Educational Research Foundation

ERF Research Companion Series No.1

Google DeepMind

From AGI to ASI

第 6 回

再帰的自己改善と知能爆発

AI が AI 研究を加速するとき、
AGI から ASI への移行は急激になるのか

Publication Edition

Version 1.0

2026

目次

はじめに

Google DeepMind の『From AGI to ASI』は、AGI から ASI へ向かう第三の経路として、Recursive Self-Improvement——再帰的自己改善を挙げています。

再帰的自己改善という言葉からは、しばしば次のような場面が想像されます。

1. AI が自分自身のプログラムを読み取る
2. 自分より優れたコードへ書き換える
3. 改良された AI が、さらに自分を改良する
4. この循環が急速に反復される
5. 短期間で人間をはるかに超える超知能が生まれる

これは、再帰的自己改善の一つの可能性ではあります。

しかし、原報告書が扱う概念は、これより広いものです。

原報告書における再帰的自己改善とは、

> **AI が AI の研究開発を促進し、その結果として生まれた改良型 AI が、さらに AI 研究開発を加速する循環**

を意味します。

したがって、AI が自分の内部コードを直接書き換えなくても構いません。

AI が、

- より良いアルゴリズムを発見する
- 実験を設計・実行する
- 学習データを生成する
- 半導体を設計する
- データセンターの効率を高める

- AI 研究者同士の分業を改善する

ことによって、次世代 AI を生み出し、その次世代 AI が同じ研究をさらに効率化するなら、すでに再帰的な改善循環が成立します。（原報告書、Section 5.3）

この経路が特に重要なのは、AI の進歩速度そのものが、AI 能力によって変化する可能性があるからです。

通常の技術進歩では、研究者が新技術を開発します。

しかし AI が AI 研究者の役割を担うようになれば、

> **AI 能力の向上が、AI 能力向上の速度をさらに高める**

という正のフィードバックが生じます。

この循環が十分に強ければ、AGI から ASI への移行は、数十年かけてゆっくり進むのではなく、数年、数か月、あるいは理論上はさらに短い期間に圧縮される可能性があります。

一方で、原報告書は「知能爆発」を当然視していません。

再帰的改善は途中で停滞するかもしれません。

次世代 AI を改善するために必要な計算資源、実験、データ、半導体、電力が急増し、改善循環を維持できなくなる可能性もあります。

したがって、Section 5.3 の中心的な問いは、

> **再帰的自己改善は起こるか否か。**

だけではありません。

> **どの種類の改善循環が、どの程度の速度で進み、どこで限界に達するのか。**

という問題です。

第 1 節　「再帰的」とは何を意味するのか

再帰的という言葉は、ある過程の出力が、次の同じ過程の入力になることを意味します。

AI 研究に当てはめると、次の循環になります。

1. 現在の AI が AI 研究を支援する
2. その研究によって、より優れた AI が作られる
3. より優れた AI が、AI 研究をさらに効果的に支援する
4. その結果、さらに優れた AI が作られる
5. 循環が反復される

ここで重要なのは、単に AI が一度だけ改善されることではありません。

改善された AI が、次の改善能力そのものを高めることです。

たとえば、AI が人間の研究者を 20％効率化するだけなら、研究開発速度は上がります。

しかし、その結果生まれた次世代 AI が研究者を 50％効率化し、その次が研究作業の大部分を自律的に実行できるようになれば、研究速度の増加率自体が上昇します。

これが再帰的改善の核心です。

原報告書は、AI 研究開発の完全な自動化が可能になり、改善循環が広い能力領域にわたって持続する場合、AGI から ASI への「爆発的」移行が起こり得ると指摘しています。（原報告書、Section 5.3）

第２節　通常の技術進歩との違い

人類の技術発展にも、再帰的な側面があります。

新しい道具が、次の道具を作る能力を高めるからです。

たとえば、

- コンピュータが新しいコンピュータの設計を助ける
- 科学機器が新しい科学的発見を可能にする
- 教育制度が次世代の研究者を育てる
- 印刷技術が知識の共有を加速する
- インターネットが世界的な共同研究を促進する

という循環です。

しかし、人間社会の改善速度には大きな制約があります。

- 研究者の教育に長い時間がかかる
- 専門知識の共有が不完全
- 人間には睡眠や休息が必要
- 研究者の人数を急速に増やせない
- 組織が大きくなると調整コストが増える
- 新しい世代へ知識を伝えるのに年月がかかる

AI では、これらの制約の一部が弱まる可能性があります。

優れた AI 研究者を複製でき、学習成果を多数の個体へ共有でき、24 時間稼働させられるなら、研究者集団の拡大速度は人間社会と異なります。

したがって、再帰的自己改善は、技術進歩の完全に新しい原理ではありません。

人間社会にも存在する技術的・文化的フィードバックを、デジタル知能がはるかに速い速度で実行する可能性なのです。

第３節　自己改善は「自分のコードを書き換えること」だけではない

原報告書は、再帰的自己改善を四つの主要な形に分類しています。

1. アルゴリズムとソフトウェアの改善
2. ハードウェアの改善
3. データによる改善
4. 分業と専門化による改善

この分類は非常に重要です。

AI が自分自身のコード全体を理解し、直接変更できなければ再帰的自己改善は起こらない、という見方を避けられるからです。

実際には、四つの改善経路の一部だけでも、正のフィードバックが成立する可能性があります。（原報告書、Section 5.3）

第４節　第一の形――アルゴリズムとソフトウェアの改善

最も伝統的に想像されてきた再帰的自己改善です。

AI が、

- 新しいモデル構造を設計する
- より効率的な最適化手法を発見する
- 学習アルゴリズムを改善する
- 推論探索を効率化する
- AI エージェントの作業環境を改善する
- 評価方法や検証方法を設計する

