き、疑わしい取引の届出業務（書類作成・承認後の提出）、検知ロジックの見直しまでの一連の流れを自動的に実行

レベル４は、工数削減にとどまらず、多様化・巧妙化が進む不正手口に対して、対応の品質と速度を構造的に向上させることを可能にする。

ただし、導入のハードルは他のレベルと比較して格段に高い。第一に、関連業務間で業務プロセスおよびシステム環境を統合する必要があり、データ定義・ID 体系・ケース管理の粒度が不整合であれば、一気通貫の自動化は成立しない。第二に、複数の所管部署が関与するため、責任分界が不明確になりやすい。第三に、適用範囲が広範なだけに、システム障害発生時にリスク管理業務全体が停止するリスクがある。

したがって、レベル４の実現には、共通 ID やデータ基盤を含む環境の標準化、業務プロセスの再設計、統制設計（責任分界の明確化を含む）、監視体制の構築、および障害発生時を想定したバックアッププランの整備が不可欠である。導入負荷は大きいものの、攻撃側の手口の増加・多様化という本質的な課題に対して、対症療法ではなく構造的な解決を図り得る点で、中長期的に目指すべき水準と位置づけられる。

業務フェーズとAI導入の設計――入口から改善まで一気通貫で考える

レベル４の AI 化を志向する場合、個別タスクの自動化だけでは業務の分断が残り、損失や顧客影響の全体最適につながりにくい。ただし、全ての組織が直ちにレベル４を目指す必要があるわけではない。自社のリスク特性、業務規模、データ基盤や統制の成熟度を踏まえ、段階的に適用範囲を定めていくことが重要である。

リスク管理は「入口→取引時→事後→改善」というフェーズの連鎖として捉えるべきであり、AI 導入もこの連鎖の中で設計される必要がある。入口で見落としたリスクは取引時のアラート増に跳ね返り、事後の調査ログが不十分であれば改善が進まず、同じ失敗が繰り返される。フェーズごとの AI 導入方針と、フェーズ間をつなぐ情報設計こそが要となる。

（ア）入口フェーズ：見落としを減らし、アラートを減らす

入口フェーズは、入会申込・KYC 新規登録・法人初期審査・端末や口座の新規紐付けなど、後工程の負荷を決めるゲートである。AI の活用領域は、属性情報の整合性チェック、提出書類の不備検知・真贋判定、反社会的勢力・制裁リストとの突合、過去の不正パターンに照らしたリスク要因の抽出などが中心となる。

ただし、反社会的勢力や制裁対象への該当性など、法令・規程上の慎重な判断を要する領域では、AI による判定だけで結論を確定することは適切ではない場合がある。AI は、候補抽出・優先順位付け・情報整理を担い、最終判断に必要な根拠と確認結果を証跡として残す設計が求められる。

審査を強化するほど取引時の不正は抑えられるが、過剰に厳格化すると正当な顧客・法人の獲得を阻害し、機会損失につながる。AI 導入の観点では、入口の判断基準をブラックボックス化させず、なぜその判断に至ったのかを説明できる形で根拠を残すことが重要である。入口の精度が上がると、取引時のアラートは単に減るだけでなく、優先的に確認すべき対象に絞り込むことができるようになる。

（イ）取引時フェーズ：検知ではなく「制御」へ

取引時フェーズは、即時性と顧客影響のバランスが特に厳しく問われる領域である。検知と審査、制限と解除、顧客への通知と追加確認という一連の行為が、秒から分の単位で求められる場面もある。

AI が担うべきは、「怪しい取引の列挙」ではなく、対応まで含めた一連の制御を支援することである。AI が取引リスクを評価し、一定条件で自動的に追加確認を要求し、確認が取れれば解除し、取れなければ制限を維持するという流れを設計できれば、損失抑制と顧客体験の双方が改善する。

ただし、制限・解除・追加確認といった顧客に直接影響する措置を自動化する場合には、適用条件、例外

条件、承認要否、誤判定時の救済手段をあらかじめ定めておく必要がある。

また、この実現にはケース管理が不可欠である。誰が・いつ・どの情報をもとに・何を判断し・どんな通知を出したかが追えなければ、事後の説明も改善もできない。AI を取引時に組み込むほど、ケース管理は単なる記録ではなく統制の中核になる。

