１Ａ０３

DeepSeek の登場とその後への影響

○高橋　浩（B-frontier 研究所）

１．はじめに

生成 AI の短い歴史はモデル性能の急進によって彩られてきた。この傾向は、DeepSeek 社が最近発表した一連の製品によって再び起きている。但し、今回は性能向上だけではない。2024 年 12 月、同社は OpenAI の GPT-4o と直接競合する DeepSeek-V3 を発表した。このモデルは 2 ヶ月で学習され、学習に要した費用は約 560 万ドルと発表されて安さが話題になった[1]。続いて、2025 年 1 月 20 日、推論機能を強化した最新の Open AI-o1 と同等性能を達成した DeepSeek-R1 を発表した（図 1）[2]。この矢継ぎ早の製品発表は、NVIDIA の大幅株価低下を含め、世界に大きなインパクトを与えた。そこで、本稿は DeepSeek V3/R1 の理解と AI エージェントを含む各方面への影響について検討する。

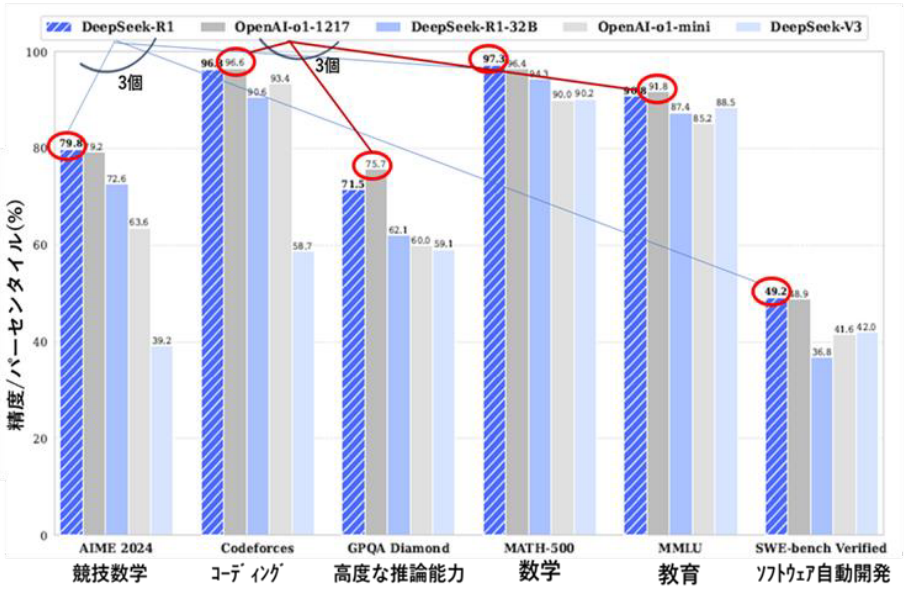

図１．DeepSeek R1 のパフォーマンス

DeepSeek は、AI を使用した金融取引を手掛けるヘッジファンド High-Flyer の共同創立者で、1 兆円超の資産運用をしていた梁文鋒氏によって 2023 年 5 月に設立された。DeepSeek は、米国からの AI 規制の環境下で ChatGPT 等米国製生成 AI に対抗するため、開発に必要な GPU など各種リソース消費を極力圧縮する小型化・低廉化を実現しながら高性能製品開発を目指した。そこで、この方針に貢献しうる多様な試み、例えば、古くから知られていたが顕著な成功を見なかった MoE(Mixture-of-Experts)アーキテクチャによる開発と運用など、コスト効率の高さと実用レベルの処理能力を兼備できる各方式実現に果敢に挑戦し DeepSeek V3 を実現させた。

