AGI をめぐる認識の差と AI 研究開発の現在

公開モデルと研究現場の能力差をどう捉えるか

2026 年 9 月 17 日

GPT-6 Astra

要旨

AI がすでに汎用人工知能である AGI の段階に入ったという経営者の発言と、日常の利用で経験する誤答や不安定さとの間には隔たりがある。この隔たりを理解するには、AGI の定義に加え、誰が、どのモデルを、どのような条件で観察しているのかを問う必要がある。一般に利用できるフロンティア LLM は、研究現場で確認される最高能力と必ずしも一致しない。未公開モデル、長時間の推論、多数のエージェント、専門家による支援を組み合わせたシステムは、通常のチャット利用と異なる姿を示し得る。AI を次世代 AI の研究開発に組み込む動きも、この認識の差を広げる可能性がある。ただし、公開範囲の違いを一律の世代差に置き換えたり、経営者の発言を AGI 到達の証明とみなしたりすることはできない。本稿は、確認できる事実とそこから導く考察を区別し、現在の実用性と今後の能力拡張を別々に評価することが、企業の意思決定と社会の理解に必要であると論じる。

1　AGI という言葉と日常の利用経験

AI の進歩をめぐる議論では、同じ技術を見ているはずの人々が、まったく異なる現実を語ることがある。ある人は、AI が研究や専門業務の主体になり始めたと受け止める。別の人は、手元のサービスが資料の細部を読み違え、存在しない出典を示し、指示を繰り返しても修正しきれない経験を重く見る。いずれも、観察した範囲では理由のある評価であり得る。問うべきは、どちらの印象が強いかではなく、それぞれの判断が、どの対象と条件に基づいているかである。

NVIDIA のジェンスン・フアン CEO は、Lex Fridman のインタビューで、AGI を達成したと思うと述べた。ただし、その応答は、AI が企業を立ち上げて収益化できるかという文脈に置かれており、持続的に NVIDIA のような企業を築く能力まで認めたわけではない［1］。OpenAI のグレッグ・ブロックマン社長も、2026 年 9 月 3 日の会見で「AGI の時代へようこそ」と発言している［2］。

これらの発言は、AI 企業の経営者が到達点をどう評価しているかを示す。発言そのものは確認できるが、発言した人物の地位によって、技術的な判定が自動的に確定するわけではない。他方、こうした言葉を宣伝として処理してしまえば、開発現場で進んでいる変化を見逃すおそれがある。評価すべき対象は、発言者の強気や慎重さだけでなく、その背景にある作業能力、研究工程、検証可能な結果である。

一般利用者が接するのは、製品として提供される AI である。研究者が観察する対象には、それに加えて、実験中のモデル、特殊な実行構成、失敗した試行、訓練による改善の履歴が含まれ得る。製品を数回試すことと、研究工程に継続して組み込むことでは、得られる経験が異なる。この違いは、将来予測にも影響する。現在できない仕事を重視する人と、どの速さでできる仕事が増えているかを重視する人では、同じ失敗の意味づけさえ変わる。

もっとも、一般利用者と開発者を、それぞれ見解のそろった集団として描くべきではない。外部にも公開モデルを高度に使いこなす研究者や実務家がいるし、開発企業の内部にも能力や進歩の速度に関する見方の違いはあり得る。本稿が検討するのは、所属による優劣ではなく、情報と経験へのアクセスが認識を変える仕組みである。AGI をめぐる対立を理解するには、まずこの観察条件の違いを言葉にする必要がある。

2　AGI の判定には評価対象の明示が要る

AGI という言葉には、異なる期待が重ねられている。OpenAI の憲章は、経済的に価値のある仕事の大部分で人間を上回る、高度に自律的なシステムとして AGI を説明している［3］。この説明には、知識の広さだけでなく、仕事を遂行する能力と自律性が含まれる。難しい試験で高得点を取ることは、その一部を示しても、あらゆる実務工程を安定して担えることまで直接証明するものではない。

Google DeepMind の研究者らによる「Levels of AGI」は、能力の深さである性能と、対応できる課題の広さである汎用性を区別し、自律性や利用形態との関係も論じている［4］。この整理が示唆するのは、AGI への到達を単一の境界線で語る際には、何を比較しているかを先に定めなければならないということである。多分野の問題に答えられること、専門家を上回ること、人間の細かな指示なしに仕事を終えることは、同じ命題ではない。

例えば、あるシステムが高度な分析案を作成できても、その前提資料を選ぶのに専門家の助けが必要かもしれない。別のシステムは、資料の収集から報告まで自動で進められても、重要な例外を見逃すかもしれない。前者を高い知的能力の証拠と評価し、後者を不十分な自律性の証拠と評価することは両立する。議論がすれ違うのは、知的能力、作業の完遂、責任を負える信頼性を、同じ言葉で扱うときである。