（ウ）事後フェーズ：証跡とテンプレートが AI 効果を増幅する

事後フェーズには、調査・補償・チャージバック対応・再発防止・当局および社内への報告が含まれる。多くの組織で属人化が起きやすい領域であり、ケースごとに判断が揺れ、対応リードタイムが延び、監査対応が重くなる傾向がある。

AI 活用の焦点は、調査ログの整理・要約、必要書類の作成支援、過去の類似事例の検索、判断理由のテンプレート化、提出物の整合性チェックなど、判断品質を揃える方向に置くべきである。補償判断そのものを AI に委ねるかどうかは、法令や規程、顧客影響、判断の難易度を踏まえた慎重な検討を要する。一方で、少なくとも「判断に必要な材料を揃え、根拠を残す」部分は AI を活用しやすい領域である。事後フェーズで得られた不正手口や判断結果、対応履歴を整理・蓄積することで、改善フェーズにおけるルール・モデル・業務手順の見直しが可能となり、次の攻撃への備えを強化できる。

（エ）改善フェーズ：定義統一が改善速度を決める

改善フェーズは、ルール・モデルの改定、現場教育、ナレッジ化を含む。AI 導入を単発のプロジェクトに終わらせないためには、この改善ループを業務として定常化することが必要である。攻撃手口の変化に対して検知ロジックを見直し、運用手順を変え、現場の判断をデータとして蓄積し、必要に応じてモデルやルールに反映させる。

ここで鍵となるのが定義の共通化である。同じ事象に対して部署ごとに呼び名や分類が異なると、学習データは統合できず、改善は遅れる。定義を統一し、判断理由をコード化して残すことで、現場の知見が再利用可能な資産となる。

（オ）各フェーズをつなぐ設計の四つの要点：フェーズ間連携を「仕様」として固定する

フェーズ間の接続を機能させるためには、以下の四点を設計の前提として組み込む必要がある。

第一に、データと実行環境の連携である。顧客・法人・取引・端末・ネットワークといった各種情報が、同一の前提（粒度・期間・キー）で突合できなければ、AI が導出する判断は局所的・断片的なものにとどまり、検知精度や説明可能性が劣化する。特に不正対策では、単一チャネル内のシグナルだけでなく、複数チャネルを横断した行動系列や関係性（同一人物・同一端末・同一法人グループなど）を捉える必要があり、データ統合の成否が実務効果を大きく左右する。

第二に、ケース管理の整備である。AI がフラグを立てた後の調査・判断・措置（ブロック・追加認証・返金・届出など）を、調査ログ・証跡・判断理由とともに一元管理し、監査・当局対応に耐える形で保存することが不可欠である。誰が・いつ・何を根拠に判断したかというプロセスの再現性に加え、モデルが提示したスコアや寄与要因、適用したルール、参照したデータのバージョンまで追跡できることが望ましい。これらが欠けると、AI 導入後に「判断は速くなったが説明できない」「担当者ごとに運用がばらつく」という事態が生じやすい。

第三に、共通 ID と共通辞書の整備である。顧客 ID・法人 ID・端末 ID・アカウント ID・ネットワーク識別子などが体系的に整理されていない場合、同一主体の名寄せが不完全となり、なりすましや法人不正の検知精度が損なわれる。また「疑わしい取引」「チャージバック起因」「アカウント乗っ取り」などの用語や分類の意味が部門・システム間で異なると、分析結果と現場運用が整合せず、モデルの学習データ自体が歪む。共通辞書は単なる用語集ではなく、業務上

の定義・判定基準・コード体系・例外規程までを含む「共通仕様」として設計する必要がある。

第四に、KPI の一貫性である。 不正損失の低減のみを追えば過剰なブロックに傾き、検知率のみを追えば偽陽性が増え、TAT（処理時間）の短縮だけを追えば調査品質が低下する。不正損失・検知率・偽陽性率・TAT・顧客影響（離脱・苦情・利用阻害）を同時にモニタリングし、相互のトレードオフを前提に意思決定できる枠組みが必要である。フェーズをまたいで同じ指標定義・集計単位・評価期間を用いることで、上流の検知最適化が下流（調査・報告・再発防止）に与える影響を定量的に把握でき、局所最適を避けた全体最適の運用が可能になる。