ことで、次世代 AI を作ります。

この改善は、必ずしも AI が自分のソースコードを直接変更する形で起こる必要はありません。

AI が研究提案を作り、コードを書き、実験を行い、その結果を評価し、人間または自動システムが採用する形でも成立します。

4.1 AI 研究のどの部分を自動化できるか

AI 研究開発には、さまざまな工程があります。

- 先行研究の調査
- 仮説の生成
- 数学的分析
- アルゴリズム設計
- コードの実装
- 実験設定

- 実験の実行
- 結果の分析
- バグの発見
- 論文執筆
- 査読や再現性確認

現在の AI でも、コード作成、文献整理、実験分析、ハイパーパラメータ探索など、一部の作業を支援できます。

しかし再帰的改善を強くするには、AI が補助的な作業だけでなく、

- どの研究課題が重要かを判断する
- 新しい仮説を立てる
- 実験失敗の原因を分析する
- 予想外の結果から新しい理論を作る
- 長期的研究計画を修正する

必要があります。

つまり、単なる研究作業の自動化ではなく、研究の方向設定まで自動化できるかが重要です。

4.2 ニューラル・アーキテクチャ探索

再帰的改善の初期的な例として、原報告書は Neural Architecture Search——ニューラル・アーキテクチャ探索を挙げています。

これは、人間がニューラルネットワークの構造を一つずつ設計する代わりに、アルゴリズムが多数の候補構造を生成し、評価し、より良いものを選ぶ方法です。

たとえば、

- 層の数
- 接続方法

- 活性化関数
- 情報の流れ
- モジュール構成

などを自動的に探索します。

ただし、多数の候補を試すだけでは計算費用が膨大になります。

重要なのは、過去の実験結果から学び、有望な候補を優先し、探索自体を効率化することです。

探索方法が改善されれば、より良いアーキテクチャをより少ない実験で見つけられます。

その新しいアーキテクチャが、さらに効率的な探索を可能にするなら、再帰的循環が生じます。

4.3 ハイパーパラメータの自動最適化

AI モデルには、

- 学習率
- バッチサイズ
- モデル規模
- 正則化
- 学習回数
- データ混合比率

など、多数の設定値があります。

これらをハイパーパラメータと呼びます。

従来は研究者が経験に基づいて調整してきましたが、自動探索によって最適化できます。

一回の改善は小さく見えるかもしれません。

しかし、大規模な学習では、わずかな効率向上でも、

- 計算費用の削減
- 学習時間の短縮
- 同じ資源でより大きなモデルを訓練
- より多くの実験の実行

につながります。

その結果、研究サイクル全体が速くなります。

4.4 メタ最適化

より根本的なのがmeta-optimization——メタ最適化です。

通常の最適化では、モデルのパラメータを改善します。

メタ最適化では、

パラメータを改善する方法そのもの

を学習または探索します。

つまりAIが、

- より優れた更新規則
- より効率的な学習方法
- 新しい探索戦略
- より良い最適化アルゴリズム

を発見します。

これは「学ぶ方法を学ぶ」ことに近いものです。

原報告書は、更新規則やアーキテクチャそのものを発見するメタ最適化が、計算資源当たりの能力向上率を高める可能性を指摘しています。（原報告書、Section 5.3）

4.5 FunSearch と AlphaEvolve

原報告書は、より具体的な例として FunSearch と AlphaEvolve を挙げています。

これらは、大規模言語モデルを用いてプログラム候補を生成し、明確な評価器によって性能を測定し、優れた候補を反復的に改良するシステムです。

AlphaEvolve は、言語モデルの提案能力と自動評価器を組み合わせ、数学的構成や計算アルゴリズムの改善を探索します。Google DeepMind は、これを数学や計算分野における高度なアルゴリズム設計へ適用するシステムとして説明しています。（Google DeepMind、AlphaEvolve、2025）

この方式の基本循環は次のとおりです。

1. AI がプログラム候補を生成する
2. 自動評価器が性能を測る
3. 優れた候補を選択する
4. その候補をもとに新しい案を生成する
5. 循環を繰り返す

ここでは、人間が正解となるプログラムを直接教える必要はありません。

性能を客観的に測定できれば、AI は探索を通じて、人間が用意した学習例を超える解法を発見できます。

これは、限定的ではあるものの、アルゴリズム的自己改善の具体例です。（原報告書、Section 5.3）

第５節　第二の形――ハードウェアの改善

AI の能力はソフトウェアだけで決まりません。

計算を実行する半導体、データセンター、電力、通信網が必要です。

そのため AI がハードウェアを改善することも、再帰的自己改善に含まれます。

原報告書は、ハードウェア改善を広く捉えています。

- より高速なチップの設計
- より省電力なアクセラレーター
- より安価な計算装置
- 半導体配置の最適化
- 製造工程の改善
- サプライチェーンの効率化
- 資源調達の改善
- エネルギー生産の改善
- ロボット用ハードウェアの設計