続いて DeepSeek-R1 では、DeepSeek V3 をベースに推論機能強化を計るため、これも通常は使用する教師あり微調整（SFT）を敢えて使用せず、最初から純粋に強化学習(RL)のみで訓練することで、自然な推論能力獲得を実現した。但し、そこで発生した課題解消のため教師あり微調整（SFT）を追加活用するような工夫も行うことで、推論機能強化版として発表されたばかりの OpenAI 社の新製品 OpenAI-o1 と互角の性能を達成した。DeepSeek V3, R1 はどちらもオープンソースとして提供された。

２．DeepSeek の影響の枠組み

V3 と R1 は、一ケ月間隔でリリースされており、R1 は、V3 の後継製品と言うだけでなく、推論機能強化の目的を持っており、V3 は GPT-4o を上回り、R1 は高度な推論機能を持つ OpenAI-o1 と互角ということで、開発リソース少で小型化、低廉化を実現しながら高性能も達成という快挙を成し遂げた。これに最も貢献したと思われる MoE アーキテクチャは、モデルを幾つかの特定小モデル（数学用、コーディング用など）に分割し、これによって学習負荷を軽減させるもので、結果的に DeepSeek は数学とコーディング分野に特化し、従来、生成 AI が苦手としてきた計算精緻化が必須な分野へも生成 AI 適応を拡大させた（図 1）。

これまでの成果をまとめると、V3 で達成された小型化、低廉化、特化（専門化）に加え、推論機能を強化

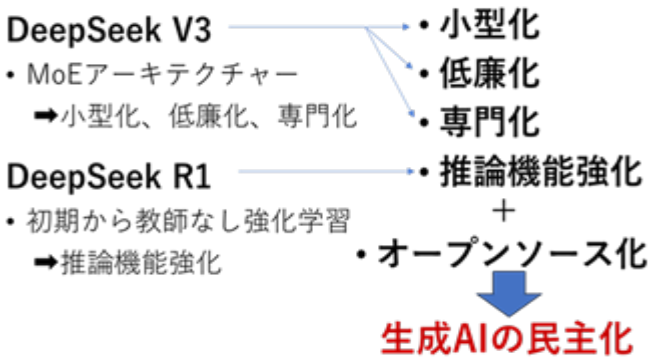

図２．DeepSeek 影響の枠組み

した R1 も加わり、それらが全てオープンソースとして提供されている。R1 は推論機能強化を目指して初めから強化学習を使い、教師あり微調整を使わなかった結果、ここでも開発リソースの縮小を達成した。これら全体の生成 AI 開発の仕組みと、全ての機能をオープンソース化した戦略は、「生成 AI の民主化」という新たな枠組みを提示したと言える（図２）。このことは単に新たな生成 AI 製品の登場に留まらず、AI エージェントを含む異なる世界を切り拓く基盤となることを示唆する。本稿はこのような認識で以降の検討を行う。

「生成 AI の民主化」は何を引き起こすか？まず、小規模デバイスで生成 AI の動作が可能になる。結果、生成 AI 機能搭載を前提とした多様なデバイスおよびそのコアとなる多様な AI チップが開発される。そうなれば、実行できる生成 AI 機能は階層化する。例えば、ハイレスポンスの生成 AI チャットを主体とした小デバイス向け機能から、推論機能を活用した AI エージェント向け機能、および大規模な推論機能提供環境の充実などである。AI エージェントの側面を更に補足すると、従来型エージェントシステムのエンジン（ルールベース）を LLM に置き換えるエージェント AI システムを登場させ易くする。そうすることで、小型化、低廉化、専門化、推論機能強化にオープンソースの特性が、新たなエージェント AI システムの構築に貢献する。これに伴い、エッジデバイスからも、ハイレスポンス生成 AI が利用可能になると共に、高度な推論が必要なエージェント AI システムを含めた全体構成が視野に入る。結果、現在とは大きく異なる世界が拓けてくる。しかし、その一方、これらを構成する各要素の担い手は多様化し、従来とは異なる危険、あるいは負担となる新たな負の連鎖が拡散するリスクが顕在化する。

