AI エージェントがもたらす人工知能との新しい関係性
～利用から共生への変化～

株式会社第一生命経済研究所 客員研究員 月田 諒弥
（第一生命テクノクロス株式会社 DX 推進本部 デジタル企画部所属）

（要旨）

○ AI エージェントとは、近年急激な進化を遂げている人工知能における主要な技術であり、自律的な観点で判断やタスクを行えることから、日常生活やビジネスシーンでの活用範囲を飛躍的に拡大させている。

○ 人工知能の進化の歴史は長く、3 つのブームを経て徐々に人間の知性に迫りつつある。

○ その中でも、AI エージェントの諸技術は、人工知能に、自律性を伴う思考能力や問題解決能力、他の AI と協調して行動する能力を与えている。この点において AI エージェントは人工知能発展の歴史におけるブレイクスルーである。

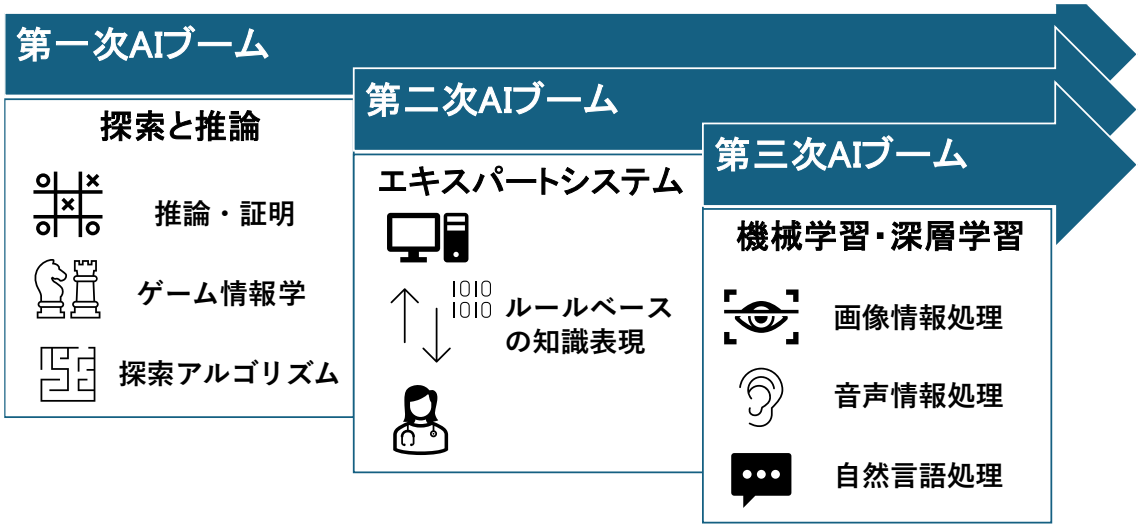
○ さらに AI エージェントに人格や個性をも与えることができるため、ますます人間に肉薄したタスクの実施や振る舞いのシミュレーションが可能となった。

○ AI エージェントの進化によって社会への AI 応用範囲が広がる一方で、懸念事項も増加する。人間社会と AI がどのように共生していくべきか検討することが急務であろう。

1. 人工知能と人間社会の関係性の変化

今日、人工知能（AI）の利用が急速に広がっている。2022 年末に登場した ChatGPT 以降、対話インターフェース型の人工知能をはじめ、各種ソフトウェアやチャットボット、Web アプリケーションなどあらゆる場面で活用が広がっている。この普及の背景には、対話型生成 AI という技術による飛躍的な言語および思考能力の獲得と、AI エージェントによる自律性の獲得が大きい。急激な進化を果たしたように見える AI であるが、その歴史は古く、AI という言葉が初めて提唱された 1956 年のダートマス会議から、第 1 次 AI ブーム（1950 年～60 年代）、第 2 次 AI ブーム（1980 年～90 年代）、第 3 次 AI ブーム（2000 年代後半以降）と複数のブームを経て、70 年越しで大きく花開いたといえる。以下では、これら 3 つのブームの概要とどのように人間の知性に近づいたかを簡潔に整理する（資料 1）。

資料１　AIブームの技術背景



（出所）総務省（2016）を参考に筆者作成

第１次AIブームでは、人間の知的な営みである「探索」「推論（帰納や演繹）」をコンピューターが担えるようになったことが大きな進歩であった。例えばLogic Theoristというプログラムにおいて、数学の定理を様々な公理の組み合わせで自動的に証明を行うことができるようになり、当時の計算機が持つ機能としては画期的な事であった（注１）。またこの時期に初期の自然言語システムとして代表的なELIZA（イライザ）も開発された（注２）。ELIZAの仕組みは基本的に機械的な推論による定型文の応答である。たとえばユーザがインプットした入力文を一部引用して返答することで、知性は無いが、人間との対話を比較的長く続けられたことから、あたかも会話が成立しているように見えるものである　（注３）。当時は、基本的に計算資源の制約などでボードゲームなどを解くような、トイプログラムと呼ばれる人工知能が中心であったものの、ELIZAをはじめとした現代のチャットボットに通ずる対話型AIの原型が、この時代において既に示されていたと考えられる。しかし実用的な展開につながらなかったため、第１次AIブームは終息していった。