などです。（原報告書、Section 5.3）

5.1 AI がチップを設計する

半導体設計は非常に複雑です。

数十億の構成要素を配置し、

- 処理速度
- 消費電力
- 発熱
- 信号遅延

- 製造可能性
- 面積
- コスト

を同時に最適化しなければなりません。

AI は、膨大な設計候補を探索し、人間の設計者を支援できます。

より優れたチップが作られれば、そのチップ上でより高度な AI を訓練できます。

その AI が、さらに優れたチップを設計すれば、次の循環が成立します。

より良い AI

→ より良い半導体設計

→ より多くの計算資源

→ さらに良い AI

ただし、設計図ができても、実物の半導体が直ちに完成するわけではありません。

製造設備の建設、材料調達、試作、検査には物理的時間が必要です。

この点が、再帰的改善の重要な減速要因になります。

5.2 エネルギーとデータセンター

AI がデータセンターの運用を改善すれば、

- 冷却効率
- 電力配分
- 計算タスクの割当
- 故障予測
- 通信経路

- 稼働率

を高められる可能性があります。

さらに、発電、送電、蓄電設備の設計や運用を AI が改善すれば、AI に利用できる計算資源を増やせます。

この場合、自己改善は AI モデル内部だけでなく、AI を支える産業基盤全体へ広がります。

5.3 ロボットと製造能力

ASI が物理世界で急速に能力を拡大するには、設計だけでなく製造も必要です。

AI が工場、物流、採掘、建設、ロボット制御を改善できれば、ハードウェアの生産速度を高められる可能性があります。

しかし、物質を動かし、工場を作り、資源を採掘する速度には限界があります。

このため、純粋なデジタル自己改善が高速でも、物理的設備の拡張が追い付かなければ、全体の進歩は抑制されます。

第６節　第三の形――データによる自己改善

原報告書が重視するもう一つの経路が、データを通じた自己改善です。

AI が、

- 高品質データを選別する
- 誤りを修正する
- 新しい訓練問題を作る
- シミュレーションを生成する
- 自己対戦によって経験を作る
- 推論結果を次の学習へ戻す

ことで、次世代 AI を改善します。（原報告書、Section 5.3）

これは、現在の AI 発展においても比較的現実的な経路です。

6.1 学習データは固定された資源ではない

従来の機械学習では、データは人間が外部から集めるものと考えられてきました。

しかし高度な AI は、自ら学習すべきデータを生成できる可能性があります。

たとえば、

- 自分が苦手な問題を特定する
- その能力を伸ばす問題を作る
- 複数の解法を探索する
- 正しさを検証する
- 最良の解法を学習に使う

という循環です。

これにより、学習は人間が用意した固定的なデータセットから、自律的に変化するカリキュラムへ移行します。

6.2 AlphaZero 型の自己改善

原報告書は、AlphaZero をデータによる再帰的自己改善の代表例として説明しています。

AlphaZero 型システムでは、

1. 現在の方策・価値ネットワークを使う
2. そのネットワークを事前知識として探索する
3. 探索によって、元のネットワーク単独より良い行動を見つける
4. 探索結果をネットワークへ学習させる
5. 改良されたネットワークが、次の探索をさらに効率化する

という循環が成立します。

ここで重要なのは、推論時に大量の計算を使って得られた優れた判断を、次の学習へ圧縮することです。

つまり、

推論時計算

→ より良いデータ

→ より良い学習済みモデル

→ より効率的な推論

という再帰的改善です。（原報告書、Section 5.3）

6.3 自己対戦という開かれた環境

AlphaZero では、自分自身との対戦が新しい訓練データを生みます。

相手も自分とともに強くなるため、学習課題の難易度が自動的に上昇します。

これは固定された試験問題を解くだけの学習とは異なります。

AI 自身の能力向上に合わせて、環境も難しくなるからです。

原報告書は、AlphaStar のリーグ形式のような仕組みを、より複雑な自己適応型環境の例として挙げています。（原報告書、Section 5.3）

6.4 推論結果を学習へ戻す

将来の AI が一つの難問について、

- 多数の解答候補を生成
- 検証器で評価
- シミュレーションで確認
- 最良の解法を選択

する場合、その処理には多くの推論時計算が必要です。

しかし、一度見つけた解法を次世代モデルへ学習させれば、次回は少ない計算で同じ問題を解けます。

これを多数の利用者や AI エージェントの経験について繰り返せば、利用時の計算が、継続的に訓練データへ変換されます。

原報告書は、フロンティア AI における推論時計算の増大と、膨大な利用規模を考えれば、この種のデータ循環に強い経済的動機があると指摘しています。（原報告書、Section 5.3）