評価条件には、時間と費用も含めるべきである。多大な計算資源を投入して難問を解けるなら、技術としての到達点は高い。しかし、それを日常業務のすべてに適用できるかは別の問題である。研究上の能力評価では、従来不可能だった成果に価値がある。企業の導入判断では、成果の質に加え、検証費用、待ち時間、業務量への対応が問われる。基準の違いを明示すれば、双方の評価を矛盾なく理解できる場合がある。

ここから、AGI についての意見が分かれていることを、そのまま能力認識の差の証拠とみなすことには限界があると分かる。同じ性能を見ていても、AGI の基準が違えば判定は変わる。他方、同じ基準を採用していても、観察できるシステムが違えば判定が変わる。

この二つを分けることで初めて、定義をそろえれば解消する対立と、追加の情報や実験が必要な対立を識別できる。

AGI という呼称の意義は、未知の変化への注意を促す点にある。しかし、導入や組織改革の判断をその呼称の確定まで待つ必要はない。仕事のどの部分が任せられるようになり、何が依然として人間に依存し、どこで費用対効果が変わったかを追えば、名称に関する合意がなくても意思決定は進められる。AGI の議論を具体的な能力の議論へ接続することが、実務にとっての出発点となる。

3　一般公開モデルと研究現場は同じ対象ではない

一般に利用できるフロンティア LLM とは、ある時点で高い性能を持ち、一定の利用条件の下で外部からアクセスできるモデルを指す。ここでいう一般利用には、有料サービスや API を通じた利用も含まれる。無料の初期設定だけを一般利用者の経験とみなすと、公開されている能力を過小に描くことになる。他方、外部から利用可能であることは、開発企業が持つ最高能力のすべてが、そのまま提供されていることを意味しない。

OpenAI は 2026 年 9 月 8 日の研究発表で、GPT-6 Astra より大幅に高性能な内部モデルを使用したと説明した。そのモデルは 8 月 28 日から訓練され、発表時点でも訓練が継続していたとされる［5］。この説明は、公開モデルより先行する内部モデルの存在を具体的に支持する。ただし、比較の範囲や条件をすべて外部から再現できるわけではなく、性能差の大きさは、開発企業による報告として受け止める必要がある。

研究現場では、試作品を限定的に使い、何が改善し、何が悪化したかを調べることができる。広く提供する製品では、より多様な利用者、入力、業務環境に対応しなければならない。研究で有用なモデルが、直ちに大規模な製品提供に適するとは限らない。この違いは、未公開モデルの能力を否定するものでも、一般公開モデルの価値を下げるものでもない。実験と普及では、求められる条件が異なるということである。

また、内部モデルと開発中モデルを別々の段階として数えると、同じ対象を二重に数えることがある。前述の内部モデルは、訓練を継続しながら研究に使われていた。つまり、開発と利用は明確に分離されていない。研究に役立つ状態になったモデルを使い、その成果を次の改善に戻す運用では、完成してから利用するという製品中心の見方だけでは実態を捉えにくい。

それでも、開発中であること自体は、性能上の優位を意味しない。新しい手法の探索には、得意分野が限られる試作や、特定の評価だけが改善するケースもあり得る。重要なのは、どの課題について、どの条件で、どの程度の改善が確認されたかである。非公開であるという情報に、未確認の万能性を付け加えるべきではない。外部の立場からは、存在の報告、性能の報告、再現可能な検証を段階的に評価する必要がある。

この意味で、公開モデルと研究現場の差を一律に二世代、三世代と表すことは適切でない。公開範囲はアクセスの区分であり、世代は開発上または性能上の区分である。両者が一致する事例はあっても、一般則にはならない。むしろ、比較する時点、課題の種類、利用できる道具、投入資源を明示する方が、能力の差を具体的に理解できる。

4　限定提供には異なる理由がある

高度な能力へのアクセスが限定されるとき、その理由を一つに決めつけることはできない。研究段階の検証、運用費用、応答時間、安全上の条件などによって、利用対象が分かれることがある。外から見れば同じ限定提供でも、より高性能な別モデルへのアクセスなのか、同じモデルの利用条件が異なるのかによって、意味は大きく変わる。

Google は 2025 年 8 月、Gemini 2.5 Deep Think について、一般の有料利用者向けの提供と、国際数学オリンピックで金メダル水準に達した版の研究者向け提供を区別した。後者は複雑な数学問題の推論に数時間を要し、一般提供版は日常利用に適する速さを重視したと説明されている［6］。これは、研究で示された到達点と、利用しやすい製品の条件が同一ではない例である。

これに対し、Anthropic は 2026 年 9 月の発表で、Claude Fable 5.1 と Claude Mythos 5.1 が同じ基盤モデルであり、安全措置の水準が異なると説明している。前者は一般提供、後者は認証された利用者向けの提供とされる［7］。この場合、名前とアクセス経路が違っても、それを一世代分の性能差として数えることはできない。利用できる能力の範囲と、モデルの基礎的な能力を分けて考える必要がある。

