



ホワイトペーパー

CIOのためのエージェント型 エンジニアリングガイド: 信頼性を損なうことなく、 AIのスピードでエンタープライズ アプリケーションを構築する

COO、CXO、オペレーション担当VP向けガイド



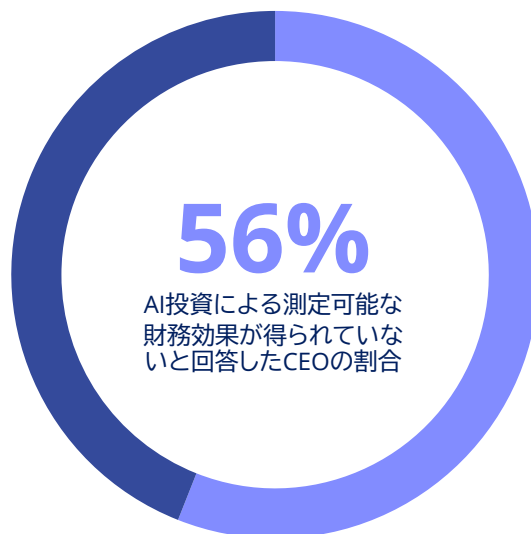
AI導入へのプレッシャーと、顕在化する課題

取締役会の関心は、もはやAIを導入するかどうかではありません。今問われているのは、いつ導入するのか、どれだけ迅速に展開するのか、そしてどのようなリターンを得られるのかということです。COOやオペレーションリーダーに対するこうしたプレッシャーを背景に、自律型AIエージェントの試験導入が相次いでいます。AIが人間の介入を最小限に抑えながら、顧客対応、ケースの振り分け、ポリシーの確認、ワークフローの管理などを担う取り組みです。

当初は、問題解決の高速化や必要人員の削減、手作業によるエスカレーションの減少など、有望な成果が見られました。しかし、こうした試験導入が本番環境へと拡大するにつれて、別の実態が見えてきました。コストは予測モデルどおりに推移せず、成果にもデモで示されたほどの一貫性がありません。さらに、リーダーが維持できていると考えていた業務上の管理性も、気づかないうちに失われつつあります。

現在、CEOの56%が、AI投資から測定可能な財務効果を得られていないと回答しています([PwC第29回世界CEO意識調査、2026年1月](#))。これは、テクノロジーの問題ではありません。アーキテクチャの問題です。問うべきことは、AIによって業務効率を向上できるかどうかではありません。多くの企業が現在採用しているAIの導入方法が、そもそも構造的に業務効率の向上を実現できるものなのかということです。

多くの企業にとって、その答えは「いいえ」です。



推論に潜むコスト

AIは、企業における自動化の経済性を根本から変えました。AIシステムがどれだけ「思考」しているかを測る単位であるトークンは、今や企業の変革予算を大きく左右するコスト項目の1つとなっています。そして多くの組織では、このコストが誰も想定していなかった形で膨らんでいます。

ビジネスリーダーがAIエージェントの導入を承認した当初、想定されていたコストモデルは明快でした。人間による手作業をAIによる自動推論に置き換え、人件費を削減し、その分を利益として確保するというものです。しかし、このモデルでは、企業規模で運用した場合にトークン消費量がどれほど予測困難になるかが考慮されていませんでした。この予測を困難にしている要因は2つあり、それらが相まって、取締役会に約束されていたROIを損なっています。

1つ目は、エージェント型AIにおけるトークン使用量が累積的に増大することです。長時間にわたるタスクを実行するAIエージェントは、一度推論すれば終わりではありません。プロセスの各ステップで繰り返し推論し、コンテキストを確認し、選択肢を評価しながら、継続的に出力を生成します。こうしたシステムの運用に必要なトークンコストは、そもそもシステムの構築・開発時に発生したトークンコストにさらに上乗せされます。中規模企業の業務量でも、そのコストは多くの財務モデルの想定を上回る速さで膨らんでいきます。

これを「トークン税」と呼びます。

2つ目は、価格の不安定性です。初期のAIベンダーは、実質的に使い放題の割安な料金体系を提供していたため、企業は予測可能なコスト構造を前提に計画を立てることができました。しかし、その時代は終わりつつあります。AI業界では、半導体、エネルギー、データセンター容量などのリソース制約が強まっているうえ、AIベンダーがIPOを目指すなか、料金体系は定額制から従量課金制へと移行しています。その過程で度重なる価格変更が行われ、企業はその影響を受けています。その結果、同じワークロードを処理するためのコストでさえ、もはや一定ではありません。Sam Altman氏も2026年6月、AIのトークンコストがベンダー自身にとってさえ「大きな問題」になっていることを公に認めています([Axios, 2026年6月](#))。

