Marubeni
丸紅ワシントン報告

2025 年 9 月 4 日

AI は「スケール」で未来を切り拓けるか

スケーリング仮説、代替アプローチ、そして社会が信じるストーリー

丸紅米国会社ワシントン事務所
シニア・マネージャー（国際関係、政府関係担当）上原　聡
uehara-so@marubeni.com

- 近年の AI 開発においては、「スケーリング仮説」が大きな推進力となってきた。しかし、その先行きには不透明感が漂う。大規模言語モデルは依然として付加価値を生み出し、普及を牽引している一方で、規模拡大による成果が逓減している兆しも顕著になりつつあり、スケーリングだけでは汎用人工知能に到達できるのかという根本的な疑念が強まっている。
- テック業界の有力企業は、リスク分散を図りつつある。スケーリングのブームで最も利益を得ている NVIDIA のような企業でさえ、代替アーキテクチャへの研究を進めており、業界内では「計算資源の力押し」だけでは限界があるとの認識が広がっていることを示している。
- AI の未来を形づくるのは技術的ブレークスルーだけでなく、政府や市場が信じる「ストーリー」である。投資、政策、インフラ整備といった巨額の意思決定は、その物語の上に成り立っており、成功も失敗も社会全体の行方を左右する。

人工知能（AI）の進化は、これまでになく加速している。スタンフォード大学の HAI（Human-Centered AI）が公表した最新の「2025 年 AI インデックス報告書」[1]によれば、2024 年にはさらなる飛躍が見られた。AI システムは単なる性能向上にとどまらず、難度の高いベンチマークテストで顕著な成果を収め、コード解読などの領域でも劇的な進展を遂げている。

その中心にあるのが生成 AI、そして大規模言語モデル（LLM）である。ChatGPT、Claude、Gemini といった代表的な LLM は、技術の変革を最もわかりやすく体現している。かつては実験的と見られた技術が、いまや消費者、企業、政府にまで浸透し、標準インフラへと変わりつつある。実際、Microsoft の Copilot や Netflix の推薦システム、Zendesk の AI サポート、Adobe の Firefly、GitHub Copilot など、AI はすでに日常のサービスに深く組み込まれている。さらに企業は、試験導入から全社的な展開へと舵を切り、本格的な収益化に踏み出している。

同報告書が特に強調するのは、AI が多くのベンチマークで人間と肩を並べ、場合によってはすでに凌駕している点である。ただし、人間が持つ優位性も依然として残されている。自由な発想を伴う推論、長期的な計画、現実世界との身体的な関わりといった領域はその代表例である。いずれにせよ、AI が急速に発展する中で問われているのは「AI が私たちの仕事や創造、意思決定を変えるかどうか」ではなく、「その変化がどの程度の速さで、どのような形で社会を作り替えていくのか」である。

こうした将来像を語るうえで避けて通れないのが「汎用人工知能（AGI）」である。しかし AGI には明確な定義がない。狭義には「経済的価値を持つ大半の認知作業で人間を超えるシステム」とされ、広義には「人間並みの適応力や因果理解をあらゆる領域で発揮できる存在」と描かれる。どの定義を採用するかで、到達への道筋は大きく異なる。

ここで浮上するのが「スケーリング仮説」である。支持者たちは、膨大なデータと計算資源で既存の LLM を拡張すれば、推論能力や知識転移に加え、より高度な適応力すら引き出せると考える。彼

[1] スタンフォード大学 HAI の「AI インデックス報告書」は、研究・産業・政策・社会的影響といった分野にわたり、人工知能の世界的な進展を追跡する年次ベンチマーク調査である。（リンク）

らにとってスケーリングは、単なる人間労働の代替ではなく、人間レベルの特性を AI に宿す道筋である。

一方で、懐疑的な立場も根強い。懐疑派は、LLM はパターン予測には優れるものの、真の推論や現実理解、因果関係の把握には限界があると指摘する。この視点からは、AGI の実現には LLM の巨大化ではなく、新たなアーキテクチャの開発が不可欠とされる。

現状ではスケーリング楽観派が主導権を握っている。資本は GPU ファームや専用チップ、大規模電力インフラに集中し、「計算資源の規模が勝敗を決める」という単純で直感的なストーリーが投資を呼び込んでいる。その結果、企業はさらなる調達のためにこの物語を強調し、物語自体が自己強化的に循環している。

