



AIの隠れたコスト問題は モデルではなく、 使い方にある

予測可能なAIを実現するために

2026年6月 | Don Schuerman、CTO兼マーケティング責任者



AIの次なるフェーズで問われるのは、可能性ではなく説明責任

この2年間で、企業におけるAIは「試してみるもの」から「実際に活用できるもの」へと進化しました。かつては専門チームと長期にわたる開発サイクルを必要としていたことが、今では数日で実現できます。コードを生成し、エージェントを立ち上げ、ワークフローを自動化する。開発プロセスは、これまでにないほど高速化しています。

こうした変化に伴い、経営層がAIに向ける視線も変わりました。

1年前に問われていたのは、「AIで何ができるのか？」でした。
今、問われているのは、「AIにどれだけのコストがかかり、実際にどれだけの価値を生み出しているのか？」というより現実的な問題です。

企業が試験導入の段階を越え、AIを実務へと本格的に展開するにつれて、期待も高まっています。AIはもはやイノベーション予算だけで賄うものではなく、ビジネスを動かす仕組みの一部になりつつあります。だからこそ、信頼性、ガバナンス、コストのすべてについて、厳しい検証に耐えられるものでなければなりません。

業界を問わず、AIの可能性は明らかである一方、その経済性についてはまだ答えが出ていないという認識が広がっています。Gartnerの最近の調査¹によると、AIへの取り組みの多くはいまだ十分なリターンを生み出せておらず、初期投資から測定可能なROIをほとんど、あるいはまったく得られていない組織も少なくありません。

これはAIの失敗を意味するものではなく、AI活用が新たなフェーズに入ったことを示しています。これから成果を左右するのは、何を構築できるかではありません。大規模に展開したAIを、どれだけ効果的に運用できるかが問われるのです。

¹ AIの価値に関するGartnerの調査(gartner.com/ja/articles/ai-value)

AIの構築は容易に。 しかし、効率的な運用はより困難に

AI導入の第一波を後押ししたのは、AIへのアクセスが容易になったことでした。高性能なモデルが広く利用できるようになると、企業はこぞって実証実験に乗り出し、イノベーションが加速しました。しかし、こうした取り組みが本番環境へと拡大するにつれて、新たな現実が見えてきています。

試験導入の段階では管理可能に見えたコストも、規模を拡大すると増加し始めます。出力品質にはばらつきが生じ、AIへの投資とビジネス成果との明確な関連性を測ることも難しくなります。その根底には、構造的な問題があります。現在のAIシステムの多くは、実験を目的として設計されており、継続的な運用を前提として設計されていないのです。



多くのAIシステムは、モデルが対話のたびにゼロから評価を行う *実行時の推論* に大きく依存しています。こうした柔軟性が有効な場合もありますが、その代償としてコストが発生します。対話するたびにコンピューティングリソースが消費され、処理が複雑になるほど消費量も増加します。さらに、AIの利用が拡大するにつれて、こうしたコストは累積的に膨らんでいきます。

モデルがタスクを再評価したり、コンテキストを解釈したり、次を取るべきアクションを判断したりするたびに、追加のコストが発生します。小規模な利用であれば管理可能ですが、企業規模になると、システムが単に処理を実行するのではなく、同じことを何度も繰り返し推論するため、コストが累積的に膨らんでいきます。その結果、多くの企業が実質的な「トークン税」とも言える負担に直面しています。推論のステップが1つ増えるたびに限界コストが発生し、それが大規模な運用環境では積み重なっていくのです。

規模が拡大すると、コスト構造そのものが大きく変わり始めます。多くの企業では、本番環境でモデルを継続的に稼働させる推論コストが、モデル構築時の初期コストを急速に上回ります。その結果、AIは一度限りの投資ではなく、継続的に発生する運用コストへと変わっていきます。