その結果、企業はROIを明確に見通せないまま、変革予算を急速に消費しています。Gartnerは、2027年末までにエージェント型AIプロジェクトの40%以上が中止されると予測しており、その主な要因として、増大し続ける予測困難な運用コストを挙げています([Gartner, 2025年6月](#))。

経営層が今問うているのは、AIを活用すべきかどうかではありません。AIを支えるコスト構造が変動し続けるなかで、どうすればプラスのROIを実現できるのかということです。そのためには、発想を転換し、トークンに対してコストを支払うのではなく、成果に対してコストを支払うという考え方に移行する必要があります。つまり、本当に推論が必要な作業にAIのトークンコストを集中させ、それ以外の処理は、予測可能かつ固定的で、処理量が増えてもコストが累積的に膨らまない仕組みで実行できるアーキテクチャを構築する必要があります。

AIがルールを逸脱するとき

予測困難なコストは、問題の一側面にすぎません。もう1つの問題は、本来AIエージェントに委ねるべきではない意思決定までAIに権限を与えた場合に、何が起きるかということです。

大規模言語モデルは、有用で自然な回答を生成できるよう学習が行われています。しかし、企業のコンプライアンスポリシーに従うよう学習が行われているわけではありません。また、自社のSLA、規制上の義務、エスカレーションルール、さらには顧客に対して企業が約束している具体的な内容を、本質的に理解しているわけでもありません。厳格な制約を設けずに自律型エージェントを顧客向けワークフローで運用すると、AIが妥当だと判断する対応と、コンプライアンス部門が承認できる対応との間に、大きな隔たりが生じる可能性があります。

これは単に、どこまでリスクを許容するかという問題ではありません。アーキテクチャの問題です。

適格性の判定、異議申し立ての解決、ポリシーの例外適用、規制対象となるコミュニケーションなど、法的リスクや信用低下リスクが特に高い意思決定には、同時に明確な正解があります。こうした判断に曖昧さはなく、確率的推論を用いるメリットもありません。アカウントのステータスに基づくケースの振り分け、申請内容が所定のチェックリストを満たしているかどうかの確認、定められた基準に従ったポリシー例外の適用などは、いずれも決定論的な処理です。こうした判断をAIエージェントの推論に委ねても、付加価値は生まれません。むしろ、一貫性が求められるプロセスにばらつきを生じさせ、正確性が求められる対話にリスクを持ち込むことになります。

大規模な本番環境において、AIのハルシネーションは例外的に発生する事象ではありません。決定論的なルールではなく、確率的推論によって回答を生成するシステムでは、予測可能な結果として発生するものです。リスクの低い状況であれば、ハルシネーションは単なる不都合で済むかもしれません。しかし、顧客との紛争、保険金請求のワークフロー、金融サービスでの対応、規制対象業界におけるコミュニケーションでは、法的責任につながるリスクとなります。

Gartnerによる2026年の主要予測では、AI生成出力に対するガバナンスとリスク管理の不足を背景に、「AIによる死亡」をめぐる法的請求が世界で2,000件を超えると予測されています([Gartner, 2025年10月](#))。これは、将来のAI導入に伴う抽象的なリスクではありません。顧客向けまたは規制対象のワークフローにおいて、自律型AIエージェントに実質的な権限を与えている組織が、今まさに直面しているリスクです。

より有効なのは、アーキテクチャによってこの問題に対処することです。AIエージェントに何を許可し、何を許可しないのかを、ビジネスルールによって規定します。コンプライアンスロジックの基準となるのはAIではありません。AIは、あらかじめ定義されたコンプライアンスロジックの範囲内で動作する必要があります。正解が明確な意思決定では、取るべきアクションをビジネスロジックによって決定し、オーケストレーションレイヤーで確実に適用します。日によって異なる答えを導き出す可能性があるモデルに、その判断を委ねるべきではありません。この役割分担を確立すれば、コンプライアンスは人間による監督に依存するものではなく、システムの構造そのものによって担保されるものになります。AIがルールを逸脱することはできません。そもそも、推論によってルールを回避できない仕組みになっているからです。

