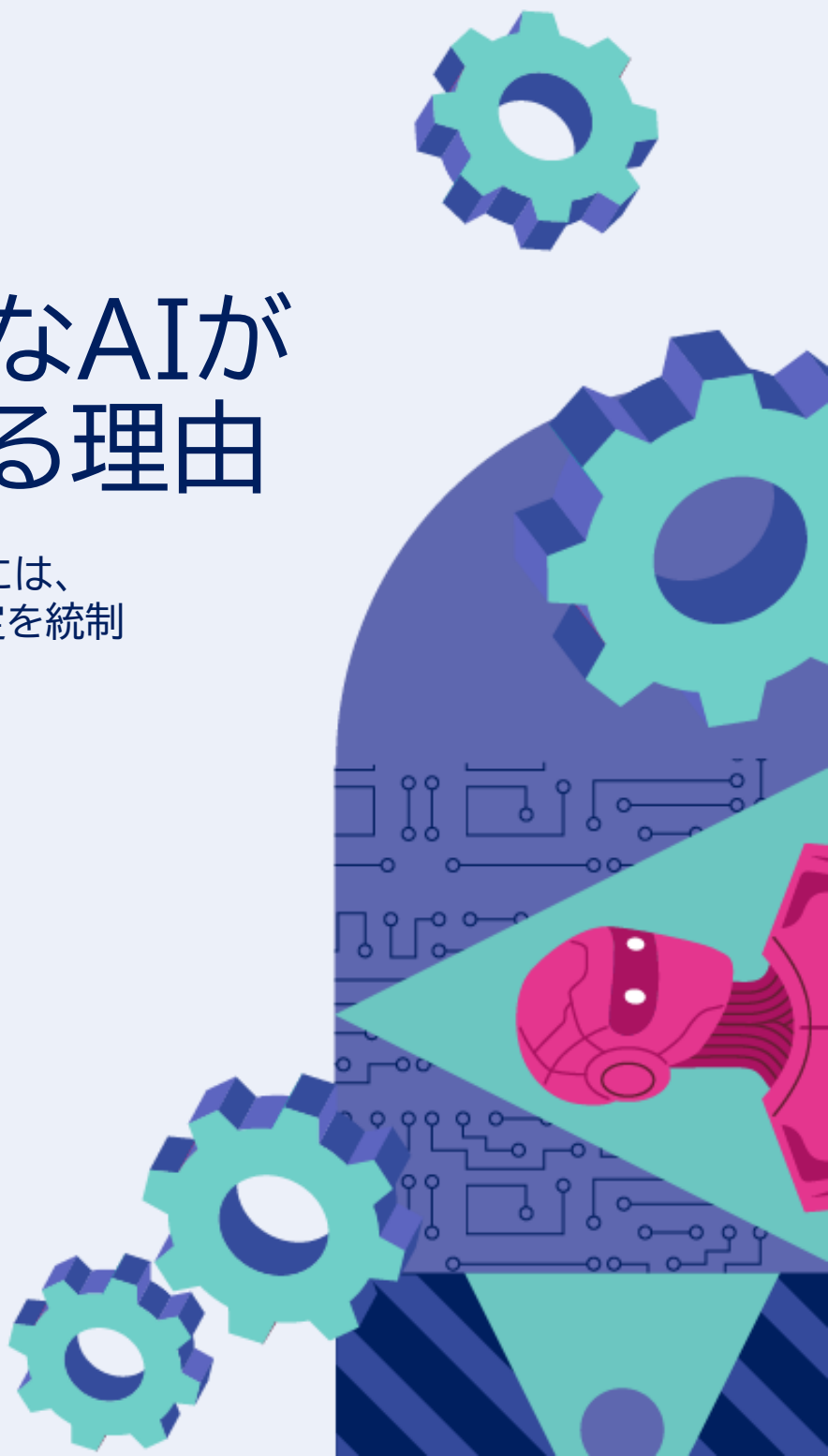


予測可能なAIが 重要である理由

トークンエコノミー時代には、
モデルではなく意思決定を統制



Pegaの構造的優位性

生成AIは、企業に新たな運用上の現実をもたらします。インテリジェンスに、測定可能で変動するコストが伴うようになったのです。プロンプト、検索、推論ステップ、エージェントループのたびにトークンが消費され、それに伴ってGPUの処理時間、エネルギー、そして設備投資も必要になります。従来のソフトウェアとは異なり、このコストは規模の拡大によって頭打ちにはなりません。利用量と複雑性に応じて増加し、場合によっては急激に膨らむこともあります。