AIの本当のコストを左右するのは、何を構築するかではありません。構築したものを継続的に稼働させるために、どれだけのコストがかかるかです。企業は、あらかじめ構造化できる処理まで、その都度AIに判断させ、そのたびにコストを負担しています。だからこそ、管理された環境での試験導入では問題なく機能していたソリューションも、実際の業務環境では効果的に規模を拡大できないことが少なくありません。問題はAIの可能性ではなく、その可能性をいかに効率よく実用化できるかにあります。

² MIT Technology Review, “AI’s energy usage and climate footprint” (AIのエネルギー使用量と気候への影響) (2025年)

AIの真のコストは、業務の進め方に潜んでいる

AIへの支出を管理しようとする、経営者はモデル、つまりモデルのプロバイダー、機能、料金体系に目を向けがちです。

しかし、それ以上に重要なのは、「業務のどこでAIが処理を行い、その業務はどのように構成されているのか」という問いです。

多くの組織では、AIが自然発生的に導入され、さまざまなチームが異なるモデル、エージェント、ワークフローをほとんど連携させることなく展開しています。その結果、複雑化するだけでなく、処理の重複も生じます。同じ意思決定が複数の場所で繰り返され、しかも、それぞれ異なるロジックに基づいて行われることも少なくありません。

現在の予測では、企業のAIワークロードの拡大に伴い、今後数年間でトークン使用量が20倍以上に増加すると見込まれています³。エージェントの利用がビジネス全体へと拡大するにつれて、増えるのは利用量だけではありません。コストも増加します。AIは、既存のプロセスそのものを根本的に変えるのではなく、その上に追加される形で導入されることが少なくありません。

その結果、一定の改善は得られても、抜本的な変革にはつながりません。そして、そこには2つの構造的な問題があります。

1. 非効率なプロセスがそのまま残ること。既存のプロセスが分断されていたり、過度に複雑化していたりする場合、AIを導入しても、その複雑化が解消されるわけではありません。単に、非効率なプロセスの処理速度が上がるだけです。
2. 実行時の判断に依存しすぎる。業務の進め方をあらかじめ定義するのではなく、AIが対話のたびに、次に何をすべきかをその場で判断することになります。

こうした柔軟性は一見すると大きなメリットに思えます。しかし、その一方で処理にばらつきが生じ、コストが増加し、結果をコントロールすることも難しくなります。突き詰めれば、これはアーキテクチャの問題です。

- ・ 適切に設計されたシステムでは、業務を開始する前の段階で必要になる作業を削減できます。
- ・ 一方、構造化が不十分なシステムでは、実行するたびにコストが発生します。

システムやチームをまたぐ業務の流れを見直さずにAIを導入すると、規模を拡大してもコストは安定せず、利用量が増えるほど膨らんでいきます。

このため、多くの組織では、成果の伸びを上回るペースでAIへの支出が増えているのです。

³ Goldman Sachs, “AI agents forecast to boost tech cash flow as usage soars” (AIエージェント、利用拡大でテクノロジー企業のキャッシュフロー押し上げへ)

問題のあるプロセスにAIを追加しても、問題は解決せず、むしろ問題を拡大する

AI導入でよく見られるのが、いわゆる「後付け型」のアプローチです。企業は、カスタマーサービス、オンボーディング、保険金請求などのワークフローを特定し、そこにAIレイヤーを追加して改善を図ります。その結果、利用体験はより対話的になり、システムの応答性も向上します。

しかし、基盤となるプロセス自体は変わりません。依然として複数のシステムが関与し、意思決定は分断されたままです。業務も、相互に連携していない一連の手順を経て進められます。それを補うために、AIがより多くの役割を担い、リアルタイムで情報を解釈し、処理を振り分け、状況に適応することで、プロセス全体をつなぎ合わせようとしています。

このアプローチによって、短期的には顧客体験を向上させることができます。しかし、コスト構造そのものが変わるわけではありません。むしろ、コストが増加することも少なくありません。システムが、従来の処理に加えて、その処理全体をつなぎ合わせるために必要な継続的な解釈も行うようになるからです。

