

2026 年 9 月 18 日 全 9 頁

AI エージェント時代、 企業は AI コストとどう向き合うべきか

“FinOps for AI” でコスト管理と事業価値の評価に取り組む

イノベーション企画部

IT リサーチャー 田中 誠人

はじめに

生成 AI の利用拡大とともに、AI サービスのコスト構造は大きく変化しつつある。特に近年は、与えられた目的や設定された権限・条件に基づき、情報収集や分析、システム操作などを実行する AI エージェントの登場により、AI が処理する情報量や計算量が急速に増加している。こうした変化に伴い、トークン（AI が文章を処理する際の単位）消費量も大幅に拡大している。AI サービス事業者にとって、こうしたトークン消費量の増加は、計算資源や電力消費の増加、ひいてはサービス提供コストの上昇を意味する。その結果、高性能モデルや AI エージェントを中心に、利用量に応じて料金が発生する従量課金型のサービスが広がりつつある。

もっとも、本稿で注目したいのは AI サービス事業者の収益構造そのものではない。むしろ重要なのは、その変化が AI を利用するユーザー企業側にどのような影響を与えるかである。従量課金モデルが広がるということは、AI の利用量と企業の支出がこれまで以上に密接に結び付くことを意味する。AI 活用による生産性向上が期待される一方、利用実態を適切に把握・管理できなければ、AI 関連支出が想定以上に膨らむ可能性もある。また、AI 関連支出がどのような成果や事業価値につながっているのかを評価することは容易ではない。AI 活用を継続的な競争力につなげるためには、コスト管理に加えて価値評価の視点が重要となる。

そこで本稿では、まず AI サービスの料金体系やコスト構造が変化している背景を整理したうえで、利用企業に求められるコスト管理の考え方を考察する。特に、近年注目される「FinOps for AI」の考え方を取り上げ、AI 活用の拡大とコスト最適化を両立するための実践的なアプローチについて整理する。

AI エージェントが変えるコスト構造

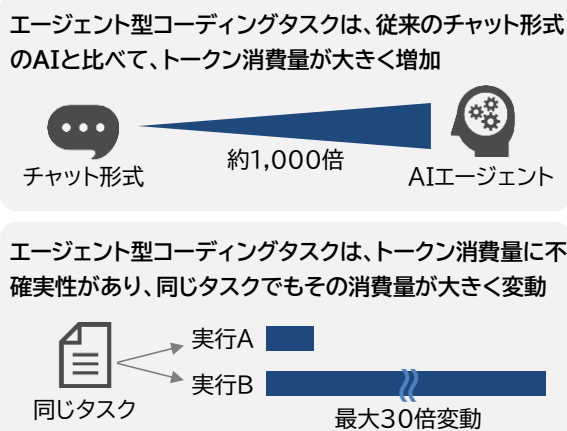
AI サービスのコスト構造が変化している一因となっているのが、AI エージェントの普及である。従来のチャット型サービスでは、利用者が入力した質問に回答を返す処理が中心だった。一方、AI エージェントは、与えられた目的や設定された権限・条件に基づき、情報収集や分

析、システム操作などを複数段階で実行する。その結果、利用者から見える依頼は1回であっても、実際には複数回の推論処理や検索、ツール実行が内部で実行される。また、コストは単純な利用回数だけで決まるわけではなく、利用するモデル、読み込む文書量、会話履歴、再試行回数、エージェント間連携などによって大きく変動する。

実際、スタンフォード大学などによる特定の実験条件下の研究では、SWE-bench Verified 上のエージェント型コーディングタスクについて、比較対象としたコード推論タスクやコードチャットよりも、トークン消費量が約 1,000 倍となるケースが報告されている。また、同研究では、同一タスクを反復実行した場合でも、実行条件等によってトークン消費量に最大 30 倍の差が確認された^{1 2}。これらの結果が示すのは、AI エージェントはより多くのトークンを消費するだけでなく、その消費量に不確実性があるということである（図表 1）。

