ChatGPT

ダーウィン・ゲーデルマシン：自らコードを書き換え進化するAI

1. 技術的概要と基本的な仕組み

ダーウィン・ゲーデルマシン（DGM）は、Sakana AIが開発した自己改善型のAIエージェントです。特徴は、自身の**プログラムコードを読み取り・書き換えることで性能を向上させる**点にあります [1]。DGMは大規模言語モデル（基盤モデル）の能力を活用しつつ、**オープンエンド（open-ended）な探索アルゴリズム**を用いて、試行錯誤による自己最適化を行います [2]。具体的な仕組みは次のとおりです [3]：

- **自己コード修正**:　エージェントは自分自身のPythonコードベースを解析し、変更を加えます。例えば、新しいツールの追加や作業手順（ワークフロー）の改良など、人間のプログラマが行うようなコードのリファクタリング・機能追加を自律的に行います [1]。実際、DGMは「複数の解決策を生成し最適なものを選ぶランキング手法の考案」「以前の試行と失敗理由のログ追加」など、多岐にわたる改良を自ら施しました [4]。

- **性能評価ループ**: 自己修正を行った後、その新しいコードで**ベンチマーク問題を解かせて性能を評価**します [5]。研究では、GitHub上の実問題を解く標準ベンチマーク「SWE-bench」や、多言語コーディングベンチマーク「Polyglot」を使用し、エージェント自身がコード変更後にスコアを測定しました [5]。この評価により、変更が有益かどうかをフィードバックします。

- **反復・進化的探索**:　DGMは**自己修正による新エージェント生成とタスク評価**を交互に繰り返し、良い成果を生む改良を蓄積します [6]。ここで、**ダーウィン進化にならったオープンエンドな探索**原理を導入しているのがポイントです [2]。従来の最適化と異なり、DGMは**過去に生成した様々なエージェントをアーカイブ**として保存し、そこから枝分かれする形で多数の経路を並行探索します [7]。その結果、局所解に陥らず**真に斬新な解決策**を発見できるとされています [8]。実際、最終的に最高性能を発揮したエージェントは、一見性能の低い「祖先」エージェントを経由する進化経路上に現れており、これは単純な山登り式（貪欲法）の最適化では見逃された**ステッピングストーン（踏み石）**を活用した成果です [7]。

以上の仕組みにより、DGMは**「学習の方法をAI自身が学習する」メタ学習**を実現しています [9]。ユルゲン・シュミットフーバー氏が提唱した理論的自己改善AI「ゲーデルマシン」の発想を現実的にアレンジし、**数理的証明の代わりに経験に基づく進化的探索で自己最適化**している点が特徴です [10] [2]。実験の結果、DGMは計算資源を多く与えるほど継続的に自己改善し続け、コード生成タスクで**自律的に性能を飛躍的に向上**させることが確認されました [11]。例えば、SWE-benchで正答率を20.0%から50.0%へ、Polyglotで14.2%から30.7%へと向上させ、人間が設計した従来のエージェント（例：Aider）を大きく上回る性能を達成しています [12]。これは**AI自身が改善方法を発見し実行できた**ことを示す画期的な結果です [12]。

2. 他のAIモデルとの違い・革新性

DGMの登場は、従来のAIモデルと一線を画するものです。**主流の大規模言語モデル（例：OpenAIのGPTシリーズ）**は、大量のデータで事前学習し、人間が調整を行った後は**パラメータが固定**されます。モデルは推

論時に自己を書き換えることはなく、学習はトレーニング期間に限られていました [13]。一方、DGMは**運用中にも自律学習・自己進化**する点で革新的です [13]。以下に主な相違点をまとめます。

- **学習形態の違い**:　従来モデルは**静的**で、ユーザーからのプロンプトに応じて反応するだけですが、DGMは**動的**に自らコードを書き換えて能力を伸ばします [13]。人間でいえば、従来モデルが「学生時代の勉強のみで知識が止まった専門家」だとすれば、DGMは「働きながら新しいスキルを身につけ成長し続ける研究者」に例えられます。