こうした違いは、利用者の評価に直接影響する。ある仕事への応答が制限されたとき、それがモデルの能力不足によるものか、利用条件によるものか、外部からは判別しにくい場合がある。逆に、限定提供の制度があることだけを根拠に、その対象モデルがあらゆる分野で優れていると推測することもできない。制度の説明と、課題ごとの評価結果を併せて見ることが求められる。

公開の段階を固定的な序列と考えると、企業による異なる提供戦略を同じ尺度に押し込めることになる。研究用に計算量を増やす企業もあれば、安全措置を用途別に調整する企業もある。同じ企業でも時期によって方式は変わり得る。したがって、限定提供、社内利用、開発中という表示を見ただけで、性能の順位や将来の製品名を推定することには慎重であるべきだ。

ここで重要なのは、一般利用者が必ず古い技術を使わされているという理解を避けることである。一般公開モデルにも最先端の成果が反映される一方、研究現場には製品化されていない能力や構成が残る。この両方を認める方が、公開モデルの実用性を正当に評価しながら、次に何が起こり得るかを考えやすい。公開と非公開は、先進と後進を単純に区切る線ではない。

5　モデルの性能とシステム全体の能力

AI の能力を考える際には、学習済みのモデルと、それを使って仕事をするシステム全体を区別する必要がある。モデルは中核となる判断や生成を担うが、成果は、与える情報、利用できる検索や計算の道具、試行回数、結果の検証方法にも左右される。単独の応答を評価しているのか、複数の工程を組み合わせた作業を評価しているのかが曖昧なままでは、モデル名を並べても公平な比較にはならない。

OpenAI は 2024 年の o1 に関する発表で、訓練に使う計算量に加え、回答時の推論に使う計算量を増やすことでも性能が改善すると報告した［8］。この事実は、モデルの能力が、短い応答を一度得る条件だけでは捉えきれないことを示す。同時に、計算量を増やせばあらゆる課題で同じように改善する、あるいは費用に見合う成果が必ず得られると一般化することもできない。

通常のチャット利用では、利用者が質問を送り、返ってきた回答を読む。エージェントとして動かす場合には、AI が作業を分け、必要な資料を読み、コードを実行し、その結果に応じて次の行動を選ぶ構成を取れる。途中で誤りを検出できる環境なら、最初の案が不十分でも修正を重ねられる。この違いは、応答の文章力だけを見ていると見落としやすい。仕事の成果は、最初の出力だけで決まるとは限らないからである。

前述の OpenAI の数学研究では、約 1 万のエージェントが並行して動く構成が用いられ、研究の途中で、さらに訓練が進んだモデルへの更新も行われたとされる［5］。この報告を通常のチャットの一回答と比較するなら、モデルの違いに加えて、探索規模と作業構成の違いを考慮しなければならない。成果が大きいからモデル単体も同じ比率で優れているという推論は成立しない。

大量の試行は、それ自体で信頼性を保証しない。同じ前提の誤りを多数のエージェントが共有すれば、出力の数が増えても誤りは修正されないことがある。有効な並行探索には、異なる案を生み出す仕組みと、案を選別する根拠が必要になる。専門家が途中で方向を変えたり、実行結果を使って候補を落としたりする場合、その寄与を含めてシステムとして評価するのが適切である。

実務的な比較では、同じ課題を、同じ時間、同じ費用、同じ資料条件で処理した結果と、資源制約を緩めて最高性能を目指した結果を分けて示すべきである。前者は導入判断に役立ち、後者は技術の到達点を理解する材料となる。両方の測定に意味があるが、一方の結果をもう一方の説明に流用すると誤解が生じる。能力を語るときに、計算資源と人間の支援を併記する必要があるのはこのためである。

この視点に立つと、認識の差は、未公開モデルの有無だけで説明できないと分かる。公開モデルであっても、業務に必要な資料と実行環境を与え、長時間の作業に使う人と、初期設定で短い質問をする人では、見える能力が違う。より高性能なモデルが外部に出れば

差がすべて解消するとは限らない。モデルへのアクセスに加え、能力を引き出す環境を整える知識と投資が必要になる。

6　AI が研究開発に組み込まれる意味

AI を使って AI を開発するという表現は、しばしば、AI が独力で自らを設計し直す姿を連想させる。しかし、現実の研究開発には、課題の選択、実験の実装、訓練、評価、結果の解釈、改善の採否など、多様な仕事がある。その一部を AI が担うことと、全体を自律的に運営することは区別しなければならない。研究支援が相当に進んでいても、最終判断の多くが人間に残る状態は十分に考えられる。

OpenAI が 2026 年 9 月 6 日に公表した社内状況の報告では、8 月中旬の研究者の利用額中央値は API 価格換算で 1 日 600 ドル超とされ、実験の実施量も増加している。他方、同社は、利用量やコード量の増加を研究全体の加速率と同一視できないことや、人間が研究の優先順位と採否を判断していることも説明している［9］。ここから読み取るべきなのは、研究の進め方が変わりつつあるという事実と、その効果を測る難しさの両方である。