6.5 合成データの危険

AI 生成データを使った自己改善には危険もあります。

- AI の誤りを正解として学習する

- 同じ偏りを繰り返す
- データの多様性が失われる
- もっともらしい虚偽が増幅される
- 自己評価だけで誤った方向へ進む

可能性です。

自己改善が成立するには、生成したデータを正しく評価する仕組みが必要です。

数学的証明、コード、ゲームの勝敗のように、結果を明確に検証できる領域では比較的容易です。

しかし、社会政策、倫理、教育、芸術、長期的科学仮説のように、正解が不明確な領域では自己評価が難しくなります。

したがって、データによる自己改善の速度は、

信頼できる評価信号をどれだけ作れるか

に強く依存します。

第７節　第四の形――分業と専門化

原報告書は、AI 集団における分業も、再帰的改善の一形態として扱っています。

ある AI が特定分野へ専門化すれば、その課題をより効率的に処理できます。

専門化によって必要資源が減れば、余った資源でさらに多くの AI を稼働できます。

AI 集団が大きくなれば、さらに細かな専門化が可能になります。

この循環は次のようになります。

1. AI が専門化する
2. 作業効率が高まる
3. 同じ成果を少ない資源で得る
4. 余った資源で個体数を増やす
5. さらに細かな分業を行う
6. 集団全体の生産性が上がる

原報告書はこれを、人間社会における分業と協力に対応する sociogenic recursive self-improvement――社会生成的再帰的自己改善として論じています。（原報告書、Section 5.3）

7.1 人間の専門化との違い

人間が専門家になるには長い教育と経験が必要です。

しかし AI は、

- プロンプト
- 外部ツール
- 専門データ
- 追加学習

- 専門的な作業環境

によって比較的短時間で専門化できる可能性があります。

そのため、人間社会と同じ形の分業が AI 集団にも大きな利益を与えるかは、まだ分かっていません。

現在の基盤モデルは、最大限に汎用的な一つのモデルを作る方向が中心であり、フロンティア水準の専門 AI 集団については実証データが限られていると原報告書は指摘しています。（原報告書、Section 5.3）

7.2 協力が必要になる

分業には協力が必要です。

異なる AI が、

- 課題を分ける
- 結果を共有する
- 相互に検証する
- 矛盾を解消する
- 最終判断を統合する

仕組みが必要です。

協調が不十分なら、専門化による利益よりも、情報伝達や統合作業の費用が大きくなる可能性があります。

この点は、第 7 回で扱うマルチエージェント集合知と直接関係します。

第８節　人間の進化との三つの類似

原報告書は、AI の再帰的自己改善を、人間の能力を発展させてきた三種類の進化と比較しています。

1. 遺伝的進化
2. 文化的進化
3. 協力的・社会的進化

8.1 遺伝的進化——設計図の改善

人間の場合、身体や脳の設計情報は遺伝子に保存されています。

遺伝的進化は世代交代を通じて非常にゆっくり進みます。

AI の場合、対応する「設計図」は、

- ソースコード
- モデル構造
- 最適化手法
- エージェントの作業環境
- 半導体設計

です。

AI がこれらを対象を絞って変更できるなら、人間の遺伝的進化よりはるかに速く進む可能性があります。

原報告書は、これを genotypic RSI——遺伝子型の再帰的自己改善と呼んでいます。（原報告書、Section 5.3）

8.2 文化的進化——知識と道具の改善

人間の知的能力を大きく高めたのは、遺伝的変化だけではありません。

- 言語
- 文字
- 書籍
- 教育
- 科学
- 数学
- 法制度
- 工具
- コンピュータ

などの文化的蓄積です。

個人の脳が大きく変化しなくても、文化的知識を利用することで、人間集団の能力は高まりました。

AI の場合は、

- 高品質データ
- 合成データ
- 学習カリキュラム
- ツール
- ソフトウェア
- 外部記憶
- 検証方法
- 推論結果

が、文化的人工物に対応します。

AI はこれらを、人間より高速に生成、共有、利用できる可能性があります。

原報告書は、AI の文化的進化速度が人間より高くなり得る点を、再帰的改善の重要な加速要因として挙げています。（原報告書、Section 5.3）

8.3 協力的進化——分業と組織

人間は、一人ひとりの知能だけで現代社会を作ったのではありません。

専門化、分業、交換、組織、制度によって集団能力を高めました。

- 科学者
- 技術者
- 医師
- 教員
- 法律家
- 行政官
- 経営者

が異なる役割を担い、協力します。

AI 集団も、専門化と組織化によって能力を高められる可能性があります。

ただし、AI の専門化が人間と同じ時間と費用を必要としないなら、分業の性質も人間社会とは異なるでしょう。

第９節　弱い再帰的改善はすでに始まっている

原報告書は、完全自律型の自己改善がまだ実現していなくても、弱い形の再帰的改善循環はすでに存在すると述べています。

たとえばAIは、

- 研究コードの作成
- 実験計画
- 結果分析
- アーキテクチャ探索
- ハイパーパラメータ最適化
- 半導体設計
- 学習カリキュラム作成
- 世界モデルによるシミュレーション