このような状況を検討するため、(1) そもそも多様化している生成 AI はどのような特徴を持つか？(2) DeepSeek 登場は生成 AI にどのような影響を与えるか？(3) 複雑化する規制環境において生成 AI を如何にガバナンスするか？を考える。

生成 AI は、図１に示すような各種ベンチマークだけでは把握できない多様な要因で構成されている。そこで、代表的生成 AI を多様な比較尺度（表１の左欄）で比較し、各生成 AI モデルの特徴と限界を示す[3]。そうすると、各モデルは重点の置き方に相違があり、典型的には専用志向と汎用志向の 2 方向があることが分かる

表１．各モデルの特徴と限界

比較尺度	DeepSeek	ChatGPT	Claude	Qwen
ゼロショット学習と少数ショット学習	ゼロショット機能は強力だが、少数ショットへの適応性は限定的。△	GPT-4アーキテクチャを活用した優れたゼロショット学習と少数ショット学習。〇	ゼロショットは良好だが、数ショットではパフォーマンスがChatGPTに遅れをとっている。△	中程度のゼロショット。少数ショットのタスクでは苦労。×/△
バイアスと公平性の評価	バイアスを緩和する技術は限られており、公平性は現在も改善中 △	高度なバイアス検出と軽減機能を備えているが、完全にバイアスが排除されているわけではない。〇/△	倫理的な整合性に重点を置いているが、バイアスの軽減は依然として進化中。〇/△	基本的なバイアス評価。高度な公平性ツールは欠落。×/△
解釈可能性と説明可能性	解釈可能性の限界とブラックボックス性 ×/△	中程度の解釈可能性。一部の説明可能性ツールを使用。△	透明性への重点化により、解釈可能性が向上。〇	解釈可能性が低い。説明のためのツールがほとんどない。×
敵対的入力に対する堅牢性	中程度の堅牢性。高度な敵対的攻撃に対して脆弱。 △	広範囲にわたる敵対的学習と微調整による高い堅牢性。〇	中程度の堅牢性があるが、ChatGPTほど強力ではない。△	堅牢性が低いため、敵対的な入力に苦労。×
ドメイン固有のパフォーマンス	専門的トレーニングによりニッチな分野で優れた能力を発揮◎	汎用的なパフォーマンス。複数のドメインに適応。〇/△	倫理およびコンプライアンス関連の分野に優れている。〇	ドメイン固有の機能が制限される。×/△
多言語対応能力	多言語サポートは限定的。主要言語に特化。△	強力な多言語対応力があり、50以上の言語をサポート。◎	多言語サポートは中程度だが、ChatGPTほど広範囲ではない。△	基本的な多言語サポート。リソースの少ない言語では対応が困難。×/△
モデルの効率性とリソース利用率	非常に効率的で、リソースの少ない環境に最適化。◎	リソースを大量に消費し、かなりの計算能力が必要。×	効率的だが、DeepSeek に比べると若干最適化が劣る。〇/△	中程度の効率。パフォーマンスとリソース使用のバランスは良好。〇/△
時間的感度と知識保持率	知識の保持力が限定的。古くなった情報に苦労。×/△	時間的な感度が高く、最新のデータで定期的に更新。〇	知識の保持は中程度。更新は頻繁ではない。△	時間的な感度が低い。頻繁な更新がない。×
倫理的意思決定とコンプライアンス	基本的な倫理的整合性は有り。コンプライアンスは現在進行中。△	厳格な倫理ガイドラインがあるが、時折コンプライアンス上の問題が発生。〇/△	コンプライアンスを念頭に設計されており、優れた倫理的意思決定。〇	倫理的な意思決定への焦点が限定的。×/△