しかし行方はいまだ定まっていない。スケーリングが今後も成果を生み出し続けるとしても（実際すでに新しいビジネスモデルや科学的進展を支えてきた）、問いは残る。曲線は AGI に向かって上昇を続けるのか、それともどこかで頭打ちになるのか。そしてもし頭打ちとなれば、スケーリングを唯一の道と信じて資金や戦略を投じてきた投資家や政府、社会はどのような帰結を迎えるのか。

1．今日における「AI」とは？

いま「AI」と聞いて多くの人が思い浮かべるのは、ChatGPT や Claude といったアプリケーションである。これらはニューラルネットワークを基盤とする LLM の一種である。LLM の根幹にあるのは、与えられた語句の続きに最も適切な単語（トークン）を予測するという作業である。膨大なテキストを用いた自己教師あり学習を通じて、言語の統計的なパターンを精緻に捉える力を身につけてきた。さらに数十億規模のパラメータに拡張され、多様なデータセットで訓練が重ねられた結果、生成される応答は流暢さと文脈把握に優れ、あたかも「理解」や「推論」をしているかのように見えるまでに進化している。

もっとも、LLM は AI 全体の一部にすぎない。ニューラルネットワークを基盤とする他の手法も、さまざまな分野で画期的な成果を支えている。畳み込みニューラルネットワーク（CNN）は顔認識や医用画像解析、自動運転車の視覚システムに用いられている。リカレントニューラルネットワーク（RNN）は、天気予測や音声認識などで長年利用されてきた。強化学習（RL）は、AlphaGo が人間のトップ棋士に勝利した出来事や高度なロボティクスの進展を支える基盤となっている。応用分野は異なっても、その基盤にあるのはいずれもニューラルネットワークであり、驚くほど高精度な予測を可能にする強力なエンジンとなっている。

突き詰めれば、現代の AI は「データ」「計算資源」「アルゴリズム」という 3 つの要素に支えられている。

- データ：高品質なテキスト、画像、音声、動画、さらには人工的に生成されたデータを含む膨大な情報が、AI システムの学習には欠かせない。
- 計算資源：GPU や TPU[2]といった専用ハードウェアが、大規模ニューラルネットワークを訓練するための処理能力を提供する。2017 年にトランスフォーマー型アーキテクチャが登場して以来、巨大な計算クラスターによるスケーリングが AI の進歩を牽引してきた。

[2] GPU（グラフィックス処理装置）：本来は画像処理用に設計されたが、現在では多数の計算を同時並行で処理できる特性から、巨大なニューラルネットワークの学習に幅広く利用されている。TPU（テンソル処理装置）：Google が開発した特定用途向け集積回路（ASIC）の一種で、機械学習の計算処理を高速化するために最適化されており、とりわけ TensorFlow を用いた深層学習に特化している。

- アルゴリズム：モデルの設計そのもの、すなわちデータとパラメータ[3]の関係をどう構築するかという選択である。トランスフォーマー、CNN、RNN、強化学習はいずれも異なるアルゴリズム上のアプローチを示している。

2.「スケーリング仮説」と「ルカンの天井」

現在、OpenAI、Anthropic、Google、DeepSeek といった最前線の研究機関は、より強力な LLM の開発競争を繰り広げている。ここでいう「強力」とは、多様な認知スキルで高い能力を発揮し、場合によっては人間の水準に近づきつつあることを指す。その実現に向けた彼らの基本戦略が「スケーリング仮説」[4]である。すなわち、データ、モデルのパラメータ、計算資源を拡大すれば（GPU や TPU を搭載した巨大データセンターを通じて）、新たな能力が次々と引き出されるという前提である。この考え方に基づき、数十億ドル規模の投資がデータセンターの拡張や先端プロセッサの確保に投じられている。

OpenAI のサム・アルトマン CEO と Anthropic のダリオ・アモデイ CEO は、このスケーリング仮説を最も強く支持する人物として知られている。モデルの規模やデータ、計算資源を増やすことで、人間に匹敵する汎用的な知能（すなわち AGI）へと近づけると確信している。彼らは、スケールの飛躍ごとに一貫して新しい、しばしば予期せぬ能力が現れてきた事実を挙げ、さらなるスケーリングを進めれば、やがて広範に人間レベルに匹敵する性能を持つモデルが登場すると見ている。