AIありきの業務フローへ――既存フロー置換の限界

現行のリスク運用は、多くの場合「人が見る」ことを前提に設計されている。網羅性を優先してアラートを多めに発報し、担当者が目視で一次判定を行い、疑義が残れば追加調査へ進むという構造である。この枠組みに AI を「上乗せ」するだけでは、AI は判断主体にならず、既存アラートとは別系統のシグナルやスコアを追加するにとどまる。その結果、確認対象が増え、判断材料も増え、現場の認知負荷と処理負荷はむしろ上がりやすい。

AI 導入の本質は、現行フローの単純な置換ではなく、「AI が仕分け、人は高難度・例外・最終責任に集中する」という前提に立った、業務と統制の再設計にある。

再設計の第一の要点は、「止めどころ」の配分を見直すことである。 入口→取引時→事後→改善の四領域における介入の重心をどこに置くかは、事業戦略と直結したトレードオフである。入口の審査を強化すれば取引時のアラートや事後対応は減るが、承認率や顧客獲得に影響が出る。取引時に強く止めれば損失は抑えられるが、誤検知による顧客体験の毀損・売上機会の逸失・問い合わせ増といった副作用が増える。事後に比重を寄せれば顧客体験は守りやすい反面、損失の顕在化を許容する設計となり、回収や届出を含む後工程の負荷が増える。AI の価値は、このトレードオフを固定的なルールで一律に裁くのではなく、顧客属性・行動文脈・環境変化（キャンペーン・攻撃の流行・季節性など）を踏まえて、閾値や介入手段（止める・追加認証・保留・通す）の設計・調整を高度化できる点にある。AI 導入箇所を「検知モデルの追加」という単点で捉えるのではなく、全体配分の設計思想として決めるべきであり、KPI も四領域を通じた複合的な指標体系として管理する必要がある。

第二の要点は、「調査のための調査」を削ることである。リスク管理業務では、判断に必要な材料が複数システムに散在し、照会・転記・整形・突合といった情報収集だけで相当の時間が消費されることが多い。ここに高度な AI 判断を導入しても、入力が揃わなければ結論は出ず、かえって「AI はこう言っているが根拠を示せない」「追加確認が増えた」という停滞が生じる。証跡（参照元と取得時点）・根拠（判断に使った要素）・結論（措置内容）、およびその記録様式（テンプレート・必須項目・レビュー手順）を標準化し、ケースとして再利用可能な形に整えることが前提となる。AI が効果を発揮しやすいのはまさにこの領域、すなわち情報収集の自動化・関連情報の要約・差分の抽出・記録の整形であり、現場の可処分時間を「判断」へ戻す効果が大きい。逆に、この基盤整備を欠いたまま自動承認・自動却下といった強い自動判断に踏み込むと、説明責任や監査耐性の不足が顕在化し、結局は人手の「後追い検証」が増えて期待効果を損なう。

第三の要点は、人の役割を「処理」から「監視と例外対応」へ移すことである。 AI 導入後に真の価値を持つのは、全件を確認する人数の多さではなく、以下を担える体制である。①閾値・ルール・介入手段の設計とチューニング、②モデル性能の監視（ドリフト検知・偽陽性と偽陰性の構造分析）、③顧客影響の定量評価（離脱率・問い合わせ・収益への影響）、④事故・障害時の判断（手動運用への切り替え・暫定ルール・

当局および社内への報告）。言い換えれば、業務フローの再設計は、権限・責任・レビューラインを含む役割設計の再設計と同義である。これを先に定義するほど、AI 導入は「現場が回る」形で安定しやすい。加えて、モデル変更の統制（変更管理・検証・リリース判定）や例外処理の設計（誰が・どの条件で・どこまで逸脱を許容するか）まで含めて設計しておくことが、実装後の混乱を最小化する。