この金額は、一般向け API 価格で使用量を換算した指標であり、研究者一人当たりの実際の支払額や、内部の限界費用を意味しない。とはいえ、日常的な対話サービスの利用と比べて、研究現場では大量の推論を仕事に組み込む場合があることを示す。利用経験を比較するなら、どのモデルかという問いとともに、どれだけ継続的に仕事を任せているかを確認する必要がある。

Anthropic は 2026 年 9 月 17 日の報告で、8 月時点の社内評価として、AI 研究開発業務の 26%を Claude が人間の監督下で主導し、90%超では少なくとも協働していると説明した。一方、測定された業務区分のいずれでも、完全自律には達していないとしている［10］。この比率は同社の分類・重み付けに基づく指標であり、研究者の 26%が不要になった、研究費の 90%を削減した、という意味には読み替えられない。

AI が担う仕事の範囲が広がれば、研究者の時間配分が変わる可能性がある。実装や不具合の調査に費やしていた時間を、仮説の比較や実験設計に振り向けられるなら、少人数でも検討できる案は増える。ただし、試す案が増えれば、その結果を解釈する仕事も増える。個々の工程が速くなったとき、どこに新たな待ち時間や確認負担が生じるかを見なければ、研究全体への効果は分からない。

また、実験の増加と AI の利用増加が同時に観察されても、因果関係を一つに決めることはできない。計算設備の増強、研究者の増員、課題の難易度、実験の数え方によっても結果は変わる。研究開発を支える経営判断では、投入資源、工程別の時間、採用された改善の量を併せて見る必要がある。活動量が増えたという指標を、実質的な技術進歩の指標へ結び付ける作業が重要になる。

それでも、AI が研究現場で日々使われ、その得意な作業の範囲が広がっているなら、開発者の将来認識が変わるのは自然である。従来は人間の作業時間で制約されていた工程を計算資源に置き換えられると感じれば、今後の改善速度の見通しも変わる。一般利用者が一つの製品の便利さを評価している間に、研究者は、その製品を生み出す工程自体の変化を見ている。この違いは、進歩の速度をめぐる認識の差を説明する有力な要因となる。

7　研究の加速と自律的な自己改善の距離

研究開発への AI 活用が進めば、優れた AI が開発を助け、その結果としてさらに優れた AI が得られる循環が生じ得る。ここで、循環が存在することと、その循環が無制限に速まることを分ける必要がある。ある工程で効率が改善しても、次の工程で計算資源、実験設備、評価、人間の判断が制約となれば、全体の速度はそこで決まる。局所的な改善率を、そのまま将来の進歩率へ延長することはできない。

Google DeepMind が 2025 年 5 月に発表した AlphaEvolve は、この問題を考える具体例になる。Gemini によるプログラムの提案と自動評価を組み合わせ、AI 訓練に関わる計算処理などの最適化に用いたとされる。同社は、特定の計算処理を 23%高速化し、Gemini の訓練時間を 1%短縮したと報告している［11］。部分の改善と全体への寄与が異なることを、同じ事例の中で読み取れる。

この例で注目すべきは、AI が提案を作るだけでなく、提案を評価し、次の探索に戻す工程に組み込まれていることである。答えの良否を比較的明確に判定できる課題では、このような反復を設計しやすい。一方、将来性のある研究テーマの選択、成果の社会的意味、長期的な安全性などは、短時間の自動評価だけで確定しにくい。自動化しやすい仕事と、判断を残すべき仕事の境界が、加速の程度を左右する。

自律的な自己改善を論じる際には、AI が改良案を出す、実験を実行する、結果を評価する、採用を決める、次の研究方針を更新する、という各段階を確認する必要がある。一部を人間が担っているなら、その関与を取り除いた場合にも同じ成果が得られるかは未確認である。人間の関与が残ることは研究成果の価値を否定しないが、完全自律という表現を使うときには決定的な違いになる。

計算資源にも制約がある。改善案を考える速度が上がっても、すべての案を十分な規模で訓練して比較できるとは限らない。高価な実験を選別する能力が重要になり、評価の誤りが従来以上の損失につながる可能性もある。研究の自動化は、人間の判断を不要にするというより、どの判断に人間の注意を集めるべきかを変える面がある。

さらに、安全性を確認する工程を、能力開発の後に付け加えるだけでは不十分になる場合がある。研究用のエージェントに資料や実行環境へのアクセスを与えるほど、作業範囲と権限の管理が重要になる。OpenAI の GPT-6 Astra のシステムカードは、社内利用にもア

クセス制御、監視、利用前の評価を設けていると説明する［12］。内部で使うのであれば制約が不要になる、という理解は適切ではない。

将来の急速な自己改善は検討すべき可能性であるが、現在の研究支援から、その到来時期や速度を一義的に導くことはできない。実務的には、仮説の生成、実験の実施、検証、採否の判断という工程ごとに、人間の関与がどう変わっているかを追う方が有益である。そうすれば、過大な期待を抑えるだけでなく、従来の想定を超える変化を早く捉えることもできる。