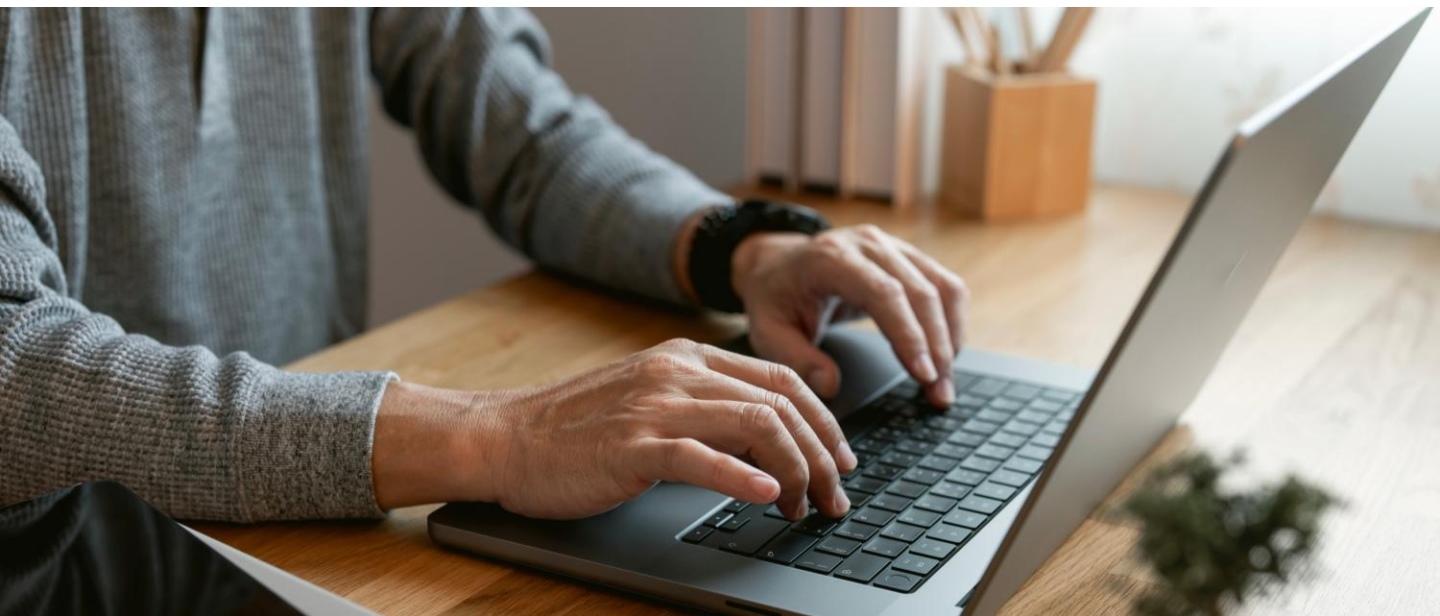
トークンを消費した時点で、コストと炭素排出はすでに発生しています。事後的なガバナンスでは効果がありません。

Pegaが構造的に異なるのは、この点です。Pegaはすでに、企業が次に何を行うかを決定する地点で機能しています。選択肢を評価し、ポリシーと制約を適用し、意思決定をリアルタイムで実行します。同じ意思決定レイヤーで、実行パス全体を確率的インテリジェンスに委ねるべきかどうか、委ねる場合はどのような制約を設けるかを判断します。委任後の処理については、その実行を許可するコスト、レイテンシー、リスクの範囲を、システムがあらかじめ把握していなければなりません。