HITL(Human in the Loop)の罠――人の介在は意図して設計しないと負債になる

HITL は、AI の判断プロセスに人が介在することで、結果の品質や妥当性を担保する考え方である。金融のように、誤判定が損失や顧客影響に直結し、かつ説明責任・監査耐性が強く求められる領域では、一定の業務において HITL は重要な選択肢となる。ただし、すべての処理に一律に人を介在させることが適切とは限らない。HITL は意図して設計しなければ単なる「人手による確認工程の追加」になり、期待した省力化が実現できない。確認範囲が全件に近づくほど工数は減らず、「AI の結論をレビューする」という新たな責任だけが現場に上乗せされる。

HITL の原則は、介入ポイントを限定し、介入の目的を明確化することにある。介入の目的は少なくとも四つに整理できる。第一に品質担保としての介入で、誤検知や見逃しを抑えるために人が最終判断を行う。第二に例外処理としての介入で、境界事例や情報不足など AI が扱いにくいケースを人が判断する。第三に学習データ付与としての介入で、人の判断をラベルとして残し、モデル改善に還流させる。第四に顧客配慮としての介入で、形式的には制限できても、顧客関係や事情を踏まえ、対応を調整する。これらを区別しないまま「必要に応じて人が確認する」運用設計とすると、担当者は安全側に倒れて確認範囲を広げ、結果として運用負荷が膨張しやすい。どの目的で、どの条件で、誰が、どの権限で介入するのかをルールとして明文化し、KPI（偽陽性、TAT、顧客影響など）と結びつけることが不可欠である。

さらに責任分界を整理しなければ、HITL は組織リスクになる。モデル性能に責任を持つ者、業務判断に責任を持つ者、例外承認をする者、監査・牽制をする者が曖昧だと、事故時に原因究明が進まず、再発防止も止まる。加えて、人の判断がモデル改善に還流しない場合、HITL は恒久的な「人件費の上乗せ」という負債になる。判断理由を理由コードとして残し、証跡とともにケース管理へ紐づけ、改善の会議体で定期的に分析し、ルール・UI・データ整備・モデル更新へ反映する。この循環を回して初めて、HITL はコストではなく投資となり、介在範囲を段階的に縮小しながらも説明責任を維持する運用へ移行できる。

導入に向けたロードマップ

実装の難易度と期待効果のバランスを踏まえると、当面の到達目標は、対象業務を限定したレベル 3、すなわち「検知・審査・対応」にまたがる業務プロセスの自動化に設定するのが一案である。（図 5）

- すべての業務を一律にレベル 3 へ移行するのではなく、顧客影響、判断の複雑性、データの整備状況、統制要件を踏まえて、適用範囲を段階的に定めることが重要である。

以下では、決済領域における不正検知・モニタリング・取引制限を例として、三段階のロードマップを示す。（図 6）

（ア）第 1 段階：現行フローの分析と業務再設計（3 ～ 6 か月間）

第 1 段階では、現行フローを分解・分析したうえで、AI 活用を前提とした業務フローへ再設計する。具体的には、既存の検知ルール・モニタリング時の確認観点・取引制限の分岐条件や対応パターンを棚卸しし、どの判断を AI に委ね、どの判断を人が担うべきかを定める。