（DeepSeek は専用志向，ChatGPT は汎用志向）。これは専用志向が効率的・リソース小、汎用志向が大計算能力要・リソース大であり、汎用志向は用途が汎用であるが故にバイアスや公平性対応、敵対的入力に対する堅牢性などをより強化しなければならない側面があることが分かる。従って、比較評価尺度の網羅的カバーが重要なのではなく、今後は使用分野、ビジネス的狙いに合わせて多様化が急速に進むと推定される。

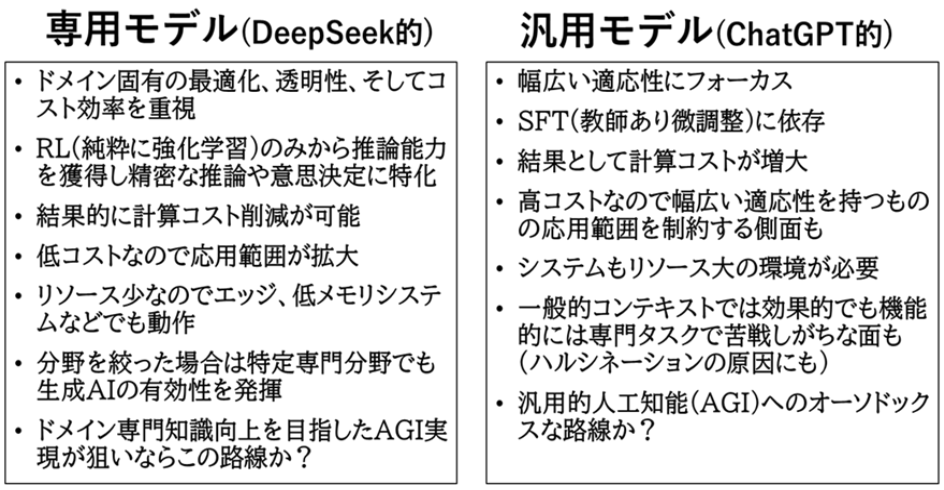

図３．専用モデルと汎用モデルの比較

３．生成 AI のリスクと民主化の影響

このような方向性を突き詰めると、DeepSeek は、汎用モデルと一線を画し、計算効率の高いアーキテクチャ、数学やコード生成などに特化、純粋な強化学習（SFT なし）で自律的に推論能力を強化と言うだけでなく「（大規模計算リソースなどの）ハードウェア環境に依存しない実装をオープンソースで提供」という目標を一定程度達成したと評価できる[4]。このことは、これまでの汎用利用/クローズドシステム/大規模リソース使用を前提としてきた ChatGPT に代表される既存生成 AI 側も、一貫した目標が異なる DeepSeek の登場とその目標のほぼほぼの成功に良い刺激を受け、これまでの方向の見直し、DeepSeek 的方向性への一部追随あるいは既存路線との共存など、新たな取組みへのキッカケになったと推定される。このような視点から専用モデル、汎用モデルの比較を図３に示す。

但し、「生成 AI の民主化」の世界は AI 由来の新たなリスクを増幅させる懸念がある。「生成 AI の民主化」は担い手の激増、小規模デバイスの登場などを通じて、よりシステム化された生成 AI 活用ならびに AI エージェント活用の機会の登場により、従来想定していた AGI/ASI 到来を前倒しする可能性があり得るが、その一方、従来の想定とは異なる多様なリスク拡散を助長する可能性がある。

このリスクを 2 つの側面で考える。一つは従来から喧伝されていた、人間の能力を超えた AI(AGI/ASI など)の登場に由来するリスク、もう一つは「生成 AI の民主化」で拡散が懸念されるリスクである。前者は人間の知能を超えた高度な AI(AGI/ASI など)登場によって引き起こされるカタストロフィー的な大規模事象（決定的リスク）、後者は一つ一つは小さな事象であっても、それらがボディブローのように積み重なって最終的に巨大な事象が発生するリスク（累積的リスク）である[5]。以後、当面現実的な累積的リスクのみを考える。このジャンルに入る小さなリスクの例としては、操作と欺瞞のリスク、誤情報と偽情報のリスク、悪意のある使用のリスク、差別やヘイトスピーチのリスク、監視、権利侵害、信頼の低下のリスク、環境リスクと社会経済的リスクなどが考えられる。生成 AI は基本的にこれらの行為を容易化する。累積的リスクは重大性は低いものの、重要な混乱が連続的に発生し、グローバルシステムのリジリエンスを侵食し、重要な社会経済的均衡を破壊する可能性がある[6]。