これに対して、Meta のヤン・ルカン[5]は「このアプローチには天井がある」との見方を示す。モデルの規模を拡大すればパターン認識の精度は高まるものの、真の推論や理解にはつながらないと考えている。現実世界に基づく入力から学習し「世界モデル」を構築できるような、根本的に新しいアーキテクチャこそが不可欠だと主張している。

3．スケーリングの将来性とリスク

2024 年、OpenAI は高度な推論能力を備えたとされる「o1」と「o3 mini」を発表した。いくつかのベンチマークでは平均的な人間の水準を上回る結果を示したものの、その実態を詳しく見れば、依然として大規模な統計的パターン認識にすぎない。ルカンは、現実世界の仕組みに関する基盤的な理解（いわゆる「世界モデル」）が欠けている限り、こうしたシステムは脆弱なままであると指摘する。学習分布の外にある問題、たとえば稀な事例や反事実的なシナリオに直面すると、しばしばつまずき、いわゆる「ハルシネーション」（幻覚）を生み出してしまう。言い換えれば、LLM は依然として重要な意思決定を担える段階には達していない。

2024 年の終盤には、「スケーリング仮説」に対する疑念が一段と強まった。11 月には、モデルを大規模化しても成果が逓減しているとの報告が相次ぎ、研究者や投資家の間に動揺が広がった。一部からは、スケーリングがもはや期待されたペースでブレークスルーを生み出していないとの警告も出ている。Google のサンダー・ピチャイ CEO もこの点を認め、「初期の段階では計算資源を投じれば

---

[3] パラメータとは選択されたアーキテクチャにおけるモデルの規模を示す。パラメータが多いほど調整できる「つまみ」が増えることになるが、それをどれだけ効果的に活用できるかは、アルゴリズムと利用可能な計算資源に左右される。

[4] メディアでは一般的に「スケーリング則」（scaling law）とされる。

[5] ヤン・ルカンは Meta の主任 AI 科学者であり、チューリング賞を受賞した計算機科学者。ディープラーニングや畳み込みニューラルネットワークの先駆者として広く知られている。スケーリング仮説に対して最も著名な懐疑論者の一人であり、真の知能を実現するには新しいアーキテクチャが不可欠だと主張している。

急速に進展するが、次の段階に進むにはより深いブレークスルーが必要になる。壁に直面していると見ることもできる」と述べている。

2025 年 1 月には不安がさらに高まった。中国の新興企業 DeepSeek が、従来のごく一部のコストと計算資源で GPT 級の性能を実現する「R1」を発表したのである。このニュースは市場に衝撃を与え、NVIDIA の時価総額はわずか 1 日で数千億ドル規模の下落[6]に見舞われた。投資家は、GPU 需要が従来想定されていたほど飽くなきものではないのではないかと疑念を抱いた。その後数日のうちに、NVIDIA のジェンスン・フアン CEO やウォール街のアナリストたちが懸念を和らげようと動き、最先端モデルには依然として膨大な訓練と推論の計算資源が必要であると強調した。こうした説明により株価は持ち直し、むしろ DeepSeek の効率性がハードウェア需要を削ぐのではなく、AI 普及を広げる契機になり得るとの見方が強まった。

現在もなお、スケーリング楽観派は「データ、計算資源、モデルの規模をさらに拡大すれば、2030 年にも AGI に到達し得る」と主張している。一方で懐疑派は、それまでに数年から数十年を要すると見ている。いずれにしても否定できないのは、現在の AI ブーム（そして OpenAI、Anthropic、NVIDIA などの天井知らずの評価額）が一つの前提に依存している点である。すなわち、スケーリングが引き続きブレークスルーを生み出すとの信念である。この前提のもと、GPU 購入やハイパースケール・データセンター、さらには「Stargate」のような超大型プロジェクトに数十億ドル規模の資金が流れ込んでいる。

では、もしスケーリングが約束通りの成果をもたらさなかったらどうなるのか。すでに研究者や企業の一部は、スケーリング曲線がいずれ頭打ちになる可能性に備え、代替的なアプローチを模索している。OpenAI の共同創業者イリヤ・サツケヴァーは「2010 年代はスケーリングの時代だったが、私たちは再び“驚きと発見”の時代に戻っている」[7]と述べている。こうした探索的アプローチへの回帰は、研究開発リスクの上昇、開発期間の長期化、そして現在進行中の巨額 GPU・データセンター投資に対する収益性の低下を伴う可能性がある。とはいえ LLM は今後も商業的に有用であり、すでに収益を生むアプリケーションを支えているのも事実である。ただし多くの AI 企業の評価額には「AGI が近い」という前提が強く織り込まれており、しかも「完璧を前提とした価格付け（priced for perfection）」[8]と言えるほど強気に見積もられている。資本が単純な計算資源への投資から、代替アーキテクチャを試みる新興企業へとシフトする可能性を考えれば、このプレミアムが急速に崩れるシナリオも想定内だろう。