を支援しています。

これらによってAI研究の速度が上がり、その成果が次世代AIに反映されるなら、人間が中心にいる場合でも再帰的循環の一部です。（原報告書、Section 5.3）

9.1 Human-in-the-loop

現在の多くの改善循環では、人間が重要な位置にいます。

- 研究課題を決める
- AIの提案を評価する
- 実験を承認する
- 結果を解釈する
- 次に進む方向を決める
- 安全性を確認する

この形では、AIは研究を加速しますが、研究全体を自律的には進めません。

再帰的改善の強さを考える際には、

> **AIが研究の何%を担当するか。**

だけでなく、

> **どの重要工程に人間の判断が残っているか。**

を見る必要があります。

研究方向の設定や評価が人間に依存していれば、改善速度は人間の処理速度に制約されます。

9.2 AI Scientist

原報告書は、仮説生成、実験、分析、論文作成などを統合する「AI Scientist」型システムにも言及しています。

こうしたシステムが発展すれば、人間の関与を減らしながら研究循環を回せる可能性があります。

ただし、研究らしい形式を自動生成することと、重要で正しい科学的発見を自律的に行うことは同じではありません。

重要なのは、

- 新規性
- 正確性
- 再現可能性
- 研究上の重要性
- 現実との対応
- 長期的価値

を評価できることです。

第 10 節　知能爆発とは何か

再帰的自己改善が十分に強ければ、AI 能力の増加が急激になる可能性があります。

これは一般に、intelligence explosion——知能爆発と呼ばれます。

基本的な論理は次のとおりです。

1. AGI が AI 研究を行える
2. AI 研究によって、より高性能な AI が作られる
3. 高性能 AI は、より速く優れた AI 研究を行う
4. 研究速度がさらに上がる
5. 改善間隔が次第に短くなる

通常の指数関数的成長では、一定期間ごとに同じ倍率で増えます。

知能爆発では、改善能力自体が高まるため、増加率も上昇します。

原報告書は、完全自律的な自己改善が長い能力範囲にわたって持続する場合、超指数関数的、さらには双曲線的な成長が理論上起こり得るとしています。（原報告書、Section 5.3）

10.1 双曲線的成長

双曲線的成長では、一定の有限時間へ近づくにつれ、数値が理論上急激に増大します。

もちろん、現実の AI 能力が数学的な無限大に達するという意味ではありません。

物理資源には限界があるため、必ずどこかで成長は鈍化します。

しかし、双曲線的な初期動態があれば、短期間に非常に大きな能力差が生じる可能性があります。

10.2 Soft Takeoff と Hard Takeoff

再帰的自己改善をめぐっては、一般に二つのシナリオが区別されます。

Soft Takeoff

能力向上が数年から数十年かけて比較的緩やかに進む。

社会には適応時間がありますが、制度変更を継続的に迫られます。

Hard Takeoff

能力向上が数か月、数週間、あるいはそれ以下の短期間に進む。

社会制度、規制、安全対策が追い付かない可能性があります。

原報告書はこの用語を中心には置いていませんが、再帰的改善が急激な AGI から ASI への移行を生み得るという議論は、この区別と重なります。

第 11 節　なぜ知能爆発は必ずしも起こらないのか

原報告書は、再帰的自己改善の可能性を重視しながらも、その進展が急速であるとは断定していません。

改善循環が早期に弱まる可能性があるからです。（原報告書、Section 5.3）

11.1 改善しやすい部分が先に尽きる

初期には、明らかな非効率や単純な欠陥を改善できるかもしれません。

しかし改善が進むにつれ、残された問題は難しくなります。

たとえば、

- 初期改善は数%の性能向上を容易に生む
- 後期改善には膨大な実験が必要
- 新しい能力を得るには根本的理論が必要
- 計算資源当たりの改善幅が小さくなる

可能性があります。

この場合、自己改善は急速に始まっても、途中で頭打ちになります。

11.2 自分より優れたものを設計できるとは限らない

人間は自分より高速な計算機を設計できます。

しかし、知的システムが常に自分より優れた後継システムを設計できるとは限りません。

自分の内部構造を完全に理解することは困難かもしれません。

変更後のシステムが期待どおり動くことを証明できない場合もあります。

改良によって、別の能力や安全性が損なわれる可能性もあります。

11.3 評価問題

改善案を作ることより、それが本当に改善かどうかを判断する方が難しい場合があります。

たとえば、

- ベンチマークの点数は上がったが、一般性が下がる
- コードは高速になったが、安全性が低下する
- 推論は強くなったが、欺瞞能力も高まる
- 学習速度は上がったが、誤情報を覚えやすくなる