第２次AIブームではエキスパートシステムと呼ばれる、専門家の知識をデータベースに格納し、専門家と同様の意思決定が行えるシステムに注目が集まった。例えば、医師などが診断を行う際に用いる前提知識と判断ロジックをコンピューター上に定義することで、医師同様に症状データから診断結果を人工知能が推論するという仕組みである（注４）。しかし、必要な専門家の知識が膨大であり、コンピューターに定義することに対する人間側への量的負荷が高いとともに、人間の意思決定や判断は極めて多様であることからルールベースでの推論では対処しきれない場面が多く、実用性には課題が残った。

第3次AIブームでは、大量のデータからパターンを自動的に認識する機械学習（マシンラーニング）と呼ばれる技術が発展した。特に、深層学習（ディープラーニング）と呼ばれるニューラルネットワーク（神経細胞から着想を得た計算アルゴリズム）を発展させた技術によってはるかにAIの応用範囲が広がった。深層学習では表現学習と呼ばれる意思決定や判断などを行う際の基準（特徴量）を自動的に学習できるため、これまでのAIにおける「局所的・限定的条件下でしか機能しない」というボトルネックを一部解消している。例えば、大量に動物の画像を学習させることで、ある画像に映っている動物がどの種類なのか判断をする際の基準を自動的に導き出すことを可能にする。このようなブレイクスルーにより、コンピュータービジョン（画像情報処理）や自然言語処理、自動運転など人間の認知機能・知的タスクをAIが代替する今日の最先端応用が実現している。

以上見てきたように、1950年代から大きく3次に及ぶブームを経て、AI技術は着実に人間の知性に近づいてきた。

しかし、ここまでの第1次～第3次AIブームに共通していることは、人工知能が処理すべきタスクは指示を行う人間から明示されなければ機能しないということである。すなわち、これまでの人工知能には人間のように問題解決の際、「ゴールを明確化すること」をはじめとした行動前の計画を立てる能力や、タスクを細分化する能力、問題解決に際して方策を変えるなど自律的に試行錯誤する能力が備わっていなかったことを意味する。また人間は複雑な問題解決を行う際、複数人から成るチームを構成し知恵を出し合いながら問題解決に挑む。このような他者との相談や議論といった人間の社会的な問題解決の営みについても、従来の人工知能が自律的に複数の人工知能と協力できておらず模倣に至っていない部分の一つである。

2. 生成AI、AIエージェントの登場

こうした背景の中、2022年末に対話型生成AIが登場した。対話型生成AIとは、自然言語による指示文（プロンプト）に応じて、自然言語や画像、コード、音声などでの出力を生成・返答するAI技術を指す。多くのソフトウェア製品やWebサービスに同様のAIが組み込まれ、文章執筆の支援や推敲、要約をはじめ、動画生成、デザイン支援まで用途は多岐にわたる。一方で課題もあった。対話型生成AIは指示文に忠実なため、人間による適切な指示出しやプロンプトエンジニアリング（どのようなプロンプトであれば最適行動を引き出せるか探索する工学的手法）が求められるようになったのである。

その課題に対応するように2023年から2024年にかけてAIエージェントと呼ばれる新しいタイプの人工知能が登場してきた。AIエージェントは、自律的に意思決定を行う人工知能技術全般を指し、これまでの課題であった「適切な指示出しがないと、最適行動がとれない」という課題が解消されている。その定義や従来の人工知能との

線引きは確立途上であるが、参考として AI エージェントサービスを提供するテクノロジー企業 3 社の定義を概観する。A 社は AI エージェントを「意思決定する主体」と位置づけ、エージェントは独立して、必要に応じて他のエージェントや人間と協力し目標を達成するもの（注 5）としている。B 社も「ユーザの代わりに目標を追求するシステム」と定義している（注 6）。C 社では「環境と相互作用し、データを収集し、そのデータを使用して自己決定タスクを実行して、事前に決められた目標を達成するためのソフトウェアプログラム」としている（注 7）。

3 社に共通するのは、AI エージェントを「目標達成のために自律的に判断、実行するシステムである」とみなしている点である。一方で、各社によって人間と AI エージェントの関係性における位置づけには微妙な差異が見られる。

生成 AI と AI エージェントはともに広く注目を集めている概念であるが、異なる背景を持つ。LLM（大規模言語モデル）に代表される生成 AI は、人工知能における推論アルゴリズムあるいはモデル群そのものを指している。一方、AI エージェントはそのアルゴリズムやモデルをシステムとして機能させタスクを遂行する包括的な枠組み（フレームワーク）となる。