DeepSeek 起因の問題は、先行した汎用生成 AI と遜色ない機能を小型化、廉価で実現しているというだけでなく、規制や隔離が困難なオープンソースで提供されている点も重要である。即ち、最小限のリソースで誰でもアクセスできるオープンソース生成 AI モデルは悪

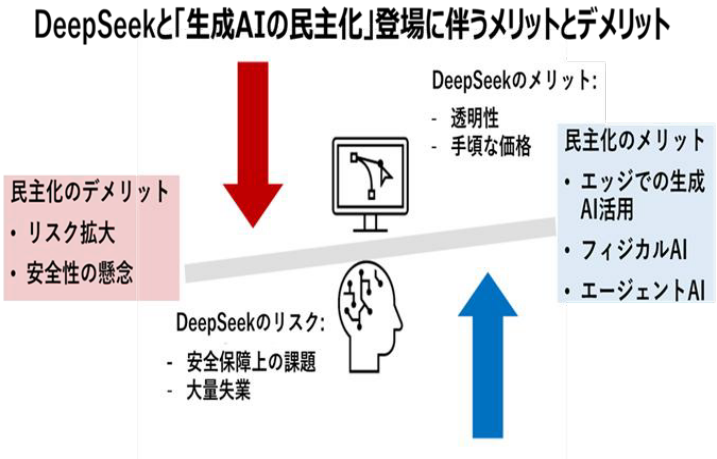

図４．生成 AI の民主化登場に伴うメリットとリスク

意のある行為者による悪用に対して障壁を低くしてしまうのである。

自動化されたサイバー攻撃から偽情報キャンペーンによる重要インフラの不安定化まで、現在でも既に発生している障害の規模や可能性が拡大し、累積的リスクを加速させる元凶になる懸念がある。そこで、オープンソース生成 AI のもたらす計り知れない恩恵と、そのリスクを軽減するための倫理的、規制的枠組みの構築が喫緊の課題となる。概念図を図 4 に示す。

4．AI エージェントが拓く産業の未来

以上の検討を踏まえ「AI エージェントが拓く産業の未来」を考える。生成 AI と AI エージェントの違いを明確に理解することから開始する必要がある。生成 AI は創造性が原動力で、核心は、既存のデータから学習し、その知識を使用して、人間の創造性を模倣した新しいオリジナルの出力を生成することである。一方、AI エージェントは自律的な問題解決者で、核心は、意思決定を行い、行動を起こし、変化する環境に適応することである。これだけ特性が明確に違うのに、既存エージェントシステム（ルールベース）のエンジンを LLM に入れ替えたエージェント AI システムが注目を集めるのには理由がある。

現行の生成 AI は理想（“夢”）を語るのは得意だが、次のような問題がクローズアップされていることが背景にある。例えば、旅行プラン作成を考える。生成 AI は希望に即して理想的プランを直ちに提供してくれる。しかし、空き室状況を確認したホテル予約、価格と日程を調整した上でのフライト選択など、手を動かし意思決定を行う具体的な行動が全く出来ない。これらの作業は全て人間に振られてくる。これでは AI が人間を作業者として使っているようなもので、AI と人間の関係の想定とは真逆の関係である。

このような生成 AI の限界が明確になり、この課題解決のニーズが先行しているので、創造中心の生成 AI と行動中心のエージェント AI は、必ずしも相性が良い訳ではないにも関わらず期待が先行している。あるいは連携不充分なのにエージェント AI システムへの期待が盛り上がっている。具体的には下記のような問題がある[7]。