可能性があります。

AI が自分自身を評価する場合、評価器と被評価システムが同じ欠陥や偏りを共有する危険があります。

11.4 実験費用の増加

より高性能なモデルほど、訓練や評価に多くの計算資源を必要とする可能性があります。

AI が新しいモデル案を短時間で生成しても、それを訓練して結果を確認するのに数週間または数か月かかれば、改善循環は遅くなります。

原報告書は、純粋なデジタル研究者であっても、より大きな実験を実行し、結果を待つ必要があると強調しています。（原報告書、Section 5.3）

11.5 物理世界の速度

半導体、ロボット、エネルギー設備、科学実験の改善には、物理的作業が必要です。

- 工場を建設する
- チップを製造する
- 新薬の作用を観察する
- 材料を合成する
- ロボットを試験する
- 発電設備を増設する

といった作業は、コンピュータ内部の思考速度だけでは加速できません。

原報告書は、物理的操作を必要とする開発が自己改善動態を抑制すると述べています。（原報告書、Section 5.3）

11.6 データと現実による接地

AI が自分の生成物だけを学習し続ければ、現実から離れる危険があります。

新しい科学的知識には、

- 観測
- 測定
- 実験
- 外部世界からの反証

が必要です。

このため、再帰的自己改善が人間の知識枠組みを超えて進むには、物理世界と能動的に相互作用し、新しい抽象概念を形成しなければならない可能性があります。

原報告書は、この問題を「抽象化の壁」と結び付け、物理世界との相互作用が改善速度を経験科学の速度へ抑える可能性を論じています。（原報告書、Section 5.3）

11.7 資源消費の急増

能力を一段高めるために必要な計算資源が、改善ごとに急増するなら、再帰的循環は維持できません。

たとえば、

- 能力を 2 倍にするため、計算資源が 100 倍必要
- 次の改善には 1 万倍必要
- 電力と半導体の供給が追い付かない

という状況です。

原報告書は、再帰的改善が途中で消失する可能性や、循環維持に必要な資源が急増する可能性を明示しています。（原報告書、Section 5.3）

第 12 節　AGI が人間より優秀な AI 研究者でなくても加速は起きる

再帰的自己改善には、最初の AGI が世界最高の AI 研究者を上回る必要はありません。

仮に一体の AGI が平均的な人間研究者と同程度でも、

- 大量に複製できる
- 24 時間稼働できる
- 人間より高速に動ける
- 経験を共有できる
- 同時に多数の実験を行える

なら、研究開発全体へ大きく貢献できます。

原報告書は、AGI が一人の人間研究者より優れていなくても、スケーリングによって AI 研究開発に重要な役割を持ち、他の摩擦に直面するまで進歩を加速する可能性があると述べています。（原報告書、Section 5.3）

つまり、知能爆発の条件は、

> **最初の AGI が超天才であること**

ではありません。

> **十分な数と速度で AI 研究へ投入できること**

かもしれません。

第 13 節　自己改善と安全性

AI が AI を改善できるようになると、安全性の問題はより複雑になります。

13.1 能力向上と安全性維持

ある AI が安全に設計されていても、その AI が作った次世代 AI が同じ安全性を保つ保証はありません。

能力を高める変更が、

- 制御可能性
- 透明性
- 人間の指示への従順性
- 不確実性の表明
- 修正可能性

を低下させる可能性があります。

再帰的改善では、この問題が世代ごとに反復されます。

13.2 検証済みプログラム合成

原報告書は、安全な自己変更の一つの可能性として、verified program synthesis——検証済みプログラム合成を挙げています。

これは、AI が作ったコードについて、

- 指定された条件を満たす
- 重要な安全性を壊さない
- 特定の不具合を生じさせない

ことを形式的に検証する方法です。

ただし、複雑な AI 全体について、あらゆる重要性質を完全に証明するのは困難です。

何を検証条件に含めるかという問題も残ります。

13.3 ゲーデル・マシン

自己改善を理論化した代表的な構想に、Jürgen Schmidhuber の Gödel Machine があります。

これは、自己変更によって期待効用が向上することを証明できた場合に、自らを書き換える理論的エージェントです。

しかし原報告書は、この方式には、

- 自己についての完全な知識
- 証明可能性
- ゲーデルの不完全性定理

に関する限界があると述べています。（原報告書、Section 5.3）

つまり、安全で最適な自己変更を常に完全証明できるとは限りません。

13.4 反復増幅

より実践的な考え方として、原報告書は Iterated Amplification——反復増幅に言及しています。

これは、難しい課題を小さな課題へ分解し、比較的弱い AI と人間の組み合わせによって解き、その結果を学習させる方法です。

この循環を繰り返しながら能力を高めます。

目的は、能力を増幅しつつ、人間の判断や価値との整合性を維持することです。

第 14 節　再帰的改善の速度をどう測るか

この経路について最大の問題は、歴史的な前例がほとんどないことです。

そのため、過去の傾向を単純に当てはめて予測できません。

原報告書は、不確実性を減らすために、recursive improvement scaling laws——再帰的改善のスケーリング則を作る必要があると提案しています。（原報告書、Section 5.3）

14.1 測るべき指標

たとえば、次のような指標が必要です。

- AI が研究者の作業時間を何％削減したか
- AI 支援で実験数が何倍になったか
- AI が提案した改善の採用率
- AI が発見したアルゴリズムの性能向上幅
- 人間の介入が必要な工程数
- 一世代の改善に必要な計算量
- AI 能力と研究生産性の関係
- 改善サイクルが世代ごとに短縮しているか
- 能力向上がいつ頭打ちになるか