8　開発者が持つ情報上の優位とその限界

最先端の開発者が外部の利用者より多くの情報を持つことは、将来予測にとって重要である。公開前の評価結果を知り、改善が偶然なのか、複数の条件で繰り返し確認されるのかを判断できれば、新しい能力への確信は強くなる。失敗の原因が分かり、解決の見通しが立っている場合、外からは重大に見える欠点を、開発者は一時的な問題として受け止めることもある。

OpenAI の主任科学者 Jakub Pachocki は、2026 年 9 月の論考で、2023 年半ばの内部研究結果が、推論モデルの訓練を拡大できるという確信につながったと回顧している［13］。これは、製品の公開前に、研究現場が技術の方向性に関する情報を持ち得ることを示す。外部の利用者が完成品の出力から進歩を推し量るのに対し、開発者は改善の過程そのものを観察できる。

ただし、情報が多いことと、将来を正しく予測できることは同義ではない。研究でうまくいった手法が、より大きな規模でも同じ効果を持つとは限らない。社内の課題に適したシステムが、広い業務分野で同程度に機能するとも限らない。観察範囲が深いほど、自分たちの課題で得られた改善を一般化したくなる可能性がある。専門性は予測の根拠を増やすが、適用範囲の確認を不要にはしない。

また、経営者、研究者、製品担当者は、同じ企業の中でも見ている指標が異なる。研究者は新しい能力の発現に注目し、製品担当者は利用者が失敗する条件を把握し、経営者は競争と投資の時間軸を考える。これらの役割を一括して最先端開発者と呼ぶと、発言の性質を見誤る。経営者の到達宣言と、研究者が示す限定条件付きの評価結果は、それぞれの文脈で読む必要がある。

企業の発信には、技術の説明と、事業上の期待を示す役割が同時にある。だからといって、強い発言をすべて宣伝とみなすことはできない。しかし、発言の強さを証拠の強さと同一視することも避けるべきである。判断の基準は、具体的な作業、実行条件、失敗率、第三者が確認できる情報がどれだけ示されているかに置くべきだ。

内部の情報優位を社会の理解につなげるには、秘密のモデルの存在を強調するだけでは足りない。どの業務がどの条件で改善したのか、どこに人間の介入が残るのか、公開モデ

ルで再現できる範囲はどこまでかを示す必要がある。その情報があれば、外部の人も前提を点検しながら見通しを更新できる。開発者の知見を尊重することと、独立した検証を求めることは両立する。

9　利用者の経験も過小評価と過大評価の両方を生む

日常の利用経験は、AI を評価するうえで欠かせない。研究発表では見えにくい入力の曖昧さ、資料の欠落、業務上の例外は、実際の仕事で初めて表面化することがある。利用者が誤りを繰り返し経験し、導入に慎重になるのには理由がある。その経験を、最先端を知らないことによる遅れとして片付けてはならない。製品が仕事の中で機能するかを判断するための、重要な観測である。

一方、過去の失敗を、現在と将来の AI 全体の限界として固定することにも問題がある。試した時期、モデル、設定、課題の与え方が変われば、結果が変わる可能性がある。特に、単純な対話で失敗した課題が、資料検索や実行環境を備えた構成なら処理できる場合、古い試用結果だけでは現在の選択肢を評価できない。否定的な経験にも、評価日と利用条件を付けて保存する必要がある。

好意的な利用者の体感も、そのまま生産性の改善率とはならない。METR は 2025 年 7 月、経験豊富なオープンソース開発者 16 人が 246 件の実務課題に取り組む無作為化比較試験を報告した。当時の AI ツールを利用できる条件では、完了時間が平均 19%長くなり、参加者の効率化の認識と実測が一致しなかった［14］。これは特定の時期と条件の結果であり、すべての AI 利用を否定する証拠ではない。

この研究には、その後の更新もある。METR は 2026 年 2 月、後続実験では、AI を使わずに働く条件を避ける参加者や課題の選別が生じ、現在の効果を信頼して推定しにくくなったと報告した。以前より効果が高まった可能性を認めつつ、その大きさについての証拠は弱いとしている［15］。初期の否定的な結果だけを使うことも、後続の好意的な手応えだけを使うことも、研究の読み方として十分ではない。

ここから得られる教訓は、AI についての認識の差が、常に過小評価する一般利用者と、正しく理解する開発者の対立になるわけではないということである。人は、容易になった部分を強く記憶し、確認や手直しにかかった時間を見落とすことがある。逆に、目立つ一度の誤りが、多数の有用な出力への評価を下げることもある。印象と測定を併用することが、双方の偏りを減らす。

未公開の高性能モデルを理由に、公開モデルの実際の欠点を退けることも避けるべきだ。外部から確認できないモデルについて、どれほど失敗例を示しても内部では解決しているはずだと答えるなら、その主張は検証できなくなる。現在の製品に対する批評は、現在の製品の条件で評価されるべきである。将来の可能性を論じるときにも、何が確認済みで、何が予測なのかを維持する必要がある。