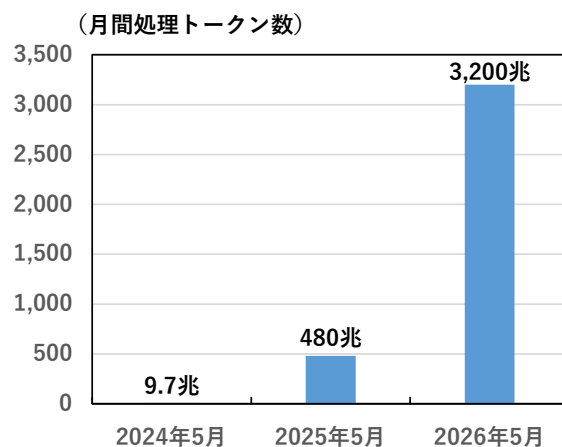
また、AI 企業によるトークン消費量の公表値を見ても、その消費量が大幅に増加していることを読み取れる。Google は 2026 年 5 月に開催された開発者会議で、自社の AI 関連サービスやモデルなどが処理するトークン数が月間 3,200 兆トークンに達し、前年の 480 兆トークンから約 7 倍になったと説明した³。これは AI エージェントによるトークン消費量の増加を直接的に裏付けるデータではないが、その傾向を示唆するものだと捉えることができる（図表 2）⁴。

図表 1 エージェント型コーディングタスクにおけるトークン消費量の比較



（出所）スタンフォード大学などより大和総研作成

図表 2 Google が処理するトークン数



（出所）Google より大和総研作成

こうしたコスト構造の変化を受けて、AI サービスの料金体系にも変化が見られる。Microsoft が 2026 年 6 月に一般提供を開始した Copilot Cowork では、ユーザー単位のライセ

¹ Stanford University “[How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks - Stanford Digital Economy Lab](#)”（2026 年 4 月 14 日付）

² SWE-bench Verified 上で 8 種類のフロンティア LLM による Agentic Coding Tasks を分析した研究。

³ Google “[I/O 2026: Welcome to the agentic Gemini era](#)”（2026 年 5 月 19 日付）

⁴ AI コストの構造変化については、下記レポートも参考にされたい。

新田 堯之、田邊 美穂「[エージェント化が迫る AI コストの二極化](#)」（2026 年 6 月 15 日付）

ンスに加え、エージェントの利用量に応じて料金が発生する仕組みが採用されている⁵。こうした動きは Microsoft に限ったものではない。Anthropic も、Claude Cowork や Claude Code などのエージェント機能を含む高度な利用について、利用量ベースで管理する仕組みを提供している⁶。また GitHub Copilot は、2026 年 6 月 1 日から、従来のプレミアムリクエスト数を基準とする仕組みに代えて、入力・出力・キャッシュトークンの消費量等に基づく GitHub AI Credits を用いた料金体系へ移行した⁷。

ユーザー企業にとってみれば、このような従量課金モデルはコスト管理の不確実性を生むものになる。そのため、利用実態を可視化し、発生する費用を継続的に把握・管理していくことが必要だ。利用状況を適切に把握できなければ、AI 活用の拡大とともに支出が想定以上に膨らむ可能性もあるだろう。

テクノロジー投資の価値を最大化する「FinOps」

AI コスト管理の重要性が高まるなか、その対応策として注目されているのが「FinOps（フィンオプス）」である。FinOps とは、Finance（財務）と DevOps（開発・運用の連携手法）を組み合わせた造語であり、エンジニアリング部門、財務部門、ビジネス部門が連携しながら、テクノロジー投資の価値を最大化することを目的としている。