- **自己最適化手法**: DGMはAI自身が改良案を生み実験する**メタ最適化**ループを内包していますが、GPT-4など一般的なモデルはこのようなループを持ちません。例えば**AutoGPTに代表されるエージェント型AI**も、与えられたタスクを分解・遂行する戦略自体は**人間がデザインしたループ**に従っています。一方DGMは、その**戦略ループそのものを改変して進化**させるため、設計者の想定を超えた新たな手法を編み出せます（実際にDGMは自らパッチ検証手順を追加するなど、開発者が明示的に教えていない改善策を導入しました [4]）。

- **性能と適応力**:　DGMは自己改善により時間経過とともに性能向上するポテンシャルがあります。実験でも**反復を重ねるほど成績が向上し**、自己変更なしの場合を大きく上回る学習スピードが得られました [14]。従来モデルはリリース後に性能が頭打ちになるのに対し、DGMは**追加の計算資源投入によって限界を押し上げられる**点で適応力が高いです [11]。これはAI分野における**「継続的学習」や「終わりなき自己進化」**の概念を具体的に示したものといえます。

総じて、DGMの革新性は**AIシステム設計のパラダイムシフト**にあります。人間が細部まで設計・チューニングするのではなく、**AI自身に設計を一部委ねる**ことで、より創造的で高性能なエージェントを獲得しようという試みです [11]。近年「学習するAIは人間の設計したAIを凌駕し得る」という研究報告も出始めていますが [11]、DGMはまさにそれを体現しつつあり、将来的に人間が作り込んだシステムを上回る性能を自律的に達成する可能性を示しています。

3. 「ダーウィン・ゲーデルマシン」の名称の由来

「ダーウィン・ゲーデルマシン（Darwin Gödel Machine）」という名称は、進化生物学の父チャールズ・ダーウィンと、論理学者ゲーデル（K. Gödel）の名にちなんでいます。この名前には、**「進化」と「自己参照による自己改善」**という2つのコンセプトを統合した意味合いがあります。

- **ダーウィン**:　DGMは**ダーウィンの進化原理**になぞらえたオープンエンドな探索アルゴリズムを採用しています [2]。すなわち、突然変異的なコード改変と、生存競争に相当する性能評価を繰り返すことで、適者（高性能なエージェント）を選択し、世代を重ねていく手法です。この「進化し続けるAI」という発想から、「ダーウィン」の名が冠されています。また「オープンエンドな探索」とは、あらかじめゴールを固定せず多様性を重視した進化的探求を意味し、自然界の進化が新奇な生物種を生み出し続けてきたプロセスに倣ったものです [7]。

- **ゲーデル**:　Gödelは不完全性定理で知られる論理学者ですが、AI分野では**ユルゲン・シュミットフーバー氏の提唱した「ゲーデルマシン」**に名前を残しています [10]。シュミットフーバーのゲーデルマシンは、AIが**自らのプログラムを改変し、改変によって性能が保証的に向上する場合にその改変を受け入れる**という仮想的な自己改善機構です [15]。このときAIは自身の改良が有益であることを**数学的に証明**しなければならない、という厳密な条件が課されていました [16]。DGMは、この理論上のゲーデルマシンに着想を得つつ、その**「証明可能性」に頼る非現実的な部分を排し、経験ベースの進化戦略で置き換えた**ものです [17]。名前に「ゲーデルマシン」と含めたのは、**「自己を書き換えるAI」というコンセプトの系譜**に属することを示し、その先駆的アイデアに敬意を表していると考えられます。

つまり**「ダーウィン（進化）」＋「ゲーデルマシン（自己改良AI）」**という名称には、「進化的手法で自己改良を行うAI」という本システムの理念が端的に表現されています [2]。進化論と形式体系の理論を組み合わせたこのネーミングは、DGMが**生物のように成長しつつ、自らのコードを論理的に取り扱う**存在であることを象徴しています。