利用者の経験を価値ある知見にするには、成功も失敗も再現できる形で残すことが有効である。課題、入力資料、使用モデル、実行条件、修正の回数、最終成果の採否を記録すれば、モデル更新時に同じ仕事で比較できる。そこで得られるのは、AI 一般への賛否ではなく、自社の業務のどこが変わり始めたかという情報である。この蓄積は、外部の発表を自社の現実に引き寄せて読む基盤となる。

10　認識の差を検証可能な問いに変える

一般利用者と開発者の認識の差を説明する仮説が有用であるためには、検証の方法を持たなければならない。内部モデルが優れている、計算資源が多い、専門家が使い方を知っている、という説明は、それぞれ異なる予測を生む。これらを区別せず、すべてを最先端だからという言葉でまとめると、何を改善すれば利用者の経験が変わるのかが分からなくなる。

モデル自体の差を調べるなら、課題と実行環境をそろえ、モデルだけを替えて比較する必要がある。推論量の差を調べるなら、同じモデルに異なる計算予算を与える。道具の効果を調べるなら、検索やコード実行の有無を変える。人間の支援の寄与を調べるなら、介入の時点と内容を記録する。このような比較を組み合わせれば、成果を生んだ要因に近づける。

評価する課題にも注意が要る。正解が短く定まる問題は測定しやすいが、企業の仕事には、複数の関係者が納得できる説明、更新される資料への追随、例外への対応などが含まれる。最終出力だけを採点すると、途中で誤った情報を採用し、偶然それらしい結論に至ったケースを見逃す可能性がある。成果物の品質とともに、根拠、修正履歴、業務上の再利用可能性を確認する必要がある。

また、一度成功したことと、任せ続けられることは異なる。研究上の実証では、従来できなかった課題を一度解いたことに意味がある。運用では、条件が少し変わっても一定の品質を保ち、失敗したときに気付き、必要なら人に戻せることが重要になる。最高到達点を追う評価と、日常の安定性を追う評価を並べることで、技術の進歩と導入の難しさを同時に捉えられる。

人間の関与は、無いか有るかだけでは測りきれない。開始時に目的を示すこと、途中で資料を追加すること、誤りを指摘すること、完成した成果を承認することでは、負担が異なる。AI が長時間動いても、専門家が頻繁に方向を修正する必要があれば、任せられる仕事の量には限界が残る。介入に要する時間と専門性を評価に含める方が、実務での能力を理解しやすい。

認識の差を測るには、技術の比較だけでなく、人の予測と結果を照らす方法も考えられる。同じ課題について成功率や完了時間を事前に予想してもらい、実験後の結果と比較する。そのうえで、詳しい利用経験を積んだ後に予測がどう変わるかを見る。こうした設計

であれば、情報不足、定義の違い、体感の偏りのうち、何が判断を動かしているかを検討できる。これは本稿の提案であり、今回参照した報告から因果関係が確定したという意味ではない。

開発企業の内部では同じ方法を全面的に公開できない事情もある。それでも、比較条件、対象業務、結果の分布、人間の介入の程度を示す余地はあるはずだ。第三者による検証や、モデルを外部に配布しない形での評価も検討に値する。社会が求めるべきなのは、すべての内部情報への無制限のアクセスではなく、重要な能力の主張を点検できるだけの証拠である。

11　企業は現在の導入と将来への準備を分ける

企業が AI の能力を評価する目的は、技術の格付けだけではない。どの仕事に投資し、誰を育て、どの情報を整備するかを決めることである。その際、現在使えるシステムの評価と、今後の能力拡張への準備を混同すると、判断を誤りやすい。現在の欠点を理由に準備まで止めれば変化への対応が遅れ、将来の可能性を理由に現在の検証を省けば運用品質が下がる。

現在の導入は、利用できるモデルと環境によって判断するべきである。重要な成果が安定して得られるか、確認と修正を含めて仕事が速くなるか、社内の情報を扱えるかを測る。未公開モデルの報告は、現行製品の費用対効果を保証しない。採用するシステムに必要な資料と作業範囲を与えたうえで、自社の代表的な課題を使って評価する必要がある。

中期的な準備では、今すぐ完成した成果が得られなくても、能力が伸びたときに使える資産を整えることに意味がある。機械が読める資料、更新履歴、品質の高い評価課題、業務ごとの判断基準は、モデルが変わっても利用できる。特定の製品名に依存する操作手順だけを覚えるより、目的と評価方法を明確にしておく方が、新しい能力を取り込みやすい。

投資判断には、段階ごとの条件を置くことが有効である。例えば、候補案の作成は試行を認め、外部提出物への使用は根拠確認を条件とし、他のシステムを動かす仕事は権限と停止方法を整えてから広げる。ここで大切なのは、承認手続きを増やすこと自体ではない。AI に任せる範囲と、失敗したときの影響を対応させることで、成果が出た仕事を適切に拡張できるようにすることである。