FinOps はもともと AI 向けに生まれた考え方ではなく、クラウドの従量課金モデルが普及するなかで生まれたものである。クラウドが普及する前の IT システムでは、サーバーやストレージなどの設備を購入することで費用の大部分が確定していた。一方、クラウドサービスでは利用した分だけ料金が発生する従量課金モデルが一般的であり、利用状況によって毎月の費用が変動する。その結果、企業は従来よりも継続的かつ動的なコスト管理を求められるようになった。

FinOps はこうした状況に対応するための考え方だ。FinOps の普及・標準化を推進する業界団体である FinOps Foundation は、FinOps を「テクノロジー投資のビジネス価値を最大化し、エンジニアリング、財務、ビジネス部門の連携を通じて、データに基づく意思決定と財務上の説明責任を実現する運営フレームワークおよび文化的実践」と定義している⁸。

図表 3 は、FinOps Foundation が提唱する「FinOps Framework」の全体像を示したものである。これは組織が FinOps に取り組むうえで、どのような体制、役割、活動を整備すべきかを体系的に整理したフレームワークである。図の上段から順に見ていくと、まず最上段には企業の事業戦略（Business Strategy）と技術戦略

（Technology Strategy）が配置されており、FinOps がその両者を結び付ける役割を担うことを示している。その下には、どのような支出領域を対象とするかを表す

⁵ Microsoft “[Copilot Cowork is now generally available | Microsoft 365 Blog](#)”（2026 年 6 月 16 日付）

⁶ Anthropic “[Manage usage credits for Team and seat-based Enterprise plans | Anthropic Help Center](#)”（2026 年 5 月 19 日付）

⁷ GitHub “[GitHub Copilot is moving to usage-based billing - The GitHub Blog](#)”（2026 年 4 月 27 日付）

⁸ FinOps Foundation “[FinOps Framework Overview](#)”（2026 年 8 月 31 日閲覧）

「Scope」、関与する部門・関係者を示す「Personas」、実践すべき活動を示す「Domains」と「Capabilities」が整理されている。

図表 3 FinOps Framework



(出所) FinOps Foundation, “FinOps Framework” を基に転載 (CC BY 4.0)。原著作物およびライセンスへのリンクは脚注に記載⁹

フレームワークから読み取れる重要なポイントを 4 つ紹介する。まず、① **全社横断で取り組む運営モデル**という点だ。FinOps は特定部門の活動ではなく、エンジニアリング、財務、事業部門などが連携してテクノロジー投資の意思決定を行う全社的な運営モデルであることが示されている。次に、② **「把握」「評価」「最適化」が主要な活動である**ことも重要だ。フレームワークでは、利用状況とコストの把握、事業価値の評価、利用・コストの最適化が主要な活動の例として位置付けている¹⁰。さらに、③ **目的はコスト削減ではなく価値最大化**であることも意識すべきである。図の最上段で事業戦略と技術戦略が強調されているのもそのためであり、コストだけでなく、成長、収益性、生産性などの事業成果とのバランスを取りながら意思決定を行う考え方が示されている。最後に、④ **適用対象はクラウドから AI へ拡大**していることも注目だ。最新のフレームワークでは、クラウドだけでなく SaaS やデータ基盤など幅広いテクノロジー領域を対象としており、近年は AI サービスや AI エージェントの利用管理にも適用が広がりつつある。

AI エージェント時代に企業が直面している課題は、クラウド普及期に企業が経験した課題とよく似ている。利用量に応じて費用が変動すること、利用実態が見えにくいこ

⁹ [FinOps Framework](#) by [FinOps Foundation](#) (CC BY 4.0 ライセンスに基づく表示)