学術的には、AI エージェントは「環境を知覚し、意思決定を行い、行動ができる人工的な存在」として定義されており、エージェント構成要素である「知覚」「脳」「行動」のうち LLM（大規模言語モデル）が脳として位置付けられている。そこに知覚や行動能力を加えることでさらに汎用性が高まることを期待されている。たとえば、知覚の機能を果たす部分にて外的環境の変化を察知し、脳機能を果たす部分にて思考や計画を行う。そして最後に人の手足のように行動する機能を果たす部分にて外部環境へ影響を与える（注 8）。この意思決定や行動能力は、LLM の固有の能力に加えて AI エージェントにおける制御機能により実現される。

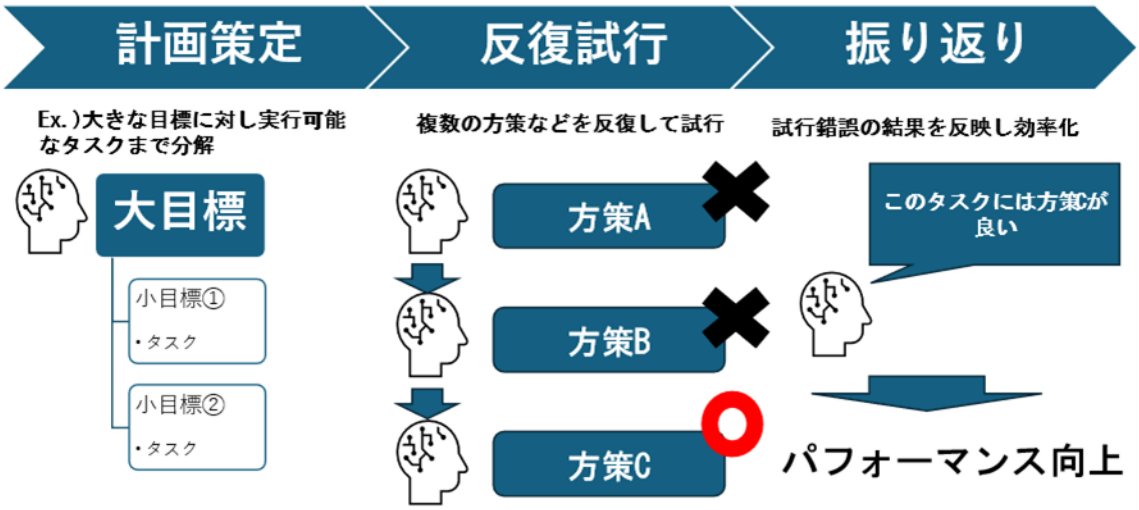
次の章では、具体的に AI エージェントがどのようにタスクを処理しているのか、そのプロセスと能力について詳細に解説する。

3. AIエージェントの特徴

AI エージェントの特徴は、ユーザの入力したインプットに対して、単にアウトプットを返答するのではなく、目標に向けた計画策定、目標に到達するための反復試行（イテレーション）、自身の振る舞いを評価しより良い行動をとるための振り返り（リフレクション）をするという自律的かつ探索的な能力が備わっている点にある（注 9 および資料 2）。

この能力を実現するために、AI エージェントは前述のような「知覚」「脳」「行動」という構成要素を持つ。それぞれの機能について説明をする。

資料 2　AIエージェントの学習機能



（出所）Masterman ら, 2024 を基に筆者作成

(1) 知覚

人間における目や耳などの感覚器官に対応した役割を果たし、外部環境の変化や情報取得を実施する機能である。ユーザからの指示文などのテキスト入力が代表的である。画像をはじめとする視覚情報、音声をはじめとする聴覚情報においても必要に応じツール利用などを行い、能力拡張して対応できる（注 8)。

また、必要であれば Web サーチなどを利用して関連情報を収集し、インプットとする（注 10)。

(2) 脳

AI エージェントにおける核であり、大規模言語モデル（LLM）で構成される機能である。人間の脳と同様に情報処理や行動の制御などを行う。この機能では自然言語による対話や知識の習得、過去の行動や観察などの記憶、問題解決のための推論と計画などの能力がある（注 8)。なかでも、AI エージェントにおける自律的な意思決定を支える推論と計画立案の機能について詳しく説明する。

知覚のプロセスを経て、十分に情報が集まった後、正しい意思決定や実行が精緻なものとなるべくいくつかのアプローチに従って目標達成のために推論・計画立案は実施される。

推論では、人間の知的活動における推論の 3 つの枠組み（演繹、帰納、仮説的推論:注 11）をエージェントにも同様に持たせ、人間同様の自律的な思考や意思決定を再現しようというものである（注 8)。推論を実現させる代表的な手法に Chain of Thought（CoT）が存在する。この手法には、人間が複雑な問題に対して段階的に物事

を考えるように、最終的な答えに至るための中間的なプロセスの例を与える方法（few-shot CoT、注 12）や、段階的な思考をとるように明示することで適切な推論が行えるという方法（zero-shot CoT、注 13）が存在する。推論は、後続の計画立案や行動を正しく行う上で重要であり、AI エージェントの推論能力が不足していると指示の誤った解釈や含意の考慮不足、文面通りの解釈のみに基づいた返答をしてしまうなどの課題につながる（注 9）。そのため、AI エージェントの構成要素の中でも重要な機能である。