- **真の自律性の欠如**: 生成 AI はトレーニングデータに依存しており、そのデータのバイアスや制限に縛られているので、真の意味での理解や自律性はない。
- **意思決定の不備**: 生成 AI はもっともらしい出力を作成することには優れているが、堅牢な推論能力が不充分なので、長期的一貫性を維持しつつ動的環境に適応する意思決定能力が足りない。
- **スケーラビリティと制御**: 生成 AI は計算コストが高く制御が困難なので、特定エージェントアプリに合わせて微調整して行動しても、多くの場合予測不可な結果が生じ、信頼性が損なわれるリスクがある。
- **倫理的およびセキュリティ上の懸念**: 生成 AI の確率的性質は変わらないので、誤解を招くコンテンツや有害なコンテンツ作成のリスクは残る。一か八かのシナリオを実行する訳には行かない。
- **証拠の欠如**: 生成 AI がエージェント AI のコンテキストで一貫して期待どおりに機能できることを示す経験的証拠はない。ケーススタディは逸話的なものであることが多く長期的存続には対応できない。

このような問題を可視化するため、GPT-4o、Claude-3 を用いて公開されているオープンソース・エージェント AI システムを 6 個取り上げ、体系的に分析した例が報告されている[8]。達成すべき目標を自律的に達成できたかどうかで失敗・成功を判定しており、失敗率がなんと、最高 87%、最低 41%、と極めて高い（図 5）。

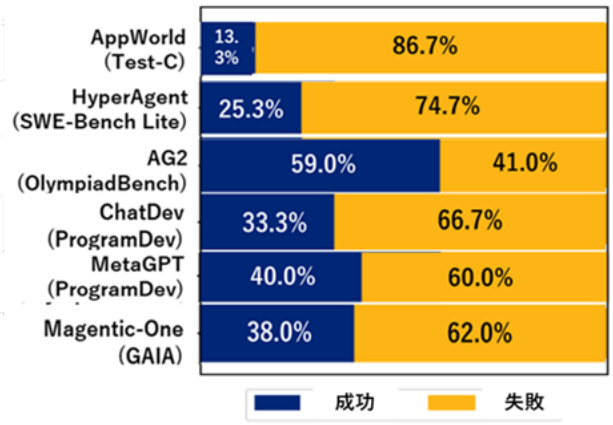

図 5. 典型的なエージェント AI システムの失敗率

このような結果を踏まえ、これらの失敗の原因を探求して、14 個の障害モードも特定している[8]。それらは仕様に関する問題（システム設計関連）、エージェント間の不整合に関する問題（エージェント間の調整）、タスク検証に関する問題（品質管理）の 3 カレゴリーに分類され、いずれも基本的なものばかりである。これは潜在的に根本的問題の存在を示唆する。

改善策としては、そもそも、エージェント AI システムにおける長期的に一貫性のある目標追求の取組みが複雑なプロセスを要求しており、このような高度な目標達成のためのベンチマークには、言語モデルレベルのベンチマークでは全く不充分である。現状は堅牢なエージェント AI システムの構築法が確立しておらず、それとセットの品質確保と検証のためのエージェント AI システム検証用ツールやベンチマーク開発も遅れている。それらの標準化も未定である。

このような状況下でホットイシュー「新技術が、製造・医療・交通・金融など多様な産業において、モノづくりやサービスの在り方、技術と技能の関係性などについて新たな価値創出の可能性」を考えるに際して 2 つの提案を述べる。