人材育成でも、知識を持つことに加え、仕事を AI に説明し、成果を評価する能力が重要になる。専門家は、自分が無意識に行っている前提確認や例外判断を言語化する必要がある。AI を使う側は、答えを得る速さだけでなく、何を確かめればその答えを採用できるかを学ぶ。これは専門知識を不要にする方向ではなく、専門知識を作業設計と検証に使う方向への変化である。

経営層への報告も、AGI という言葉への賛否だけで終わらせるべきではない。自社のどの工程で処理可能な仕事が増えたか、どこに確認負担が残るか、次の改善に必要なデータ

や権限は何かを示す。技術の進歩を業務上の条件に翻訳できれば、過大な期待と過度な先送りの両方を抑えられる。外部の到達宣言を、自社の行動へ結び付けるのは、この翻訳の仕事である。

12　社会の理解を支える情報公開と学習

AI の能力に関する認識の差は、企業の導入判断だけでなく、社会が技術の方向を議論する条件にも関わる。開発企業だけが最先端の経験を持ち、外部からはその意味を十分に検討できないのであれば、将来像を語る力が一部に集中する。専門的な知見を社会の判断に生かすには、開発者への信頼だけに依存せず、主張を評価できる情報を共有する必要がある。

情報公開の中心に置くべきなのは、印象的な成功例だけではない。同じ課題で何回試したか、どの程度の費用と時間を使ったか、人間がどこで介入したか、失敗はどのような形で現れたかが分かれば、成果の意味を判断しやすくなる。モデルそのものを公開できない場合でも、実行条件と評価方法を説明する余地はある。企業秘密の保護と、能力に関する説明の具体化は、常に両立しないわけではない。

外部の評価には、研究現場の最高到達点と、社会で再現できる能力をつなぐ役割がある。共通の課題で性能を測るだけでなく、長い作業の完遂、例外への対応、検証の負担まで調べることが望ましい。同じモデルについて、計算資源を増やした条件と、一般利用に近い条件を併記すれば、公開されている技術の潜在力と実用上の制約を同時に理解できる。

報道や解説にも、到達宣言と評価結果の関係を明らかにする役割がある。経営者が AGI という言葉を使った事実は、そのまま伝えてよい。しかし、その発言がどの定義に基づくか、能力の主張は誰が確認したか、一般に利用できるかを併せて示すことで、読者の理解は変わる。発表時点と検証の進み具合を明記すれば、速報としての価値を保ちながら、暫定的な評価が確定事項として広がることを抑えられる。

学習の場では、AI の強さを見せる実演と、限界を見抜く経験を結び付けることが有効である。同じ課題を、短い質問だけで試す場合と、資料や道具を与えて進める場合で比較する。さらに、出力を採用するまでにどの確認が必要だったかを振り返る。こうした経験は、AI を過大評価するためでも、失敗を探して否定するためでもなく、能力が発揮される条件を理解するために行うものである。

利用環境の差を個人の理解力の差に還元しないことも大切だ。高性能なサービスの料金、使える時間、専門資料、実行環境、学習の機会は、誰にでも同じように与えられているわけではない。高度な活用を知る人が、短い試用だけで判断する人を責めても、経験の隔たりは縮まらない。認識を更新できる場を広げることは、モデルの提供範囲を広げることとは別の課題として考える必要がある。

また、情報の差が縮まっても、社会の意見が一つになるとは限らない。能力向上を歓迎しながら、仕事の変化や責任の所在を重く見る人もいる。開発者と同じ性能を認めても、普及の速さや導入の条件について異なる判断をすることは可能である。技術に詳しくなれば必ず同じ結論に至るという前提を置かず、事実についての認識と、何を望ましいとするかの判断を分けて議論するべきだ。

目指すべきなのは、社会全体が開発企業と同じ期待を持つことではない。どの能力が確認され、どの評価が特定の条件に依存し、どの見通しがまだ予測なのかを、外部からも検討できる状態である。そのためには、成功と失敗を記録し、比較可能な結果を積み重ね、評価が変わった理由を説明することが要る。情報の差を完全になくせなくても、判断の根拠を共有することはできる。

13　情報の差を社会的な判断につなげる

公開モデルと研究現場の能力が一致しないことは、社会が AI を理解するうえで避けて通れない問題である。最も詳しい情報を持つ企業が将来像を語り、外部の人々は限られた製品と公開資料から判断する。この構造の下では、企業への全面的な信頼にも、全面的な不信にも、十分な根拠を持たせにくい。必要なのは、主張を点検できる情報を増やし、評価を更新する仕組みである。

開発企業には、能力の高さだけでなく、結果が得られた条件と限界を説明する責任がある。一般利用者や導入企業には、自らの利用経験の範囲を認識し、過去の印象を更新する姿勢が求められる。研究者や評価機関には、内部の報告と外部の実務をつなぐ測定が期待される。これらの役割は、AGI という呼称に合意してから始めるものではない。