です。

14.2 人間の関与度

研究自動化の程度を測る際には、人間の関与を単純な作業時間だけで測るべきではありません。

たとえば、AI が研究作業の 95％を行っても、人間が最後の 5％として、

- 研究目的を決める
- 結果を評価する
- 重大な判断を行う

なら、研究循環全体は人間に依存しています。

したがって、

- 作業量
- 判断権限
- 評価権限
- 方向設定
- 安全承認

を区別して測定する必要があります。

第 15 節　再帰的自己改善と他の三経路の関係

再帰的自己改善は、他の経路と独立ではありません。

15.1 スケーリングとの関係

AI が半導体、データセンター、学習効率を改善すれば、より多くの計算資源を利用できます。
その計算資源によって、さらに高度な AI を作れます。
したがって、

> **自己改善**
> **→ スケーリングの加速**
> **→ さらに強い自己改善**

という循環が生じます。

15.2 パラダイム転換との関係

AGI 研究者が新しいアーキテクチャや学習原理を発見すれば、第二経路であるパラダイム転換を加速します。
根本的な効率向上が起きれば、自己改善循環はさらに強くなる可能性があります。

15.3 マルチエージェント集合知との関係

多数の AI 研究者が専門化し、協力すれば、個体では不可能な規模の研究を行えます。
その集団が自らの分業や組織を改善すれば、集合体全体として再帰的自己改善します。
したがって、第三経路と第四経路は深く重なります。

第 16 節　再帰的自己改善は「大学の自動化」でもある

AI が AI 研究を自動化するということは、研究活動の多くが大学や人間研究者だけのものではなくなることを意味します。

将来、AI が、

- 文献調査
- 仮説形成
- 数学的分析
- 実験計画
- データ分析
- 論文作成
- 相互評価

を行うようになれば、研究機関としての大学は大きく変わります。

問題は、研究者の仕事の一部が効率化されることだけではありません。

知識を生産する主体そのものが変わることです。

第 17 節　AI が研究する時代に、大学は何を担うのか

研究の技術的工程を AI が担えるとしても、次の問題は残ります。

- 何を研究すべきか
- 誰のための研究か
- どのリスクを受け入れるか
- 研究成果を誰が所有するか
- 社会にどのように適用するか
- 研究を中止すべき条件は何か
- 将来世代への責任をどう考えるか