¹⁰ 「把握」「評価」「最適化」に加え、FinOps 活動自体を運営する領域も設けられている。

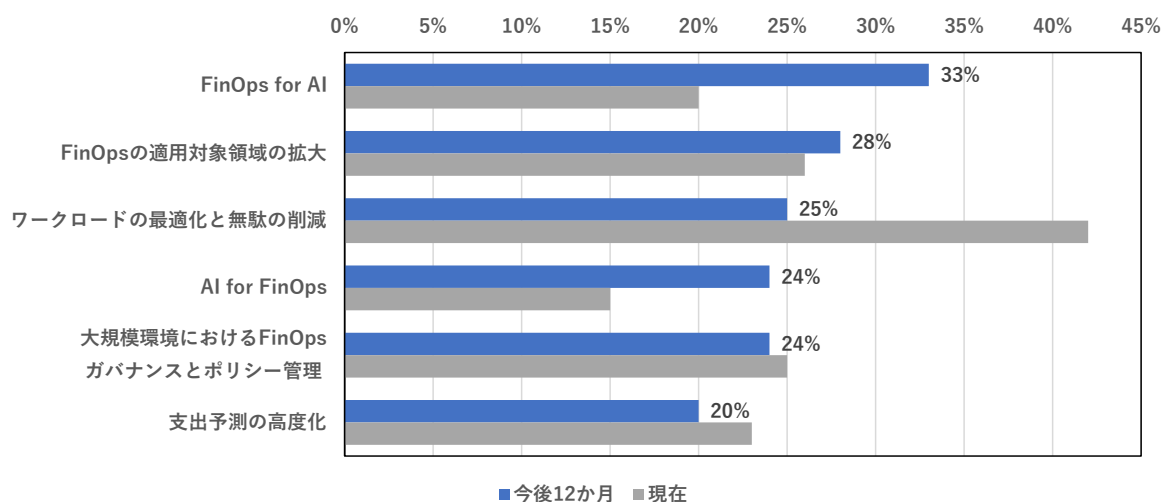
と、想定以上の支出が発生し得ることなどがその例である。前述のとおり、FinOps Framework でも AI が対象領域として位置付けられており、AI サービスや AI エージェントについても、利用量、コスト、事業価値を継続的に管理することが求められるようになっている。こうした AI 領域への FinOps の適用は、近年「FinOps for AI」と呼ばれている。次章では、この FinOps for AI について詳しく見ていく。

AI 活用とコスト最適化を両立する「FinOps for AI」

従来の FinOps がクラウド利用の最適化を主な対象としてきたのに対し、FinOps for AI は AI サービスや AI エージェントに焦点を当てた考え方である。FinOps の考え方を AI 領域へ適用し、AI 利用に伴うコストを可視化・最適化しながら、事業価値の向上を図る。

FinOps に取り組む企業を中心に、FinOps for AI への関心は急速に高まっている。図表 4 は、FinOps Foundation が実施した調査において、FinOps の領域で現在注力しているテーマと、今後 12 か月間で優先的に取り組むテーマを比較したものである¹¹。これを見ると、現在の優先事項としては上位ではないものの、今後 12 か月間の優先事項では、「FinOps for AI」が 33% となり、調査項目の中で最も高い回答割合を示している。AI 活用の拡大に伴い、AI コストの管理や最適化が企業にとって重要な経営課題として認識され始めていることがうかがえる。

図表 4 FinOps における現在および今後 12 か月の優先事項（上位 6 項目）



(注1) 今後 12 か月の優先事項について回答割合が高かった上位 6 項目を抽出し、それらの項目に対応する現在の回答割合を併記した。現在の上位 6 項目と今後 12 か月の上位 6 項目は必ずしも一致しない。

(注2) 複数回答可のアンケートであるため、回答割合の合計は 100%にならない。

(出所) FinOps Foundation より大和総研作成

では、FinOps for AI にはどのように取り組むべきだろうか。まず重要なのは、従来の FinOps の対象であったクラウドを AI へ置き換えるだけでは不十分だという点であ