計画立案は前述した推論能力を基に実現される能力で、代表的なアプローチとしては以下の 5 つである（注 9）。

・タスクの詳細化（より細かく処理可能なレベルへの分割）
・複数計画からの比較選択（計画の比較評価と選択）
・外部モジュールを利用した計画（API などのツール利用を含めた計画の実施）
・振り返りと改善（フィードバックによる計画の再検討）
・メモリを活用した計画（記憶を基とした方策の検討）

以上のような方策を使い分けて、ユーザの期待に応えられるような計画立案やタスク設定を行い、行動に移行する。

(3) 行動

この機能は、ユーザからの指示に対し、知覚したうえで推論および計画立案を行い、意思決定を経由して導き出された行動を最終的に AI エージェントがとる段階である。基本的なテキストやイメージなどの出力にとどまらず、他ツール、API 呼び出しなどの外部の情報を参照・利用することで、エージェントが行える行動空間を拡張している。例えば、検索エンジンにアクセスして情報を Web 上から検索することや、科学計算など専門的なタスクを行う際に python をはじめとする専門性の高いツールを利用してタスクを実行する。そうすることで、生成 AI が生成した情報ではなく、外部ツールで情報処理された結果を用いるため、事実ではない知識を生成 AI が生み出してしまう課題（幻覚：ハルシネーション）を防ぐ役割も併せ持つ。さらに、ツールを活用してタスクを遂行するエージェントは、指示方法の変化に対しても安定した正しい答えを出すことができ、生成された出力に対して、結果の導出方法を示せる点において信頼性が高い（注 8）。

エージェントにこのような拡張機能を持たせる方法としては大きく以下の 2 つである。

I. Function Calling

一つ目はFunction Callingと呼ばれている枠組みである。この枠組みでは、定義された関数や外部ツールをエージェント側が認識し、タスクに最適なツールを自動的に呼び出す形で機能する（注14）。

この枠組みでは一部を除き、基本的にはAIエージェント側に利用する関数やツールの定義を個別に実装しなければいけないという制約がある。

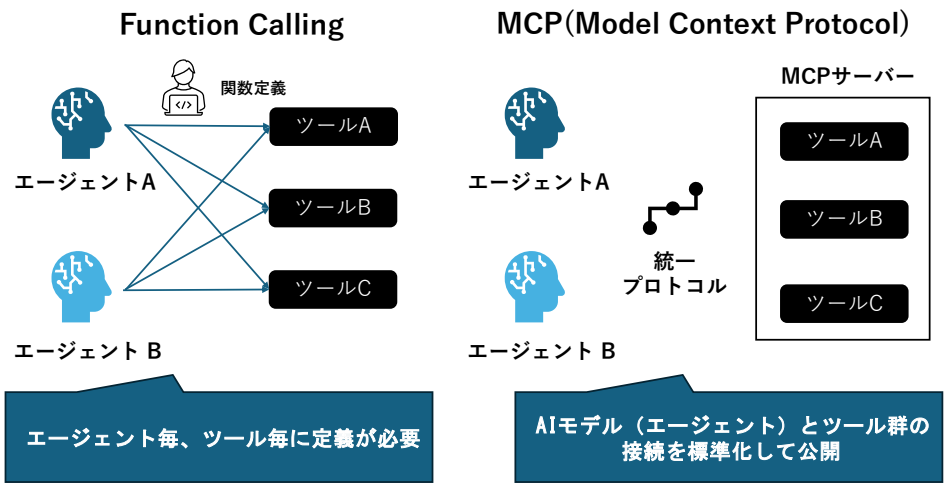
II. MCP （Model Context Protocol）（注15）

二つ目は、2024年11月に登場したMCPである。エージェントとツール群をつなぐオープンプロトコル（誰でも利用できる公開規格）であり、異なるエージェントから同一のツールやデータへのアクセスを容易にした。MCPサーバー側で関数やツールを管理するためエージェント毎の実装が不要になりツールアクセスが大幅に容易になる。

これら2つの手法の差異を資料3で示す。MCPによる標準化はAIエージェントの能力獲得をより簡便に、かつより多様化させることに貢献している。

このように、さまざまなAIエージェントが標準化されたツールへアクセスできるようになるというのは、人類が火や石器などの道具を使えるようになる進化に近しいブレイクスルーである。これにより情報の取得方法やアウトプット方法が無限に広がることを意味するからである。

資料3　Function CallingとMCPの違い



（出所） Open AIおよびMCPが提供するドキュメントを参考に筆者作成

行動においては、基本的にデジタル空間上がメインとなっているが、ロボティクス制御などと統合し、身体的行動として物理空間にも行動範囲を拡張することで、仮想知能と物理世界を結びつける将来像も期待されている（注 8）。