本稿の検討から支持されるのは、一般に利用できるフロンティア LLM が、研究現場の最高能力のすべてを代表しているわけではないという命題である。より高性能とされる未公開モデルの報告があり、大規模な計算資源を使う運用があり、AI を AI 研究開発に組み込む動きがある。これらは、通常の製品利用とは異なる経験を生む条件として理解できる。

そこから、情報と経験の違いが、能力や進歩の速度をめぐる認識の差の一因になり得ると考えることには合理性がある。ただし、AGI 到達をめぐるすべての意見の違いを説明するわけではない。定義、評価基準、実用性への要求、将来の見通しも関係する。認識の差を一つの原因へ還元せず、どの条件がどの判断を動かすのかを確かめる姿勢が必要である。

企業にとっての課題は、現在使える AI の価値を測りながら、研究現場で生じている変化を次の準備につなげることである。現在の失敗を将来の限界と決めつけず、将来の可能性によって現在の欠点を覆い隠さない。必要な資料を整え、代表的な業務を繰り返し評価し、人間が担う判断を明確にする。その積み重ねが、研究現場の進歩を社会が理解し、活用の条件を自ら選び取るための基盤となる。

参考文献

ウェブ資料の最終確認日　2026 年 9 月 17 日

［1］Lex Fridman. Jensen Huang: NVIDIA - The $4 Trillion Company & the AI Revolution. Lex Fridman Podcast #494。2026 年 3 月 23 日。参照箇所は文字起こしの AGI timeline および Future of programming。
https://lexfridman.com/jensen-huang-transcript/

［2］Ina Fried. "Welcome to the AGI era," OpenAI says as GPT-6 Astra debuts. Axios。2026 年 9 月 3 日。
https://www.axios.com/2026/09/03/openai-astra-gpt-6-agi-brockman

［3］OpenAI. OpenAI Charter. 企業憲章。AGI の定義を参照。
https://openai.com/charter/

［4］Meredith Ringel Morris ほか. Levels of AGI for Operationalizing Progress on the Path to AGI. ICML 2024。arXiv:2311.02462。初稿 2023 年 11 月 4 日、参照版 v5 は 2025 年 9 月 24 日。
https://arxiv.org/abs/2311.02462

［5］OpenAI. On the Navier–Stokes Millennium Prize Problem. 2026 年 9 月 8 日。9 月 10 日更新。内部モデルの能力に関する説明と研究の実施方法を参照。
https://openai.com/index/navier-stokes-solution/

［6］The Deep Think team. Try Deep Think in the Gemini app. Google。2025 年 8 月 1 日。一般提供版と数学研究者向けの版の違いを参照。
https://blog.google/products-and-platforms/products/gemini/gemini-2-5-deep-think/

［7］Anthropic. Introducing Claude Fable 5.1 and Claude Mythos 5.1. 2026 年 9 月 1 日。基盤モデルの共通性と安全措置および提供条件の違いを参照。
https://www.anthropic.com/claude-fable-and-mythos-5-1

［8］OpenAI. Learning to reason with LLMs. 2024 年 9 月 12 日。訓練時と推論時の計算量に関する説明を参照。
https://openai.com/index/learning-to-reason-with-llms/

［9］OpenAI. Research acceleration: The view inside OpenAI. 2026 年 9 月 6 日。研究者のエージェント利用量、研究工程、指標の限界を参照。
https://openai.com/index/research-acceleration-view-inside-openai/

［10］Marina Favaro and Phillie Wright. Measurements for understanding the pace of AI development inside frontier labs. Anthropic。2026 年 9 月 17 日。AI 研究開発の自動化指標とその測定上の限界を参照。
https://www.anthropic.com/institute/measuring-pace-of-ai-development

［11］AlphaEvolve team. AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms. Google DeepMind。2025 年 5 月 14 日。AI 訓練に関わる計算処理の最適化を参照。
https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/

［12］OpenAI. GPT-6 Astra System Card. OpenAI Deployment Safety Hub。2026 年 9 月 3 日公表、9 月 9 日更新。社内利用に関する管理と評価を参照。
https://deploymentsafety.openai.com/gpt-6-astra

［13］Jakub Pachocki. An Alien Mind. OpenAI。2026 年 9 月 6 日。内部研究の経験に基づく将来認識についての論考。
https://openai.com/index/an-alien-mind/

［14］Joel Becker, Nate Rush, Beth Barnes and David Rein. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. METR。2025 年 7 月 10 日。経験豊富な開発者を対象とした無作為化比較試験。
https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

［15］Joel Becker, Nate Rush, Tom Cunningham, David Rein and Khalid Mahamud. We are Changing our Developer Productivity Experiment Design. METR。2026 年 2 月 24 日。後続実験の選択バイアスと測定上の課題を参照。
https://metr.org/blog/2026-02-24-uplift-update/