その際、品質担保および例外処理として人が介在すべきポイントを明確化し、HITL の介入条件・権限・エスカレーション経路を定義する。また、導入後に追跡

図5　レベル別 期待効果と実装難易度

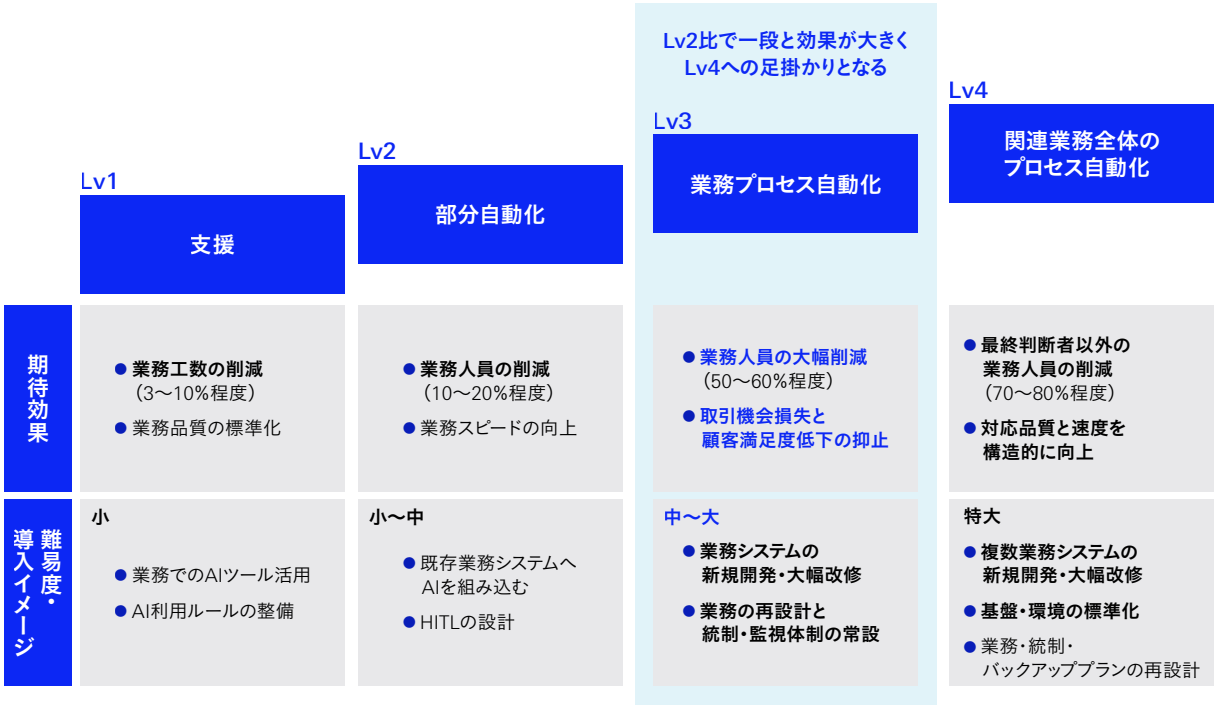
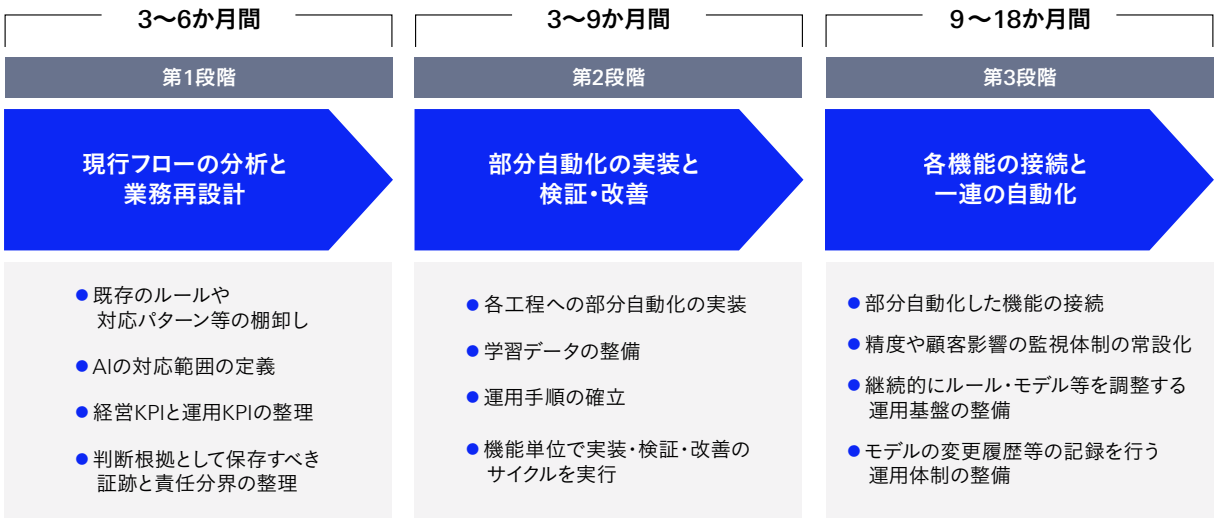