したがって、予測可能なAI(Predictable AI)はモデルの機能でも、プロンプトのフレームワークでもありません。これはアーキテクチャ上の位置づけです。Pegaは、生成AIの経済性をリアルタイムでガバナンスできるコントロールプレーンに位置します。

この考え方では、予測可能なAIをトークンレベルの経済ガバナンスとして捉えます。個々のアプリケーションに委ねるのではなく、意思決定が行われる時点で、設計時に評価し、実行時に適用します。

設計時の評価で経済的な意図とポリシーを定義し、実行時の適用によってコストを制御します。



トークンエコノミーが デジタルコスト曲線を覆す

何十年もの間、エンタープライズソフトウェアには、規模の拡大に伴って限界コストが低下するという予測可能な経済パターンがありました。

生成AIは、この曲線を逆転させます。トークンは、コンピューティング、エネルギー、インフラストラクチャのコストに直接反映されます。組織がAIの利用を拡大すると、より豊富なコンテキスト、長い履歴、検索拡張生成、エージェント型ワークフローによって、トークン消費量も増加します。品質向上には通常、コンテキスト、検索、推論のステップ追加が必要であり、その一つひとつがインフラクションあたりのトークン数を増やすためです。

重要なのは、これらのコストが一時的ではなく、構造的なものであることです。生成AIを支えるインフラストラクチャ、つまり高密度GPUクラスター、電力消費量の多いデータセンター、アクセラレーテッドハードウェアの減価償却は、効率化だけでは相殺できない、

上昇基調のコスト構造を企業にもたらします。

AIインフラストラクチャのコスト構造は、従来のソフトウェア経済とは根本的に異なり、限界コストが低下するのではなく継続的に上昇する投資パターンを示すことが、業界アナリストの間でも広く認識されています。

そのため、生成AIは技術上の課題であるだけでなく、経済上、アーキテクチャ上の課題でもあります。

ガバナンスの対象をモデルから 意思決定へ移すべき理由

生成AIに関する議論の多くは、どのLLMがより正確か、速いか、トークン単価が安いかというモデル選定に集中しています。これは重要ですが、本質的な問題を見落としています。コストとリスクを大きく左右するのは、モデルの選択よりも、呼び出し頻度と意思決定のコンテキストです。

ここでいう「意思決定」とは、フロー内の単一の分岐点ではありません。プロンプトの構築、検索、推論ループ、ツールの使用、エージェントによる反復を含み得る、エンドツーエンドの実行パスを選択することです。実行が始まると、トークンは継続的に消費され、その実態はほとんど見えなくなります。したがって、経済的リスクは個々の意思決定ノードに限定されません。ワークフロー全体を確率的システムに委ねる行為そのものにあります。

ガバナンスがなければ、企業は「至る所でLLMを使う」アーキテクチャへと流れがちです。

ルールや小規模言語モデル、決定論的ロジックで十分な場合でも、大規模モデルが呼び出されます。トークン消費量は予測不能になり、予算は膨張し、持続可能性への影響も積み重なります。

トークンは実行時に動的に消費されるため、予算策定や調達によってガバナンスを事後的に適用することはできません。意思決定が行われる時点で適用する必要があります。

これにより、ガバナンスの対象はモデルレイヤーから意思決定レイヤーへ移ります。

意思決定の階層化: 予測可能なAIの実践

トークンエコノミーへの実用的な対応策が、意思決定の階層化です。「どのモデルを使うべきか」と問う代わりに、複雑性やリスク、ビジネス価値に基づいて意思決定を分類します。