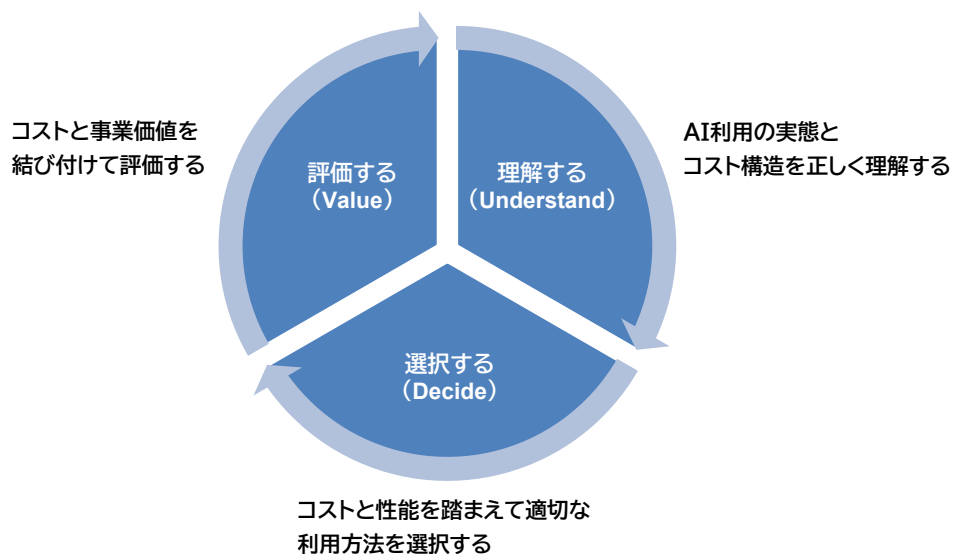
¹¹ FinOps Foundation “[State of FinOps 2026 Report](#)” (2026 年 8 月 31 日閲覧)

る。クラウドでは、サーバーやストレージ、ネットワークといったインフラ利用量が主な管理対象であった。一方、AI では、利用するモデル、入力・出力トークン数、推論回数、読み込むコンテキスト量、エージェントの実行回数など、多様な要素がコストに影響する。

もっとも、AI 活用の拡大に伴ってコストが増加すること自体は必ずしも悪いことではない。AI 活用が生産性向上や業務効率化、新たな価値創出につながるのであれば、そのための支出増加は合理的な投資と考えることもできる。重要なのは、発生したコストを抑制することではなく、そのコストによってどのような成果や事業価値が生み出されているのかを把握し、継続的に評価できる状態を構築することである。

こうした考え方をもとに、本稿では利用企業に求められる取り組みを 3 ステップで整理した（図表 5）¹²。AI 利用の実態やコスト構造を「理解する（Understand）」、コストと性能を踏まえて最適な利用方法を「選択する（Decide）」、そして事業価値の観点から投資効果を「評価する（Value）」の 3 ステップである。まず利用実態を理解し、そのうえで適切な選択を行い、最終的に事業価値の観点から評価するという一連のサイクルを継続的に回していくことが重要だ。以下、各ステップで取り組むべき事項について説明する。

図表 5 FinOps for AI で利用企業に求められる取り組み



（出所）大和総研作成

① 理解する（Understand）

AI 利用の実態とコスト構造を正しく理解する。AI では、利用するモデルやトークン量、コンテキスト量、エージェントの実行回数など、多様な要素がコストに影響する。

¹² この考え方は、下記の Web サイトで説明されている「FinOps for AI」の内容をもとに、ユーザー企業が利用しやすい形で当社が整理したものである。

FinOps Foundation “[FinOps for AI - FinOps Framework Technology Category](#)”（2026 年 8 月 31 日閲覧）

そのため、誰が、どの業務で、どの AI サービスを利用し、なぜそのコストが発生したのかを把握できる状態を構築することが重要となる。

例えば、同じ「レポートを作成する」という業務であっても、簡単な要約であれば少量のトークン消費で済む一方、大量の資料を参照しながら AI エージェントが情報収集や分析を繰り返す場合には、コストが大きく増加することがある。利用回数だけでは実際のコスト構造を把握できないため、まずは AI 利用の実態を可視化し、そのコスト構造を正しく理解することが FinOps for AI の出発点となる。