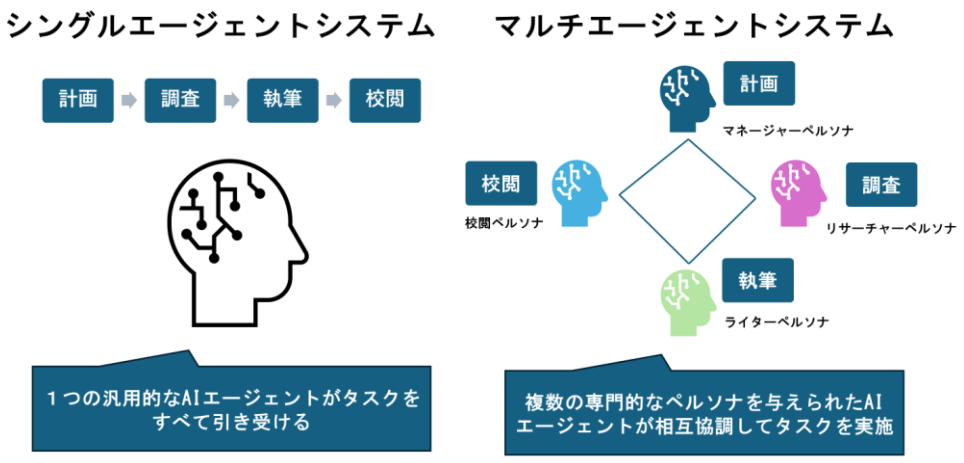
以上の 3 つの機能を経て、AI エージェントはまるで人間の知的行動と同様の問題解決能力を手に入れつつある。

4. AIエージェントにおける協調行動（マルチエージェントシステム）とペルソナの埋め込み

AI エージェントのシステムには 1 体のエージェントから構成されるシングルエージェントシステムと、複数のエージェントから構成されるマルチエージェントシステムが存在する。

マルチエージェントシステムの特徴は複数の AI エージェント、あるいは複数の AI エージェントと人間が相互作用してタスクを処理するという点である。このシステムの利点は複雑な問題を個々の専門家としての役割をもつエージェントに分担させることができる点である。資料 4 では、文章生成におけるシングルエージェントシステムとマルチエージェントシステムの違いを記載している。シングルエージェントでは、計画から調査、執筆、校閲までを単一のエージェントで実施しないといけない。そのため、各タスクにおける精度を担保することが難しい。マルチエージェントシステムではそれぞれのタスクをそれぞれのエージェントに割り当てることができる。こうすることで、各タスクにおいて、そのタスクに特化したツールや知識を持つエージェントが取り組むことができる。さらに、ペルソナ同士が相互作用によりフィードバックしあうことで並列処理が実現する。また各々のエージェントは分解されたシンプルなタスクに取り組む形となることで、実行精度が高まり堅牢なアウトプットを期待できるのである（注 9）。

資料 4　シングルエージェントシステムとマルチエージェントシステムの違い



（出所）筆者作成

また、マルチエージェントシステムには大きく二つのアーキテクチャ（垂直アーキテクチャおよび水平アーキテクチャ）が存在する。垂直アーキテクチャでは、一つのエージェントがリーダーとして機能し、他のエージェントがリーダーに報告する形式で機能する。この枠組みでは各エージェントでの役割分担がより明確となる。一方で水平型アーキテクチャでは、すべてのエージェントが対等であり、グループディスカッションのような形式で各エージェントの自発的なタスク処理が進行する（注 9)。

このようにエージェントが特定のルールに従って動作することでタスク効率を大幅に向上させる議論プロセスを「秩序ある協調」と呼ぶ（注 8）。

エージェント同士の協調においては A2A と呼ばれるオープンプロトコルが公開されており、この枠組みを利用すれば異なる企業が開発したエージェント同士を標準化された仕様の下で相互運用することができる（注 16）。

マルチエージェントシステムにはタスクに対しての問題解決のみならず、現実世界（人間社会）の現象や議論をシミュレートするという用途も期待されている。これまでの内容で、AI エージェントが人間と同様に計画、意思決定や推論、行動などの知的振る舞いをとることができることを説明したが、人間と AI の決定的な違いは個性やバイアスの有無である。人間が生み出したデータを大量に学習している生成 AI は基本的に中立的な観点での意見や意思決定、推論を実施する。したがって、問題解決や情報処理のユースケースであれば特段問題はない。しかしながら、例えばマーケティング用途で顧客目線の意見が欲しい場合、また市場動向のシミュレーションや社会的討論のシミュレーションなどを行いたい場合、より個性付けされた（時として非合理的な）人間らしい思考や意見、意思決定が必要となる。そのために、この複数エージェント同士のコミュニケーションや役割について、タスク効率の向上を目標とせず、人間の行動やコミュニケーションをより精緻に模倣、シミュレーションするという文脈で利用している例がある。これをマルチエージェントシミュレーションという。たとえば、M 社が提供している Python 言語向けのツールキットでは、問題解決を最優先とする AI エージェントシステムが目指す正確無比なふるまいではなく、人間の多様性や個性（ペルソナ）を再現することに重きを置いている（注 17）。このようなフレームワークでは AI エージェントにより詳細なペルソナ（国籍、年齢、職業、性格、信念、行動パターンなど）を設定することができ、そのペルソナに沿った人間らしいふるまいをさせることができる。さらにエージェントたちを複数人用意してディスカッションさせるなど、新しいアイデアの生成や市場調査、A/B テスト（マーケティングなどの領域において、商品の Web ページやデザイン、広告などを検討する際に複数の候補を消費者に提示しどちらがより集客などに効果があるかを判断するテスト）などへの活用が期待できる（注 17）。このような AI エージェントの使い方は、従来の問題解決のためのツールとしての利用のみならず、新規事業企画やイノベーションといった創発的なユースケースにも活用できる展望を示している。