決定論的で変動の少ない意思決定はルールで処理し、トークンコストをほぼゼロに抑える



小規模でタスクに特化した予測モデルにより、中程度の複雑性へ効率的に対応



曖昧性、情報の統合、推論が必要な場合にのみLLMを使用



継続的な推論に見合う価値がある場合にのみ、エージェント型ワークフローを使用

意思決定を段階的に高コストな階層へルーティングすることで、企業は実行時に、要件を満たす中で最も低コストなインテリジェンスを一貫して適用できます。

ここで予測可能なAIが実際の運用に組み込まれます。LLMを避けることが目的ではありません。エスカレーションを明示し、ガバナンスと説明責任を確保することが目的です。このルーティングを個々のアプリケーションに組み込まず一元管理することで、ケースごとに再実装するのではなく、企業全体で標準化して再利用し、一貫したガバナンスを実現できます。

経済ガバナンスそのものをAIに委ねられるという反論もあります。一見すると、コスト管理を事後対応から事前対応へ変えることで、問題を解決できるように見えます。しかし実際には、品質、網羅性、正確性とコストをどう両立させるかという最も重要な経済的意思決定を、不透明で非決定論的なシステムに委ねるだけです。

組織にとってガバナンスの対象は成果ではなく意図になり、モデルがその意図を適切に解釈することを期待するしかなくなります。

ビジネスとテクニカル両方のアーキテクチャの連携が不可欠

これは、テクニカルアーキテクチャだけの問題ではありません。



ビジネスアーキテクチャの観点から、企業が定義すべきこと:

- 高コストな推論を正当化できる意思決定の種類
- トークンとエネルギーの消費を上回るビジネス価値が得られる領域
- 高コストまたはエージェント型のAIパターンを誰が承認するか
- 意思決定単位のコストを成果やKPIとどう整合させるか



テクニカルアーキテクチャの観点から、プラットフォームに求められる機能:

- ルーティングとポリシーを一元的に適用
- 意思決定単位、ワークフロー単位でトークン使用量を測定
- コスト、レイテンシー、リスクのしきい値を適用
- エスカレーションが妥当でない場合に、決定論的なフォールバックパスを提供

Pegaの優位性は、この2つのアーキテクチャが同じ場所で交わることです。Pegaのプラットフォームは、ポリシー、ガバナンス、実行をすでに大規模環境で運用可能にしています。トークンエコノミーによって、この連携は避けて通れないものになります。

この連携には、明確なオーナーシップも必要です。AIの経済性を統制するには、AI利用を単なる技術課題として扱うのではなく、アーキテクチャ上の選択を財務的影響へと置き換え、意思決定の設計、意思決定単位のコスト、ビジネス価値を結び付けられるチームが必要です。





迫りくる現実：経済ガバナンスから規制圧力へ

現在、多くの企業では、生成AIの利用に対する実効性のあるガバナンスがほとんどありません。LLMはワークフロー、アプリケーション、製品に直接組み込まれていますが、いつ、どの程度の頻度で、累積コストがいくらになるまで呼び出すかを実行時に制御する仕組みは限定的です。支出が実験段階にとどまる限り、こうした統制の欠如は容認されます。しかし、その状態は長く続きません。

近い将来、避けて通れない最初の段階は、規制ではなく経済ガバナンスです。生成AIが実証実験から大規模運用へ移行するにつれ、経営陣は次の根本的な問いへの回答を求めるようになります。

AIの支出が利用量を上回るペースで増えているのはなぜか。

高コストな推論が本当に正当化される意思決定はどれか。

意思決定単位のコストはいくらか。処理量の増加に伴い、そのコストはどう変化するか。

AIの導入拡大に伴うコスト構造を、どの程度予測できるか。

トークン消費量は変動し、増加は非線形で、実行時に発生するため、従来の財務統制ではこれらの問いに答えられません。主に制御できるのは、呼び出し前の段階です。予算策定、調達、事後的なFinOpsレポートでは、すでにトークンが消費された後となり、対応が遅すぎます。