さらに根本的な問題は、チューリング賞受賞者や著名な言語学者、認知科学者を含む AI の先駆者たちの間でも、スケーリングによって本当に AGI に到達できるのかについて意見が一致していない点である。半導体、バイオテクノロジー、航空宇宙といった巨額投資を伴う他の先端技術では、基盤となる科学は比較的確立されており、賭けの対象は商業化にある。しかし AI は異なる。ここでは、学問の最前線でなお議論が続いている仮説に対して、数兆ドル規模の資金が投じられている。

---

[6] NVIDIA の株価は 17%急落し、米国企業史上最大となる 1 日の下落幅を記録、時価総額で約 6,000 億ドルが吹き飛んだ。

[7] 2024 年 11 月 15 日付、ロイター通信記事。（リンク）

[8] 「Priced for perfection（完璧を前提とした価格付け）」とは、投資家が企業の事業運営や成長見通しに一切の不備がないと織り込み、最大限の成長を前提に株価を評価している状況を指す。この場合、わずかな失望や期待の変化でも、市場は過剰な修正に動きやすい。

こうした不確実性は、投資家だけの懸念にとどまらない。ワシントンの政策当局者や戦略立案者にとっても、国家の競争力や地政学的戦略をこの技術の行方に結びつけている以上、極めて大きな意味を持つ。

４．AGI への別のアプローチ

もし「ルカンの天井」が現実となれば、それは単なる科学上の寄り道にはとどまらない。研究開発の方向性そのものを根本から転換させる事態となる。そして実際、AGI に至るための代替的な道筋はすでに模索され始めている。

研究最前線では、Meta が「Joint Embedding Predictive Architecture（JEPA）」[9]を開発している。これは世界の仕組みを内部に「地図」として構築することを目指し、単なるパターン記憶ではなく、AI が一種の内面的な世界観（メンタルモデル）を形成し、学んだ知識を異なる文脈に応用できるよう設計されている。人が学校で学んだ知識を仕事や生活に応用するのと同じ発想である。Google DeepMind も関連する試みを進めてきた。AlphaZero はチェス、将棋、囲碁で自己対戦のみを通じて人間を超える技能を獲得し、MuZero[10]はさらに進んでルールを与えられずに環境の内部モデルを自ら構築しながらゲームを習得した。さらに DeepMind の「Dreamer」[11]は、世界を簡略化したモデルを構築し、その内部で「想像上の」シナリオを走らせて学習を進める。可能な未来を思い描くこの能力は、効率性を高めると同時に、人間の計画や学習のプロセスに近づけるものとなっている。

業界の大手企業もこうした可能性を無視してはいない。スケーリングの波で巨額の利益を上げながらも、その限界に備えて記録的な収益を研究に振り向けている。中でも注目すべきは、スケーリング仮説の最大の受益者である NVIDIA 自身が代替路線を追求している点である。同社は主力の GPU に加えて、データ処理ユニット（DPU）やニューロモーフィック（脳型）研究といった新しいチップ設計に投資し、さらに CUDA や Omniverse といったソフトウェア・エコシステムを拡充して、シミュレーションや効率的なアーキテクチャを支えている。「ポスト・スケーリング」の未来に備える姿勢は明確であり、業界リーダーがスケーリングに全てを賭けるのではなく、知能の実現方法をめぐるパラダイム転換を見据えていることを示している。

ただし、これら代替アプローチが本格化した場合のリソース面への影響は不透明である。スケーリングに限界が訪れれば、巨大 LLM やそれを支える GPU・TPU への飽くなき需要は、新たなハードウェアへと移行する可能性がある。インテルの「Loihi」に代表されるニューロモーフィック・プロセッサ[12]はその一例であり、脳の構造をモデル化することで効率性の向上を狙っている。研究はまだ初期

---

[9] Meta は最近、新しい AI モデル「V-JEPA2」を発表した。これは AI エージェントが周囲の世界を理解することを支援する「ワールドモデル」として設計されている。（リンク）