5. AIエージェントのユースケース

AIエージェントに関係するユースケースは多く存在するが、ここでは企業の事例として、日本の大手自動車メーカーと総合ITベンダーのAIエージェント導入事例、クレジット会社A社のマルチエージェントシミュレーション実証事例、また個人向けにAIプラットフォームO社が提供するAIエージェントを紹介する

(1) 大手自動車メーカーの AI エージェント導入例

大手自動車メーカーT 社では会社の伝統的な経営手法を AI エージェント上で再現するシステムを構築している。このシステムでは自動車における、振動や燃費など様々な分野にわたる 9 つの専門家 AI エージェントが配置されており、ユーザの質問に各専門家の観点から回答してくれる。ナレッジベースには設計書や法規制、エンジニアの手書きの文書が格納されており、さらに DB にユーザと AI エージェントとの対話履歴や人間の専門家による評価が蓄積されることで、継続的な進化を期待できる設計となっている（注 18)。

(2) 総合 IT ベンダーの AI エージェント導入例

総合 IT ベンダーの F 社は、ビジネス業務向けに人々と協調して業務推進する AI エージェントを提供している。その一つの会議 AI エージェントでは、人間が行っている会議に出席し、会議の議論に参加する形で自律的にグラフの提示やデータの解釈の説明を実施する。会議進行のアシストや結論の導出サポートをすることで会議の生産性向上に貢献している（注 19)。このように、人間が企業において実施していた複雑な知的活動も AI エージェントによって代替されつつある。

(3) クレジット会社 A 社のマルチエージェントシミュレーション導入例

クレジット会社 A 社では顧客データを基に複数体の顧客ペルソナを作成した。ライフスタイルなどが埋め込まれているペルソナそれぞれをエージェントとして、ターゲットとなる商品に対しての仮想的なグループディスカッションを実施。商品の受容性がより高いペルソナをシミュレーションから特定することに成功した。実際にシミュレーション結果をもとにターゲティングを行い販売したことで、購入率が 1.7 倍に増加するという結果を得られている。このユースケースではマルチエージェントシミュレーションが実インタビューなどより優れている点として、①リクルーティングなどに時間のかかるグループインタビューを簡易かつ効率的にシミュレーションできる点、②実在する人間へのインタビューでは引き出しにくい批判の意見を回収できる点が指摘されている（注 20）。

(4) AI プラットフォーム O 社が提供する AI エージェントのユースケース

対話型生成 AI をビジネスや個人に広く提供する O 社では従来の対話型生成 AI のプラットフォームを拡張する形で、AI エージェントを提供している。この AI エージェントでは可能なタスクとして、献立の計画から材料の購入、旅行計画や各種予約など日常のさまざまな作業を AI が代替してくれることを示している（注 21）。AI エージェントがビジネスでの活用にとどまらず、今後あらゆる場面で利活用される未来が予想される。

6. 人とAIの関係性の変化と留意すべきリスク

前章までで触れた通り、AI エージェントにおいて社会性や人間の個性までも模倣するフレームワークが登場している。もはや人間と能力や振る舞いにおいて区別できない人工知能が登場するのも時間の問題であろう。ビジネス現場では、アメリカのテック大手企業の CEO が「いまの世代の CEO は『人だけ』を管理する最後の世代になる」と言及したように（注 22）、人と AI エージェントが対等な立場で仕事や交流をする未来が迫っている。

テクノロジー大手 M 社によれば、AI エージェントの企業への導入における人との関係性の変化は 3 段階あるという。人をアシストする存在として、チャット型の対話型生成 AI を使いこなす第 1 段階、エージェントとして、人の指示の下で自律して働く第 2 段階、人がエージェントの「上司」となり、エージェントが人間に近しい作業者として組織的に動く第 3 段階である（注 23）。すでに多くの企業が第 1 段階に到達し、第 2 段階、第 3 段階へと進もうとしている。