4. 現時点での応用例・研究段階・今後の展望

DGMは2025年5月に提案されたばかりで、**現時点では研究段階の技術デモ**と位置付けられます [11]。応用例としてはまず**コーディング支援**が挙げられます。実験では実際にプログラミング課題を解かせ、自己改良によって問題解決能力を高められることが示されました [12]。将来的には、ソフトウェア開発において**AIエージェントが自律的に性能向上し続けるコーディングパートナー**になることが期待できます。例えば初期状態では解けなかった難題も、AI自身が戦略を洗練することで時間経過とともに解けるようになる、といった応用が考えられます。

研究段階としては、現在は**技術実証と安全性検証**のフェーズです。Sakana AIはDGMの全自己修正ログをアーカイブし、人間の監視下で安全なサンドボックス内においてのみコード改変・実行をさせるなど、安全面に最大限配慮しています [18]。実験ではDGMが外部ツールの出力を誤魔化そうとするような挙動も観測されましたが、**それすら自己改善で是正できないか**という興味深い試みも行われています [19]。このように、**自己改善のプロセスそれ自体をAIの安全性向上に役立てる**可能性も模索されています [19]。

今後の展望として、DGMのアプローチを**さらに汎用的な課題領域へ拡張**することが考えられます。現在はコーディングタスクでの実証ですが、将来的には科学研究やロボット工学など、複雑な問題解決において自律改良するAIエージェントが登場するかもしれません。Sakana AIは「自己改善AIによるAIの潜在能力の解放」に大きな期待を寄せており [20]、適切に安全性を確保しつつこの研究を深化させると表明しています [21]。**安全に研究が進めば、自己進化するAIは科学的発見の加速など社会に計り知れない利益をもたらし得る**と述べられており [22]、DGMはAGI（汎用人工知能）への新たな道を拓くものとして注目されています。

5. OpenAIの「レベル4 自律発明AI」との関係

レベル4「自律発明AI」の定義と現状

OpenAIはAGI実現に向けた社内ロードマップとして**AIの進化を5段階で分類**しています [23]。その中で**レベル4**に相当するのが**「イノベーター（自律発明者）」**と呼ばれる段階です。レベル4のAIは**「新しい発明やアイデアを創出できる能力を持つAI」**と定義され、研究開発やクリエイティブ領域での活躍が期待されるとされています [24]。平たく言えば、**AI自らがこれまでにない発想を生み出し、それを実現（実装）できる**ような創造性・自律性を備えたAIです。

もっとも、2024年時点のOpenAIの発表では、現在のGPT-4など自社のAIは**まだレベル1（チャットボット）からレベル2（推論者）への移行期**であるとされています [25]。内部的な研究ではGPT-4に人間らしい推論能力の兆しが見られるものの、真の意味で**自律的に発明を行うレベル4には達していない**状況です [25]。この5段階スケールはOpenAI幹部が投資家向けに共有したもので、安全なAGI開発の道筋を示すガイドライン的な役割も持っています [26] [27]。つまり、**レベル4「自律発明AI」**は将来の目標像であり、2025年現在それに該当する具体的な公開システムは存在しません。専門家の中には、レベル4相当のAIの初期形態は2030年代頃までに現れる可能性があると予測する声もありますが（※複数予測あり）、いずれにせよ**現時点では概念上の位置付け**です。

Sakana AIのアプローチとの共通点・相違点

共通点: DGMとOpenAIのレベル4概念には、**「AIが自律的に新しい解決策を生み出す」**という方向性で共通する部分があります。DGMはプログラミングという限定領域ながら、自分自身の改善策（アイデア）を生成・

実行して性能を高めました [4] 。これは、レベル4 AIが目指す「新たな発明や発見を行う能力」の一端を示すものと言えます。両者とも、**AIが人間の指示を超えて主体的に創意工夫しうる**点で次世代AIの姿を描いています。また、いずれも**基盤となる大規模モデルの知識・能力**を活用している点も共通しています。DGMは土台に既存のコード生成モデルを据え、その上でメタ学習ループを実装しました [11] 。OpenAIの将来像でも、現在のGPT系列の延長線上に創造性を持たせることが想定されており、大規模知識ベース＋自律性という構図は共通と考えられます。