そのため企業はまず、実行前にAI利用を制御する仕組み、すなわち経済ガバナンスを導入せざるを得なくなります。これは、要件を満たす中で意思決定を最も低コストなインテリジェンスヘルペティングし、しきい値を適用するとともに、高コストなモデルへのエスカレーションを暗黙的ではなく意図的に行うことを意味します。

意思決定単位のコスト、つまり成果を生み出すために必要な推論、検索、オーケストレーションの総コストが、予測可能なAIのユニットエコノミクスになります。モデル価格を巡る抽象的な議論を、運用上の実態へと置き換える指標です。

こうしたコスト管理の規律は、やがて規制圧力に対応するための基盤となります。AIのエネルギー消費が企業規模、さらには国家規模で無視できない水準に達し、自動化による労働力の変化と従来の労働所得を基盤

とする税収の減少を受け、公共収入モデルの見直しを求める圧力が高まる中、政府や規制当局がコストの透明性を土台に正式な要件を設ける可能性は、ますます現実味を帯びています。その対応には、炭素排出量に連動した報告、計算負荷の高いワークロードに対するエネルギー課金、生成AIをどこで、なぜ使用しているかの説明を組織に求める監査義務などが含まれる可能性があります。

そうなれば、実行時の経済ガバナンスを持たない組織は対応に苦慮します。意思決定レベルの制御がなければ、高コストまたはエネルギーを大量消費する推論が妥当で、必要かつ相応のものだったことを証明できません。コンプライアンス対応には、ダッシュボードだけでは不十分です。代替手段が評価され、その結果が適用されたことを示すアーキテクチャ上の証跡が必要になります。

この観点から見ると、予測可能なAIは将来の規制に対する事後的な対応ではありません。企業が混乱なくその将来像を実現するための前提条件です。まず経済ガバナンスを確立します。その結果として、規制に対する耐性が生まれます。



予測可能なAIの重要性が 今後さらに高まる理由

精度向上、ハルシネーションの低減、プロンプトフレームワークなど、現在の生成AIの差別化要因の多くは、いずれ収束していきます。モデルとツールが改善されるにつれ、こうした差は縮まります。

しかし、AIの経済性が収束することはありません。

トークンコスト、インフラストラクチャへの負荷、エネルギー使用量、規制当局の監視は今後も続き、さらに強まります。

こうした懸念が一般化し、コモディティ化する前に、ガバナンスと経済性を軸として予測可能なAIを定義できる戦略的な機会は限られています。

これは、アナリストやアドバイザリー企業の調査ですで見られる、より大きな傾向とも一致します。AI導入初期の優位性は機能によって生まれますが、長期的

な優位性は運用規律、経済的統制、大規模環境でのガバナンスによって生まれます。

こうした環境で競争優位を得るのは、最も多くの生成AIを導入する組織ではなく、最も賢く、コスト効率よく導入する組織です。

意思決定レベルのガバナンスを基盤とする予測可能なAIは、財務、運用、持続可能性の各面で責任ある形で生成AIを大規模展開する手段を企業に提供します。

まとめ

予測可能なAIは、イノベーションを遅らせるためのものではありません。インテリジェンスを意図的かつ一貫して、経済的な意図をもって適用するためのものです。

モデルではなく意思決定を統制し、ビジネスアーキテクチャと技術的な実行を整合させることで、企業はコスト、炭素排出、コンプライアンスに対する統制を失うことなく、生成AIを大規模に活用できます。

この点において、Pegaのアーキテクチャは単なる選択肢ではなく、不可欠な存在です。





ありがとうございました

Pegaについて

Pegaは、エンタープライズ変革を支える業界有数のAI搭載プラットフォームを提供しています。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

© Copyright 2026 Pegasystems Inc.All Rights reserved.

すべての商標は、それぞれの所有者に帰属します。記載されている機能の開発、リリース、提供時期は、Pegasystemsの単独の裁量に委ねられます。Pegasystemsは、本情報に関して一切の責任を負いません。詳細については、Pegaの担当者にお問い合わせいただくか、pega.comをご覧ください。