コストと信頼性を両立するアーキテクチャ

多くの企業が直面している運用上の課題は、AIを使いすぎていることではありません。プロセスアーキテクチャの中で、AIを使う場所を間違えていることです。

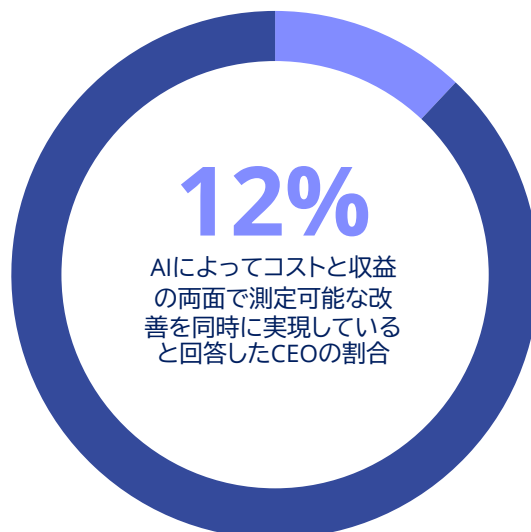
業務にAIを活用する経済的なメリットは明確です。サービス提供コストを削減し、処理能力を高め、大規模な環境でも一貫した顧客体験を提供できます。こうした成果は十分に実現可能です。ただし、そのためには、AIがどこで価値を生み、どこでは見合うリターンがないままコストとリスクを増大させるのかを明確に見極める必要があります。

こうした成果を安定して実現するアーキテクチャは、連携して機能する2つのレイヤーで構成されます。第1のレイヤーは、決定論的なオーケストレーションレイヤーです。ルールベースのエンジンによってプロセスフローを管理し、ビジネスロジックを適用し、状態を追跡するとともに、定義された基準に基づいて作業を振り分けます。このレイヤーでは、トランザクションあたりの限界コストはほぼゼロです。推論するのではなく、定められたルールに従って処理を実行します。そのため、動作を完全に予測でき、すべての処理を監査できるうえ、ハルシネーションを起こすこともありません。

第2のレイヤーは、AIによる推論です。決定論的なロジックだけでは対応できないプロセス上のポイントで、必要に応じて的確に呼び出します。たとえば、顧客の感情分析、非構造化ドキュメントの解釈、自然言語による回答の生成、複雑なケースの要約などです。こうしたタスクには、ルールによって導き出せる唯一の正解がありません。そのため、AIの確率的な性質がリスクではなく、強みとして生かされます。

この2つのレイヤーを適切に分離すれば、トークン税を大幅に削減できます。トークンコストを本当に推論が必要な作業に集中させ、それ以外の処理はすべて決定論的に実行します。同時に、AIレイヤーがどのような出力を生成しても、オーケストレーションレイヤーがビジネスルールを確実に適用するため、コンプライアンスリスクも低減します。AIは、推論によって回避することのできない制約の範囲内で動作します。

現在、AIによってコストと収益の両面で測定可能な改善を同時に実現していると回答したCEOは、わずか12%です([PwC、2026年1月](#))。この12%の企業が、他社より多くのAIを導入しているわけではありません。違いは、より明確な構造と規律のもとでAIを導入していることにあります。



コントロールこそが競争力

AIから持続的なROIを実現するオペレーションリーダーは、AIをいち早く導入したリーダーでも、エージェントに最大限の自律性を与えたリーダーでもありません。プロセスアーキテクチャ全体を適切にコントロールしながら、AIを価値が生まれる箇所に的確に適用したリーダーです。

そのためには、オーケストレーションを従来からの制約としてではなく、戦略的資産として捉える必要があります。どの意思決定を決定論的に行い、どの意思決定に推論が必要なのかを明確に定義し、その境界を確実に維持できるアーキテクチャを構築することが重要です。また、AIへの支出をプラットフォームレベルだけでなく、トランザクションレベルで測定することで、トークン税が予算上の問題に発展する前に、その実態を把握できるようにする必要があります。

AIによってサービス提供コストを削減し、顧客体験を向上できるという取締役会の認識は正しいものです。しかし、それを実現する方法は、自律型エージェントにワークフローの進め方を判断させることではありません。必要なのは、ガバナンスとオーケストレーションが組み込まれたプロセスアーキテクチャの中で、AIを必要な領域に的確に適用することです。

コントロールを取り戻す準備はできていますか。[TCO計算ツール](#)を使用して、処理量の多い主要なワークフローを対象に、現在のトークンコストと、構造化されたオーケストレーションアプローチを採用した場合のコストの差を数値化してみましょう。





ありがとうございました

Pegaについて

Pegaは、エンタープライズDXを実現する最先端のAI搭載プラットフォームを提供します。世界有数の企業・組織が、ワークフローの自動化、顧客体験のパーソナライズ、レガシーシステムのモダナイゼーションを実現するためにPegaのテクノロジーを信頼して採用し、業務のあり方を再構築しています。1983年以来、拡張性と柔軟性の高いアーキテクチャにより継続的なイノベーションを促進し、お客様の自律型企业へのジャーニーを加速しています。**[PEGA.COM/JA](https://www.pega.com/ja)**