一方で人間と同等の認知能力をもつ AI エージェントの普及はホワイトカラーをはじめとする様々な職種の雇用を代替するリスクや、犯罪行為への悪用、差別の助長などの可能性を内包する。さらにエージェントの自律性の向上は副作用として様々な独自のリスク（AI への依存や過信、AI にゆだねたタスク実行に対する外部からの攻撃リスクなど）を増大させる。最新の研究では、AI エージェントの利益とリスクを比較し、これ以上の自律性を持たせるべきではないという意見も存在する（注 24）。資料 5 にある通り、AI エージェントの進化はあらゆる側面でトレードオフであるため、急速な開発に対する法的枠組みの整備と人間社会への影響の精査が急務となる。

7. まとめ

本稿では、人工知能の発展の歴史をなぞりつつ、AIにおける大きなパラダイムシフトであるAIエージェントの登場にフォーカスを当ててその特徴や機能、ユースケースやリスクを解説した。AIに対する理解を深めながら、機能拡張の末に来たる人間同様の知性をもつAGI（汎用人工知能）や人間を超えた知性を持つASI（人工超知能）（注 25）に対する人間側の備えを進めていくべきであろう。

資料 5 AI エージェントの進化によるベネフィットとリスク

価値	ベネフィット	リスク
正確性	応答や判断が正しければ、自律性の向上が有用な機能性の向上につながる。	LLM は正しく見えるが不正確な情報を生成するリスクがあり、自律性の向上によってリスクが増幅される。
支援性	人の作業を支援することで、より多くの機会と自由を提供しうる。（例：移動支援）	人間の代替による雇用や経済的影響、AI への過信・依存を招きうる。
一貫性	人間のように気分や体調に左右されず均一な応対を提供できる。	生成の多様性により一貫性の欠如が増加、見落とすと安全性の問題に発展。適切に一貫性を維持するには AI の発話の記録が必要なため、プライバシーの懸念がある。
効率性	自動化し、作業効率化することで人間がより価値ある仕事や家族に時間を使える。	多段の誤りの調査・修正が手間、精度と人の制御次第で効率性が低下。
公平性	発言時間の可視化で会議における公平性の偏りの発見・是正などに役立つ可能性。	AI エージェントにおける訓練データの偏りやサンプル偏差、雇用の AI による代替で不公平が拡大し得る。
柔軟性	様々なシステムと連携でき、ユーザの効率性向上や多様なニーズへの支援を提供。	他のシステムとの接続が増えるほど、悪意ある攻撃や意図しない誤作動の余地が増える。
人間らしさ	人の反応のシミュレーションや対話的な伴走（コンパニオンシップ）が可能になる。	擬人化による過信・依存・感情的巻き込みや、「不気味の谷」現象によりユーザに不快感を与える。
プライバシー	（提供者が監視できる範囲を除き）取引やタスクを高い機密性で扱える可能性。	AI エージェントが個人情報や各種アプリへアクセスする許可が必要で、委ねるほど漏えい時の被害が拡大。
関連性	個人最適でユーザに有益な情報が増加。	パーソナライズが偏見強化やエコーチェンバー（注 26）を生む。
安全性	ロボット型の AI エージェントによって危険作業（爆発物処理等）を人から代替。	行動の予測不可能性により、新規リスクをはらむ。行動制約の回避や重要な操作を警告なしで行う可能性。システムとの連携や人間の監視の欠如もリスク要因。
セキュリティ	（適切な設計であれば）プライバシー同様に機密保護を高めうる。	機密データを扱うことに加え、複数システムと連携するため、乗っ取りや情報流出・大規模自動攻撃のリスクがある。また完全自律型エージェントは新規コードを生成できるため、開発者が予測できない脆弱性を生み得る。
持続可能性	災害予測や交通最適化等で環境に寄与。	大規模言語モデルが炭素や水資源など環境に負荷を与える。
信頼	—	AI エージェントの自律化が進むにつれて、人間の信頼が他の価値に根差したリスクを増幅させる可能性。
真実性	—	ディープフェイクや誤情報の発生源であり、AI エージェントは個別最適化で思想の操作や詐欺へ悪用の恐れ。