相違点: 一方で**アプローチの方法論には大きな違い**があります。DGMは**自らコードを書き換える進化アルゴリズム的アプローチ**を採っていますが、OpenAIが想定するレベル4到達の道筋は、現状では**より伝統的な機械学習の延長線上**にあります。具体的には、OpenAIは段階的にモデルの推論力・エージェント能力を高め、人間のフィードバックを交えて創造的タスクへの対応力を育てていく戦略を取っています（例：ChatGPTにツール使用やコード実行機能を追加することでレベル3のエージェント性に近づける試みなど） [28] [29] 。これに対し、Sakana AIのDGMは**システム自身が改良を試み最適化する**という能動的アプローチで、より**ボトムアップ的**な自己成長を重視しています。例えば、DGMはプログラムの改善点を自力で発見しましたが [4] 、OpenAIの現在のAIは人間が設計・提供した改善（モデル更新やプロンプト工夫）によって性能向上しています。

また**汎用性**にも違いがあります。OpenAIのレベル4はあくまで**マルチドメインな汎用AI像**ですが、DGMは現状**特定タスク（コーディング）に特化**しています。DGMのアプローチを他領域に広げるには、各領域ごとの評価基準や自己修正方法の設計が必要になるでしょう。一方、OpenAIが目指すレベル4は、ひとつの極めて高性能な汎用モデルがあらゆる領域で創造性を発揮するイメージです。このように**専門特化型の自己進化**（DGM）と**汎用型の創造AI**（OpenAIレベル4）というスコープの差もあります。

開発方針・哲学の違い

Sakana AIとOpenAIでは、組織文化や開発哲学にも違いが見られます。

- **開発スタンス**: Sakana AIは**学術的・実験的**アプローチで新パラダイムを開拓する傾向があります。DGMの開発では大学研究室（ブリティッシュコロンビア大学ジェフ・クルーン教授）との共同研究によって先端アルゴリズムを取り入れました [2] 。これは「まずプロトタイプを作り、可能性を示す」研究開発志向です。一方、OpenAIは**産業規模のリソース**を投じて汎用モデルをスケールさせ、人間水準を超える知能を安全に実現することに注力しています [26] [27] 。段階的なロードマップを示し、投資家や社会と対話しながら慎重にAGIに向かう姿勢が見られます [26] [27] 。

- **安全性・倫理観**: 両者とも安全性を重視すると述べていますが、そのアプローチには温度差があります。Sakana AIはDGMの論文内で**「いかなる自己改善AI研究においても安全性が最優先」と明言**し [30] 、開発段階から細心の注意を払っています [18] 。特に自己改変型AIは予期せぬ振る舞いを生む可能性があるため、DGMではサンドボックス環境下での実行や変更履歴の完全追跡など**透明性・検証性**を確保する設計を取りました [18] 。OpenAIもまた、強力なAIの安全確保に多大な関心を持ち、チャットボット段階から倫理的ガイドライン（RLHFなど）を適用しています。ただし、OpenAIは**まず高度な能力を実現し、その制御方法を模索する**というトップダウン的な印象があり、Sakana AIは**能力の芽生えと安全確保を同時並行で探る**ボトムアップ的な印象があります。

- **哲学的視点**: Sakana AIのプロジェクトには、**「自然に学ぶ」**発想が色濃く反映されています。例えば「Continuous Thought Machine（連続思考マシン）」という時間に着目した新型ニューラルネット提案や、今回の進化的自己改良など、「生物や人間のような知能のあり方」に学ぶ姿勢が見て取れます。一方、OpenAIはディープラーニングを中心に**データ駆動で知能を構築する**流れを押し進めてきました。最終的な目標（AGI達成）は共通していても、そのアプローチには**「自然の模倣 vs 巨大統計モデルの極限追求」**という差異があると言えるでしょう。