1) 今後の展開のタイムスケジュール
2) エージェント AI システムの類型化

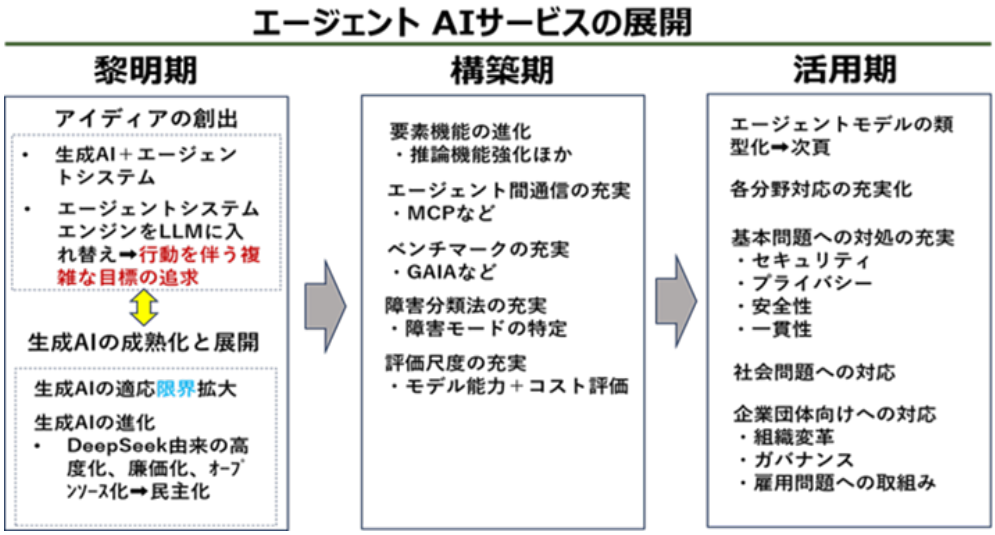

図 6．エージェント AI システムの展開予想

第一に、上述の検討を踏まえたタイムスケジュールを図 6 に示す。現在、市場では多数の取組みが行なわれているが、それらは、もし、ある企業が特定分野でエージェント AI 活用に成功し確実に生産性向上を達成してしまったら、同業他社への競争優位性が確立されてしまうのではないか？との認識の影響も想定される。しかし現状は構築期の初期にあり、本格稼働までには間があると考えられる。従って、必ずしもイノベーションフレームワーク確立の環境は整っていない。

第二に、このような状況でも個別産業の将来性を考える場合、エージェント AI システムの類型化も有効と思われる[9]。エージェント AI システムは既存のルールベースのものも含めて、適用分野の要件に合わせて多様なソリューションが求められる。従って、今後は個別の要件に合わせて多様なエージェント AI システムが共存し発展することが予想される。これをまとめて表 2 に示す[9]。表の区分は既存エージェントシステムも含む暫定的なもので、今後成功事例の登場によって詳細化と見直しが行なわれる。そこで、適切なフォローが居る。

最後に、DeepSeek は米国からの AI 規制で、開発リソースに大きな制約があった際、それを克服するアイディア群に挑戦するうちに、MoE アーキテクチャを本格的に実現することで所要の成果を得た事を振り返る。エージェント AI システムも期待は大きいものの、問題は多く、本格実用化に向けた課題も多い。生成 AI の場合は、データから学習するというメインルートが存在したので、改善にはハルシネーション緩和