② 選択する (Decide)

コストと性能を踏まえて適切な利用方法を選択する。AI は、利用するモデルやサービス形態、システム構成によってコストと性能が大きく変化する。したがって、利用開始後のコスト最適化だけでなく、どのモデルやアーキテクチャを採用するかといった利用前の意思決定も重要な管理対象となる。

例えば、すべての業務に最高性能のモデルを利用することが最適とは限らない。用途によっては小型モデルや専用モデルで十分な成果を得られる場合もある。また、SaaS 型サービスを利用するのか、API を活用するのか、自社環境で運用するのかといった選択もコストと価値に大きな影響を与える。

③ 価値で評価する (Value)

コストと事業価値を結び付けて評価する。FinOps for AI の目的はコスト削減そのものではない。AI による生産性向上や業務改善、新たな価値創出によって十分な成果が得られているのであれば、コストの増加は必ずしも問題ではない。

例えば、AI 導入によって月間の利用コストが増加したとしても、提案書作成時間の短縮や問い合わせ対応件数の増加などによって、それを上回る効果が得られているのであれば、その投資は合理的であると考えられる。重要なのは、その支出によってどのような成果が生み出されているのかを継続的に評価し、より高い価値を生み出す領域へ投資を振り向けていくことである。

FinOps for AI の実践を支援する製品・サービス

ここまで、FinOps for AI への取り組み方について説明してきた。最後に、FinOps for AI の実践にあたって参考となる製品・サービスの例を紹介する。これらの製品・サービスは、クラウド事業者が自社のクラウド環境向けに提供する「i. クラウドベンダー型」と、複数のクラウドや AI サービスを横断的に管理する「ii. ベンダー横断型」に分けて考えることができる。なお、現時点では FinOps for AI を前面に押し出した製品があるというより、クラウドのコスト管理、AI の利用状況の監視、費用配賦、投資効果の評価といった複数の機能を組み合わせて FinOps for AI を実現するケースが多い。また、投資効果を「評価する (Value)」領域については、市場形成の途上であり、関連する製品・サービスはまだ限定的である。

i. クラウドベンダー型の製品・サービス

クラウドベンダー型の例として、Microsoft、AWS、Google Cloud が提供するコスト管理機能が挙げられる（図表 6）。これらは、各社のクラウド上で AI サービスを利用している企業にとって導入しやすく、既存の請求・監視・権限管理の仕組みを活用できる点に特徴がある。

図表 6 クラウドベンダー型の製品・サービス

企業名	関連するサービス	機能例
Microsoft	Microsoft Cost Management ／Azure Copilot	Azure および Azure OpenAI の利用量・費用の可視化、予算管理、アラート通知、費用配分。
AWS	Amazon Bedrock／AWS Cost Explorer／AWS Budgets	Amazon Bedrock を含む AWS サービスの利用量・費用の可視化、予算管理、ショーバック／チャージバック。
Google Cloud	Google Cloud Billing／FinOps Hub	Google Cloud サービスの利用量・費用の可視化、予算管理、コスト要因分析、最適化提案。

（出所）各社資料より大和総研作成

Microsoft では、Microsoft Cost Management などを通じて、Azure サービスの利用量や費用の可視化、予算管理等の機能を提供している。Azure OpenAI Service を含む AI 関連コストの分析や試算については、利用するサービスや構成に応じて関連機能を組み合わせることになる。AWS では、Amazon Bedrock の利用コストをユーザーやチーム、アプリケーション単位で把握できる¹³。AI 利用に伴う費用を組織内へ適切に配分し、ショーバックやチャージバック¹⁴へ活用しやすい仕組みを提供している。Google Cloud では、FinOps Hub を中心にコスト最適化を支援する機能を提供している。またプロジェクト単位の支出管理に関する機能として発表された Spend Caps¹⁵などにより、AI 利用の拡大に伴うコスト統制を支援する仕組みも示されている。

ii. ベンダー横断型の製品・サービス

複数のクラウドや AI サービスを利用している企業においては、クラウドベンダーごとの管理画面だけでは、AI 関連支出の全体像を把握しにくい場合がある。その場合には、ベンダー横断型の製品・サービスを利用し、クラウドや AI モデルを横断して管理することが選択肢となる（図表 7）。