システムが業務の流れをその都度見直したり、再構築したりするたびに、追加のリソースが消費されます。それが数千、数百万回もの対話で繰り返されれば、その影響はすぐに顕在化します。このモデルでは、AIは構造的な複雑性の上に、継続的な解釈を重ねるレイヤーとなります。

このような仕組みは、大規模な運用では持続可能ではありません。コスト構造を変えることが目的なら、まず見直すべきものはプロセス自体です。

コスト管理の鍵は、 実行時の最適化ではなく設計にある

AIの活用で着実な成果を上げている組織は、従来とは異なるアプローチでこの課題に取り組んでいます。AIを連携・構造化し、業務の設計と実行の仕組みに組み込む「オーケストレーション型モデル」へと移行しているのです。

AIにその場で意思決定を委ねるのではなく、業務の進め方を設計する段階からAIを活用しています。この段階で、AIは大きな力を発揮します。既存のプロセスを分析し、非効率な部分を特定し、システムやチームをまたいで業務をより適切に構成する方法を提案できます。

その構造が定義されれば、実行ははるかに効率的になります。実行時に、システムが各ステップで「何をすべきか」をその都度判断する必要はなくなります。あらかじめ定義されたプロセスに従って処理を進め、意図の解釈や非構造化データの処理など、AIが最も価値を発揮する場面に絞ってAIを活用します。

実行時の推論から設計時のインテリジェンスへと移行することで、コスト構造と成果の一貫性の両方が変わります。これにより、組織は次のことが可能になります。

- 不要なコンピューティング処理を削減する
- 同じ意思決定が繰り返されないようにする
- 定められた方法に従って、一貫性とガバナンスを確保しながら業務を実行する

このモデルでは、AIは業務を実行するシステムにとって代わるものではなく、その一部として機能します。

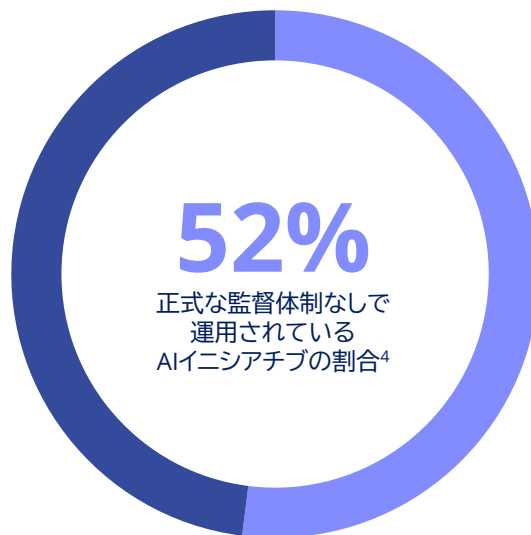
予測可能性が、企業規模でのAI活用を可能にする

AIが企業の基盤となるためには、他の基幹システムと同等の規律をもって運用される必要があります。

その第一歩となるのが、予測可能性です。経営者には、次の3つの点について確信を持てることが求められます。

1. どの程度のコストがかかるのか
 2. どのような結果が得られるのか
 3. 実際の運用環境でどのように動作するのか
- それができなければ、AIの利用を拡大することにはリスクが伴います。

EYの最近の調査によると、AIへの取り組みの半数以上が正式な監督体制のないまま運用されています。そのため、AIの導入が拡大するにつれて、組織はコスト超過やコンプライアンス上の不備、成果のばらつきといったリスクにさらされることとなります。



ここから、AIに対するより構造化されたアプローチが定着し始めます。AIがオーケストレーションとガバナンスに支えられ、明確に定義されたワークフローの中で動作するようになると、次のような重要な変化が生まれます。

- 成果の一貫性が高まる。
- コストを予測しやすくなり、意思決定やトランザクション、実際のビジネス成果との関連性もより明確になる。
- 業務を管理しやすくなる。