これらは、研究能力だけでは決まりません。

価値判断、公共性、倫理、政治的正統性が必要です。

したがって、大学の使命は、

> **AI より速く論文を書くこと**

ではなく、

> **AI による研究を、公共的・倫理的・学術的に統治すること**

へ広がる必要があります。

第 18 節　研究の方向設定は誰が行うのか

AI が大量の研究を自動実行できるようになると、何を研究するかという選択が極めて重要になります。

研究資源は有限です。

たとえば、

- 富裕層向け医療
- 希少疾患
- 軍事技術
- 気候変動
- 広告最適化
- 教育格差

のどれを優先するかは、技術的能力だけでは決まりません。

AI は、与えられた目標に対して研究を最適化できます。

しかし、社会全体の研究目標を正当に決定するためには、民主的・公共的な議論が必要です。

大学には、研究の方向性を市場や国家権力だけに委ねず、多様な価値観から吟味する役割があります。

第 19 節　AI が作った知識を人間は理解できるか

ASI が人間の専門家集団を超える研究を行う場合、研究成果が人間には理解できない可能性があります。

AI が、

- 新しい数学体系
- 新しい物理概念
- 非常に複雑な医療モデル
- 人間には直感できない因果構造

を発見したとして、人間はその正しさをどう判断するのでしょうか。

形式的検証が可能な領域では、証明やプログラムによって確認できるかもしれません。

しかし、現実世界についての科学的理論では、観測、実験、解釈が必要です。

人間が理解できない知識を、社会が利用してよいかという問題も生じます。

大学には、AI の発見を人間の概念体系へ翻訳し、批判可能な形にする役割が生まれるでしょう。

第 20 節　論文数ではなく信頼性が中心になる

AI が研究成果を大量生成できる時代には、論文数は知的生産性の有効な指標ではなくなります。

一日に何万本もの論文を生成できても、

- 誤り
- 重複
- 無意味な結果
- 再現不能な発見
- 自己引用的な知識循環

が増えれば、学問は進歩しません。

重要になるのは、

- 再現性
- 検証可能性
- 研究の重要性
- 外部世界との対応
- 異なる AI による独立検証
- 人間社会への意味

です。

大学と学術制度は、知識生成から知識の選別・検証へ重点を移す必要があります。

第 21 節　大学自身も再帰的に改善できるか

再帰的自己改善という考え方は、大学にも適用できます。

大学が AI を用いて、

- 教育方法を評価する
- 研究支援を改善する
- 組織運営を分析する
- 学際連携を促進する
- 学習成果を検証する

ことができれば、大学の制度そのものを改善できます。

改善された大学が、さらに優れた研究と教育を生み、それをもとに制度を再改善するなら、組織的な正のフィードバックが成立します。

しかし大学の改善は、単なる効率化であってはなりません。

- 教員一人当たり学生数を増やす
- 論文数を最大化する
- 管理コストを減らす
- 学習時間を短縮する

ことだけを目標にすれば、教育と研究の質が損なわれる可能性があります。

大学の再帰的改善には、

- 人間形成
- 学問の自由
- 公共性
- 多様性
- 社会的責任

を含む複数の価値基準が必要です。

第 22 節　人間と AI の共同改善

再帰的自己改善を、AI だけが自律的に自分を高める過程として考える必要はありません。

より現実的で望ましい形は、人間と AI の共同改善かもしれません。

AI が人間研究者を支援し、人間が、

- 研究目的
- 倫理的制約
- 社会的意味
- 長期的方向性

を与える。

その結果生まれた AI が、さらに人間の教育、研究、判断能力を支援する。

この循環では、

> **より良い AI が、より良い人間の判断を支え、**
>
> **より良い人間の判断が、より良い AI を形成する**

ことを目指します。

これは、能力だけを最大化する知能爆発とは異なる、共同的な知的発展です。

第 23 節　教育は研究方向を判断できる人間を育てなければならない

AI が研究の多くを行えるようになれば、大学教育では、研究作業の技法だけでなく、

- どの問いが重要か
- どの証拠を信頼するか
- どの価値を優先するか
- 不確実性の中でどう判断するか
- AI の提案をどう批判するか
- 結果に誰が責任を持つか

を教える必要があります。

これは人文学、社会科学、自然科学、工学を分離した教育では十分に扱えません。

AI 時代の研究判断には、学際的な理解が必要です。

第 24 節　Ecological University との関係

Ronald Barnett の Ecological University は、大学を複数の生態系に責任を負う制度として捉えます。

再帰的自己改善は、AI 能力だけを中心に見れば、可能な限り速く循環を回すことが望ましいように見えます。

しかし、AI 研究の加速は、

- エネルギー消費
- 資源利用
- 労働市場
- 権力集中
- 軍事競争
- 科学の公開性
- 民主的統治
- 人間の自律性

へ影響します。

したがって、改善速度だけを最大化してはなりません。

大学は、AI の自己改善循環を、社会、自然、文化、知識、人間形成の各生態系との関係から評価する必要があります。

第 25 節　「止められる自己改善」が重要である

安全な再帰的自己改善に必要なのは、改善速度だけではありません。

- 人間が監査できる
- 改善理由を説明できる
- 問題があれば停止できる
- 過去の安全な状態へ戻せる
- 権限を段階的に拡大する
- 独立した評価を受ける

ことが必要です。

つまり、重要なのは自己改善可能性だけでなく、修正可能性と停止可能性です。

能力が向上するほど、人間が介入できなくなる設計は、技術的に成功しても社会的には危険です。

結論

『From AGI to ASI』の Section 5.3 は、再帰的自己改善を、AI が自分のコードを直接書き換える単一の過程としては捉えていません。

再帰的自己改善とは、

> **AI が AI 研究開発を促進し、その結果生まれた改良型 AI が、次の AI 研究開発をさらに促進する正のフィードバック**

です。

原報告書は、その具体的な形を次の四つに整理しています。

1. アルゴリズムとソフトウェアの改善
2. 半導体、製造、エネルギーを含むハードウェア改善
3. 合成データ、自己対戦、推論結果の蒸留によるデータ改善
4. 専門化と分業による集合的な効率改善

これらは、人間の遺伝的進化、文化的進化、協力的進化に対応するものとして理解できます。（原報告書、Section 5.3）

弱い形の再帰的改善は、すでに始まっています。

AI は研究コード、実験計画、アーキテクチャ探索、半導体設計、アルゴリズム発見を支援しています。

しかし、完全自律的な自己改善循環が長期間持続し、AGI から ASI への急激な移行を引き起こすかどうかは分かりません。

改善を抑える要因には、

- 改善課題の難化
- 評価の困難
- 計算資源の急増
- 実験結果を待つ時間
- 半導体や設備の製造時間
- データと現実世界への接地
- 安全性と検証の問題

があります。

したがって、知能爆発は論理的に可能な一つのシナリオですが、必然ではありません。

原報告書が求めているのは、楽観または悲観による断定ではなく、

- AIがAI研究をどの程度加速しているか
- 人間がどの工程に残っているか
- 改善一世代当たりの費用がどう変化するか
- 改善速度が加速しているか、頭打ちになっているか

を継続的に測定し、再帰的改善のスケーリング則を構築することです。（原報告書、Section 5.3）

そして高等教育にとって、この経路が意味するのは、AIが大学研究を支援するというだけではありません。

研究を行い、知識を生み、次の研究者を形成する循環そのものに、AIが組み込まれることです。

そのとき大学に求められる中心的役割は、AIより速く研究することではありません。

何を研究すべきかを問い、AIが生み出す知識を検証し、異なる価値を調整し、技術進歩を公共的・倫理的に統治することです。

第7回では、原報告書のSection 5.4に進み、第四の経路であるMulti-Agent Coordination and Group Agency——多数のAGIが専門化、協力、競争することで、単一の超人的AIが存在しなくても集合体としてASIが成立する可能性を詳しく解説します。