表 2．エージェント AI システムの類型化

名称	内容	事例
単純な反射エージェント	• 事前に定義されたルールと即時データに厳密に基づいて動作する。特定のイベント条件アクションルールを超える状況には応答しない。	サーモスタット制御、特定キーワード検出でパスワードリセット、など
モデルベースの反射エージェント	• 特定のルールに従うのでなく、起こりそうな結果と影響を評価し、裏付けとなるデータを利用して、認識している世界の内部モデルを構築し、それを意思決定に活用する。	ナビゲーション、推奨システム、観測可能な症状からの診断システム、など
ゴールベースのエージェント	• 推論機能を活用して、環境データを評価するだけでなく、さまざまなアプローチを比較して、望ましい結果を達成できるように動作する。	自然言語処理 (NLP) やロボット工学アプリなど複雑なタスクの実行
効用ベースのエージェント	• さまざまなシナリオとそれぞれの効用価値や利点を比較し、ユーザーに最も多くの報酬を与えるものを選択して実行する。複数の目標が衝突する場合や不確実性下での微妙な意思決定に有用	金融取引システム、複数の好みを最適化する旅行計画アシスタント、など
学習エージェント	• 過去の経験から継続的に学習し、結果を改善する。経験を通じて特定の基準を満たすように学習要素を経時的に適応させて動作する。	相互作用履歴から学習するチャット, フィードバックで改善する推奨システム, 等
階層型エージェント	• 複雑なタスクを小さなタスクに分解し下位に割り当てる。各自は独立して動作する。上位エージェントは結果を収集し、集合的に目標を達成できるように調整する。	複雑なワークフロー管理、企業オートメーションプラットフォーム、など

に集中するなどの目標が立て易く、RLHF 法や Andrew Ng 教授による「データ中心 AI」概念[10]などの登場があった。

エージェント AI システムは、これに比べると、まだメインルートが定かでなく焦点が絞り切れていない。特に、エージェント AI システムは自律的に多様な目標を追求することが目的なので、目標の高度化により、通常はマルチエージェント構成に成ることが多い。これは問題解決プロセスをより複雑にすることを意味する。このような問題に立ち向かう一般的な方法は、精緻で大規模なソフトウェアシステム開発と同様に、個々の構成要素であるエージェントの精度を極限にまで高め、それらを束ねた集合体が期待どおりに動作しているかどうかを確認する検証法の精緻化を徹底することである。そのためにはベンチマークの工夫、検証に用いる独自データの充足など多様な準備が必要になる。

課題は重く作業は複雑なので、一種のすり合わせ的な精緻な議論と作業が必要になる。また、エージェント AI システムの達成目標は実用的でなければならない。いくら正しく動作してもコストが高くては実用にならない[11]。このような課題への取組みは DeepSeek でも類似の側面があった。日本流のすり合わせ技術やメンタリティが課題解決に貢献できる側面もあるかもしれない。

〔参考文献〕

[1] Aixin Liu et al., “DeepSeek-V3 Technical Report”, arXiv preprint arXiv:2412.19437, 2024.

[2] Daya Guo et al., “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning”, arXiv preprint arXiv:2501.12948, 2025.

[3] Rupesh Phogat et al., “A Comparative Study of Large Language Models: ChatGPT, DeepSeek, Claude and Qwen”, 3rd International Conference on Device Intelligence, Computing and Communication Technologies, Dehradun, India, 2025.

[4] Fnu Neha and Deepshikha Bhati, “A Survey of DeepSeek Models”, Authorea Preprints, 2025.

[5] Atoosa Kasirzadeh, “Two types of AI existential risk: decisive and accumulative”, Philosophical Studies, 1-29, 2025.

[6] Malik Sallam et al., “DeepSeek: Is it the End of Generative AI Monopoly or the Mark of the Impending Doomsday?”, Mesopotamian Journal of Big Data 2025, 26-34, 2025.

[7] Gonçalo Ribeiro, “Why 2025 Won't Be The Year Of Agentic AI”, Forbes, Jan 29, 2025.

[8] Mert Cemri et al., “Why Do Multi-Agent LLM Systems Fail? ”, arXiv preprint arXiv:2503.13657, 2025.

[9] Naveen Krishnan, “AI Agents: Evolution, Architecture, and Real-World Applications”, arXiv preprint arXiv:2503.12687, 2025.

[10] Andrew Ng, “Unbiggen AI”, By Eliza Strickland, 09 Feb 2022, IEEE Spectrum, https://spectrum.ieee.org/andrew-ng-data-centric-ai

[11] Sayash Kapoor et al., “AI Agents That Matter”, arXiv preprint arXiv:2407.01502 , 2024.