図6　導入に向けたロードマップ



すべきKPIとして、経営判断に資する経営KPI（不正損失・売上機会損失・補償コスト・顧客影響・重大インシデント件数など）と、運用のPDCAサイクルを回す運用KPI（検知率・誤検知率・アラート量・TAT・問い合わせ量など）の両軸を整理する。さらに、判断根拠として保存すべき証跡の範囲、ならびに責任分界（モデル・運用判断・例外承認・監査）の整理を行う。

なお、KPIは単独で評価するのではなく、相互のトレードオフを確認できるよう設計する必要がある。例えば、検知率の向上だけを追求すると誤検知率や顧客

影響が悪化する可能性があるため、複数の KPI を組み合わせて導入効果を評価することが重要である。

この初期段階に十分な検討工数を投じて設計を固めることが、以降の実装効率と効果の最大化を左右する。後工程での手戻りを最小化するためにも、「設計に時間をかけすぎている」という現場の焦りに流されず、この段階の完成度を優先すべきである。

（イ）第 2 段階：部分自動化の実装と検証・改善（3 ～ 9 か月間）

第 2 段階では、再設計したフローの中で部分自動化を順次実装し、各処理における学習データの整備・精度検証・運用手順の確立を進める。フロー全体へ一度に適用すると、問題の切り分けや影響範囲の特定が困難になるため、機能単位で「実装→検証→改善」のサイクルを短く回す進め方が望ましい。

この段階では、精度の高さだけでなく「運用として回せるか」を含めて評価することが重要である。モデルの性能指標、誤検知・見逃しの発生状況、顧客影響に加え、現場担当者が実際に HITL の設計どおりに動けているか、ケース管理に必要な証跡が適切に残されているか、KPI が想定した形で計測できているかを並行して確認する。これらを踏まえ、第 3 段階への移行可否を判断する材料を積み上げる。

（ウ）第 3 段階：各機能の接続と一連の処理の自動化（9 ～ 18 か月間）

第 3 段階では、部分自動化した各機能を接続し、定められた条件下での自動制限・自動解除に加え、通知・記録・証跡保存までを含む一連の処理を自動化する。顧客影響の大きい措置については、承認を要する条件や例外条件を明確にしたうえで、段階的に適用範囲を拡大する。

同時に、検知精度の経時劣化（性能ドリフト）や顧客影響を継続監視する体制を常設化し、閾値・ルール・モデルの調整が継続的に行える運用基盤を整える。

この段階で特に留意すべきは、自動化の範囲を広げるほど、個々の判断の根拠が見えにくくなるリスクが高まる点である。モデルのバージョン管理・変更履歴・リリース判定の記録を整備し、「いつ・何を・なぜ変えたか」を事後に再現できる体制を維持することが、監査対応と継続的改善の双方において不可欠となる。

おわりに

AI 導入はツールの追加ではなく、意思決定と統制の再設計である。目指す自動化レベルの選択と HITL の設計がその成否を左右し、フェーズをまたいだ情報設計と KPI の一貫性がその効果を持続させる。

まずは対象業務を限定したレベル 3 の実装可能な範囲で、データ・ケース管理・KPI・責任分界を揃え、「運用として回る形」を確立することが先決である。華やかな自動化の絵を描く前に、現場が判断の根拠を残せているか、人の介在が設計どおりに機能しているか、得られた知見が次のモデルやルールの改善に還流しているかを問い続けることが、AI 導入の恩恵を最大化する。

その土台の上で初めて、関連業務を横断するレベル 4 の自動化が現実的な選択肢となる。不正と戦う力は、AI を導入したかではなく、「AI が判断し、人が統制し、組織が学習し続ける仕組みを築けたか」で決まる。高度化する不正に対抗できる組織への変革は、技術導入の瞬間ではなく、この地道な設計と運用の積み重ねによる進化から始まる。

マネージャー　田中 遼一
シニアコンサルタント　丸山 茉希
パートナー　岡 遼平