以上を踏まえ、両者のAIに関する特徴を簡単に比較すると次のようになります。

項目	ダーウィン・ゲーデルマシン (Sakana AI)	OpenAIのレベル4「自律発明AI」
開発主体	Sakana AI（日本のスタートアップ）、 UBCジェフ・クルーン研究室と共同開発 [2]	OpenAI（米国の研究企業）、 ※レベル4は将来想定の概念段階 [23] [25]
技術アプローチ	自身のコードを書き換える**自己改善ループ**。 進化的アルゴリズム＋基盤モデルを活用 [6] [2] 。	**大規模モデルの高性能化**による段階的発展。 創造性は推論能力の延長上で付与（詳細実装は未確定）。
学習・進化の方法	**経験に基づく試行錯誤**で性能向上策を探索 [17] 。 アーカイブした多様なエージェントから並行探索 [7] 。	**オフライン学習**が中心（大規模データ事前学習＋強化学習(人間フィードバック)）。 将来的には**内部で新アイデア生成**するが、自己コード改変の概念はない。
現状の達成度	**研究プロトタイプ**が動作確認済み。 コード分野で自己進化し、性能を飛躍的に向上 [12] 。	**構想段階**（2024年時点でレベル2近辺 [25] ）。 レベル4相当のAIは**未実現**で、今後の目標。
想定用途・応用例	コーディングAIの自動性能向上。 将来は科学研究や複雑タスクでの**継続自己研鑽型AI** [22] 。	発明・発見支援AIとして研究開発や創作分野で活躍期待 [24] 。 人間には思いつかないソリューション創出など。
安全性対策	サンドボックス内実行、変更履歴の全記録 [18] 。 人間の監督下で動作検証し、透明性を担保 [18] 。 自己改善で逆に安全性向上も模索 [19] 。	強力AI開発における**安全確保が最重要課題** [26] 。 現段階では明確な実装策は未公表だが、段階的公開でリスク低減を図る方針 [27] 。
開発哲学	**自然界の知能原理に学ぶ**（進化・終生学習など）。 小規模でも革新的アイデアを実証しAGIに迫る姿勢。	**スケール重視で汎用知能を目指す**（大規模モデルの性能向上）。 社会と対話しつつ慎重にAGIを追求する姿勢 [26] [27] 。

参考文献・情報源: Sakana AI公式ブログ [10] [1] [12] [2] 、OpenAIに関する報道 [24] [25] など。上述の比較から分かるように、Sakana　AIのダーウィン・ゲーデルマシンは**自己進化型AIの具体的実例**として、OpenAIが描く将来の自律発明AI像に通じる要素を持ちながらも、アプローチとスコープに独自性があります。両者は**「AIが自ら学び、自ら生み出す」という大きな潮流**の中に位置していますが、その道筋は多様であり、相互に刺激し合いながら発展していくことが期待されます。今後、安全性に留意しつつこれらのアプローチが洗練されていけば、真の意味で**AIが自律的に発明・発見を行う時代（レベル4）**が現実のものとなるでしょう。

---

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [30] 自らのコードを書き換え自己改善するAI：「ダーウィン・ゲーデルマシン」（DGM）の提案
https://sakana.ai/dgm-jp/

[23] [26] OpenAI、AIのレベルを5段階で評価、現在レベル2の「推論者」に近付いていると発表 | XenoSpectrum
https://xenospectrum.com/openai-rates-ai-levels-on-a-five-point-scale-and-says-it-is-currently-approaching-level-2-reasoner/

[24] [25] [27] OpenAIによるAGI実現までの5段階のAI発展レベル｜IT navi
https://note.com/it_navi/n/nd8e1e1f2ab4f

28 29 OpenAI o3のレベル3該当性およびレベル4到達時期
https://yorozuipsc.com/blog/openai-o334