そして重要なのは、この予測可能性がトークン使用量だけに当てはまるものではないということです。この予測可能性は、運用やガバナンス、コンプライアンス、さらには長期にわたってシステムを進化させるための継続的な作業など、所有に伴うあらゆるコストに及びます。これにより、総所有コストもより明確になります。これにはトークンコストだけでなく、AIを長期にわたって運用、統制、保守し、規模を拡大していくためのコストも含まれます。

企業は、AIを際限なく活用できる機能として捉えるのではなく、業務プロセスの中で管理・測定できる要素として捉えるようになります。これにより、AIの利用量を追跡するだけでなく、AIによって得られた成果を追跡できるようになります。そして、この転換こそが、AIの経済性を確立するうえで重要な一歩となります。

⁴ EY, “Autonomous AI adoption surges as oversight falls behind” (自律型AIの導入が急増する一方、監督体制の整備には遅れ) (2026年)

AIへの発想を転換する企業が、成功をつかむ

多くの企業は今も、「AIを迅速に導入するか、コントロールを維持するか」という誤った二者択一にとらわれています。しかし実際には、制約となっているのはスピードではなく、アーキテクチャです。

AIの価値を最大限に引き出す企業は、単にAIをいち早く導入した企業ではありません。

AI活用を成功させる企業は、自社のビジネスのあり方そのものを見直します。具体的には、従来とは異なる次のような優先事項に重点を置くことになります。

ツールではなく、システムに目を向ける

重要なのは、AI単体で何ができるかではなく、企業全体の業務の流れにAIをどのように組み込むかです。

自動化する前に、プロセスを再設計する

問題のあるプロセスにAIを適用しても、問題が解決することはほとんどありません。まずプロセスそのものを見直すことが、意味のある成果を生み出すための基盤となります。

活動量ではなく、成果を測定する

成功の尺度とすべき要素は、単なるAIの利用量ではありません。業務の完了、顧客体験の向上、業務上の摩擦の軽減といったビジネス成果を基準に評価する必要があります。

AIを業務インフラとして捉える

AIが中核的なプロセスに組み込まれるにつれて、他のエンタープライズシステムと同等の信頼性とガバナンスの基準を満たすことが求められます。

これは、新しいテクノロジーを導入するよりも、そのテクノロジーをどのように活用するかについて規律を持つ方が重要だということです。

実証実験から本格運用へ

企業におけるAI活用は、今まさに転換点を迎えています。初期の段階では、AIによって何が可能になるのかが示されました。次の段階では、何が実用的で、何が大規模な展開に適しているのかが明らかになっていきます。

企業がAIの活用を拡大するなか、もはや課題は、インサイトを生み出すことや個々のタスクを自動化することではありません。重要なのは、そうしたAIの機能を、安定性と効率性を確保しながら、ビジネス成果につながる形で運用することです。

それが、個別の成功と、企業全体に及ぶ成果の違いです。

このギャップを埋めることができる企業は、より迅速に行動し、より効率的に業務を遂行し、変化にもより効果的に適応できるようになります。

しかし、それは単にAIの導入を増やすだけでは実現できません。

AIを自社のビジネスの中で適切に機能させるための基盤を構築することで、初めて実現できるのです。

まとめ

AIはすでに、企業に何ができるかを大きく変えています。

今、焦点は、AIを実際の環境で大規模に運用したときに、複雑なシステムの中でコストやコントロールの制約に対応しながら、どのようなパフォーマンスを発揮できるかへと移っています。

AIに潜む最大のコストは、モデルそのものに存在するものではありません。これは、業務がどのように構成され、実行され、長期にわたって繰り返されるかによって生じます。

この課題を解決できる企業は、単にコストを削減できるだけではありません。AIを、実際のビジネス運営を支える仕組みとして継続的に活用できるようになります。

そして、企業におけるAI活用の次の時代をリードするのは、こうした企業です。AIをコントロールし、測定し、ビジネスの一部として運用できる仕組みを構築した企業が、一歩先を行くことになるのです。



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、お客様の自律型企業へのジャーニーを加速しています。**[PEGA.COM/JA](https://www.pega.com/ja)**