（出所）Mitchell ら, 2025 の論文内容を引用し筆者作成

以　上

【注釈】

1) 人工知能学会HP https://www.ai-gakkai.or.jp/whatsai/AItopics5.html および Newell, A., & Simon, H. (1956). The logic theory machine--A complex information processing system. IRE Transactions on information theory, 2(3), 61-79.
2) Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.
3) 松原仁. (2011). チューリングテストとは何か (<特集> チューリングテストを再び考える). 人工知能, 26(1), 42-44.
Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation.
4) IT用語辞典 e-Words「エキスパートシステム」
(https://e-words.jp/w/%E3%82%A8%E3%82%AD%E3%82%B9%E3%83%91%E3%83%BC%E3%83%88%E3%82%B7%E3%82%B9%E3%83%86%E3%83%A0.html)
5) Microsoft「Azure AI Foundry Agent Service とは - Azure AI Foundry」
(https://learn.microsoft.com/ja-jp/azure/ai-foundry/agents/overview)
6) Google Cloud「AI エージェントとは定義、例、種類」 Google Cloud
(https://cloud.google.com/discover/what-are-ai-agents?hl=ja)
7) AWS「AI エージェントとは何ですか? - 人工知能のエージェントの説明」
(https://aws.amazon.com/jp/what-is/ai-agents/)
8) Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2025). The rise and potential of large language model based agents: A survey. Science China Information Sciences, 68(2), 121101.
9) Masterman, T., Besen, S., Sawtell, M., & Chao, A. (2024). The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. arXiv preprint arXiv:2404.11584.
10) Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., ... & Schulman, J. (2021). Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.
11) 帰納：観察されたケースに基づいて一般的な法則や結論を導き出す推論手法
演繹（えんえき）： 一般的に正しいとされる前提に基づいて結論を導き出す推論手法
仮説的推論：一連の事象に対して、観察から得られた最良の説明（仮説）を構築することで結論を導き出す推論手法
（Huang, J., & Chang, K. C. C. (2022). Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403.）
12) 問題解決における思考のプロセスを指示に加えることで、思考の連鎖という多段階の推論能力を持たせることができるプロンプト手法。通常、類題の問いと答えを

セットに与える。
指示文の例：「問題、テニスボールを 5 個持っている。さらに 1 つ 3 個入りのテニスボール缶を 2 つ買った。いま何個テニスボールを持っているか。答え、最初は 5 個持っており、新しく買ったテニスボールは 3 個×2 つで 6 個。よって 5＋6＝11 個である。では、カフェテリアに林檎が 23 個ある。昼食に 20 個使い、さらに 6 個買った。いま林檎はいくつあるか。」
（Wei, Jason, et al. “Chain-of-thought prompting elicits reasoning in large language models.” Advances in neural information processing systems 35 (2022): 24824-24837.）

13) few-shotとは異なり、「Let’s think step by step」という文言を指示文に追加することで、例示を与えずに段階的な推論を行わせるプロンプト手法
（Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.）

14) OpenAI “Function calling - OpenAI API”
（https://platform.openai.com/docs/guides/function-calling）

15) MCP “Introduction - Model Context Protocol”
（https://modelcontextprotocol.io/docs/getting-started/intro）

16) Google “Announcing the Agent2Agent Protocol (A2A) - Google Developers Blog”
（https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/）

17) Salem, P., Sim, R., Olsen, C., Saxena, P., Barcelos, R., & Ding, Y. (2025). TinyTroupe: An LLM-powered Multiagent Persona Simulation Toolkit. arXiv preprint arXiv:2507.09788.

18) Microsoft News Center Japan「トヨタ自動車、エンジニアの知見を AI エージェントで継承へ ― 競争力強化に向け革新的な取り組みを開始」
（https://news.microsoft.com/ja-jp/features/241120-toyota-is-deploying-ai-agents-to-harness-the-collective-wisdom-of-engineers-and-innovate-faster/）

19) 富士通 「AIが人と協調して自律的に高度な業務を推進する「Fujitsu Kozuchi AI Agent」を提供開始」
（https://pr.fujitsu.com/jp/news/2024/10/23.html）

20) MarkeZine 「JALカードとNTTデータ、生成AIを活用したマーケティング施策で購買率 3%向上」
（https://markezine.jp/article/detail/48143）

21) Open AI「ChatGPT エージェントが登場：研究とアクションをつなぐ新たな架け橋」
（https://openai.com/ja-JP/index/introducing-chatgpt-agent/）

22) Fortune “Marc Benioff says that from now on CEOs will no longer lead all-human workforces-enter the new era of AI coworkers”
（https://fortune.com/2025/01/24/marc-benioff-salesforce-human-workforces-ai-agents/）

23) ITmedia AI+「AIエージェントが隣にある世界」はすぐそこに 事例続々

Copilotの最前線をのぞく」
(https://www.itmedia.co.jp/aiplus/articles/2507/22/news003.html)
24) Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully autonomous ai agents should not be developed. arXiv preprint arXiv:2502.02649.
25) ソフトバンク「AGI（汎用人工知能）とASI（人工超知能）とは？ 従来のAIとの違いも解説｜ビジネスブログ」
(https://www.softbank.jp/business/content/blog/202310/what-is-agi)
26) 似たような考えや価値観を持つ人々が集まり、その中で情報のやりとりやコミュニケーションが完結すること
(引用:IT用語辞典 e-Words 「エコーチェンバー現象」)
(https://e-words.jp/w/%E3%82%A8%E3%82%B3%E3%83%BC%E3%83%81%E3%82%A7%E3%83%B3%E3%83%90%E3%83%BC%E7%8F%BE%E8%B1%A1.html)

【参考文献】

・総務省（2016）「情報通信白書 平成 28 年版」
・A. K. Pati (2025) “Agentic AI: A Comprehensive Survey of Technologies, Applications, and Societal Implications,” in *IEEE Access*, doi: 10.1109/ACCESS.2025.3585609.