¹³ 厳密には、呼び出し元のユーザーやロールに付与したタグを利用し、部門、チーム、プロジェクト、アプリケーションなどの単位で費用を把握・配賦できる。

¹⁴ ショーバックは利用部門ごとの費用を可視化する仕組み、チャージバックは利用部門へ実際に費用を配賦する仕組み。

¹⁵ 2026 年 4 月に限定公開プレビューとして発表され、その後パブリックプレビューへ移行。

図表 7 ベンダー横断型の製品・サービス

企業名	関連するサービス	機能例
IBM (Apptio)	IBM Cloudability／AI TCO & Usage／AI Value & ROI	AI 利用量・費用の可視化、費用配分、総保有コスト（TCO）分析、事業成果との関連付け。
Datadog	Agent Observability／Cloud Cost Management	AI 利用量・費用の可視化、トークン利用分析、性能・品質評価、クラウド・SaaS・AI 費用の統合分析。

（出所）各社資料より大和総研作成

まず、IBM 傘下の Apptio が提供するサービス群が挙げられる。IBM Cloudability はマルチクラウド環境のコスト管理を支援する製品であり、AI TCO & Usage では AI 利用量や費用に加え、クラウドや人件費などを含む AI 投資全体の総保有コスト（TCO）を把握できる。さらに、Apptio AI Value & ROI は、AI 関連支出を売上、コスト、スピード、生産性、リスクなどの成果指標と関連付ける機能として発表されている¹⁶。もう一つの例が、Datadog の Agent Observability／Cloud Cost Management である。Datadog は、AI アプリケーションや AI エージェントの性能、品質、コストを統合的に監視できる点に特徴がある。トークン利用量や推定コストなどを把握できるほか、クラウドや SaaS を含めた費用分析にも対応している。

おわりに

AI エージェント時代の到来により、企業における AI 活用は新たな段階へ入りつつある。AI の利用量や適用範囲が拡大するなか、関連コストを適切に把握・管理することは、企業にとって不可欠な取り組みになりつつある。

FinOps for AI は、こうした環境下で AI 活用とコスト最適化を両立するための実践的なアプローチである。ただし、本稿で紹介した製品・サービスからも分かるように、AI の利用量やコストの可視化、費用配賦、最適化を支援する仕組みが整いつつある一方、AI 活用が生み出した価値を測定し、投資判断に結び付ける手法やサービスは、いまだ発展途上にある。

AI 活用による効果は、業務時間の短縮や処理件数の増加のように定量化しやすいものだけでなく、意思決定の高度化や顧客体験の向上など、短期的には数値で捉えにくいものも含まれる。こうした多様な効果をどのような指標で捉え、コストと関連付けて評価するかが、FinOps for AI を実践する企業にとって次の課題となる。今後は、コストを可視化・抑制するだけでなく、AI 活用の価値を継続的に測定し、その結果を次の投資判断や資源配分につなげる仕組みを確立することが重要となるだろう。

¹⁶ Apptio AI Value & ROI に関する記述は、2026 年 8 月時点で公表されているパブリックプレビュー版の情報に基づく。一般提供開始は 2026 年第 3 四半期を予定している。詳細は下記を参照。

IBM [“IBM Introduces Apptio AI Value & ROI to Close the Gap Between AI Spend and Business Results”](#)（2026 年 8 月 6 日付）