[10] MuZero のようなシステムは、汎用 AI への進展を示す強力な強化学習（RL）アルゴリズムである。真の汎用人工知能（AGI）ではないものの、環境モデルを学習し、多様な状況で計画を立てられる能力は、より適応性と柔軟性を備えた AI システムへの道筋を示している。（リンク）

[11] Google DeepMind は、人気ゲーム『Minecraft』をプレイするよう Dreamer をプログラムした。Dreamer は画像からワールドモデルを学習する強化学習（RL）エージェントである。（リンク）

[12] インテルの Loihi に代表されるニューロモーフィック・チップは、脳が情報を処理する仕組みを模倣し、必要なときだけ信号を発火させる設計を採用している。このアプローチは、現行の GPU に比べてはるかに高い効率性と低いエネルギー消費を実現する可能性を秘めている。ただし技術はまだ初期段階にあり、大規模な AI 学習用ハードウェアを置き換える準備が整っているわけではない。（リンク）

段階にあるものの、量（巨大モデルや GPU）から質（賢いアーキテクチャと効率的な計算）への重点シフトを示唆している。

こうした転換は電力やインフラにも波及する。新しいアーキテクチャが効率的であれば、大規模な電力計画の必要性は和らぐかもしれない。逆に計算効率が低ければ、電力需要をさらに押し上げる可能性もある。電力事業者や政府にとって、この不確実性は長期的な計画を難しくする要因となっている。

資金の流れも同様に変化し得る。現在、投資家の間では「勝ち筋はスケーリング」という見方が支配的で、GPU、データセンター、そして最大規模のモデル訓練を行う研究機関に資金が集中している。しかしスケーリングが行き詰まれば、資本は新しいアーキテクチャを追求するスタートアップや研究機関へと移動するだろう。各国政府もまた、エネルギーを大量に消費するスケーリングへの補助金から、アルゴリズムや脳を模倣したハードウェアといった基盤研究支援に軸足を移す可能性がある。

AI はいま、明確な岐路に立たされている。短期的には LLM が引き続き価値を生み出し、巨額の計算資源投資を正当化すると思われる。しかしその限界こそが新しいアプローチへの研究を促し、長期的な進路を形づくる可能性を高めている。これらの道は必ずしも排他的ではなく、LLM が広範な「AGI 技術スタック」の重要な一部として残ることもあれば、完全に取って代わられることもあり得る。いずれに進むのか、その答えはまだ誰にも分からない。

5.「ストーリー」に左右される未来

総じて言えば、注目すべきなのは科学そのものだけではなく、そこに付随する「ストーリー」である。分野の先駆者たちの間で深い意見の相違が残っているにもかかわらず、市場はいま、スケーリングが未来を切り開くと確信している。本来なら投資家に慎重さを促すはずの状況だが、ストーリーは自己増殖的に広がっている。語り手の OpenAI や NVIDIA、さらにはワシントン自身もスケーリングに巨額の賭けをしている当事者であり、このビジョンを掲げることで政治的・金融的資本を一層 GPU や電力インフラ、データセンターに流し込み、利害が複雑に絡み合うエコシステムを形成している。

歴史は警鐘を鳴らしている。ドットコム・バブルもまた魅力的なストーリーの上に築かれ、大規模なインフラ投資、目を見張る企業評価額、そして「必然」であるかのような熱狂を生み出したが、最終的には崩壊を免れなかった。もしスケーリングが限界に突き当たるなら、現在の AI ブームも同様の清算を迫られるかもしれない。

それでも、スケーリングがスーパーインテリジェンスや AGI に直結しなかったとしても、LLM の短期的な進展は依然として大きな可能性を秘めている。これらのシステムはすでにビジネスモデルを作り替え、新しいツールを生み出し、数年前には想像できなかった生産性や創造性を実現しつつある。その価値が最終的に計算資源やエネルギー、インフラにかかる莫大なコストを正当化するかどうかは不透明だが、これまでの成果が揺るぎない事実であることは間違いない。

結局のところ、AI の物語は科学や工学の問題にとどまらない。政府や市場がどのストーリーを信じ、どの価値に対して代価を払うのか、そして社会がそのストーリーにいかに巨額の賭けをするのか。その選択が未来を形づくっていくことになりそうだ。

丸紅米国会社ワシントン事務所

1717 Pennsylvania Avenue, N.W. Suite 375, Washington, D.C. 20006
https://www.marubeni.com/jp/research/