perplexity

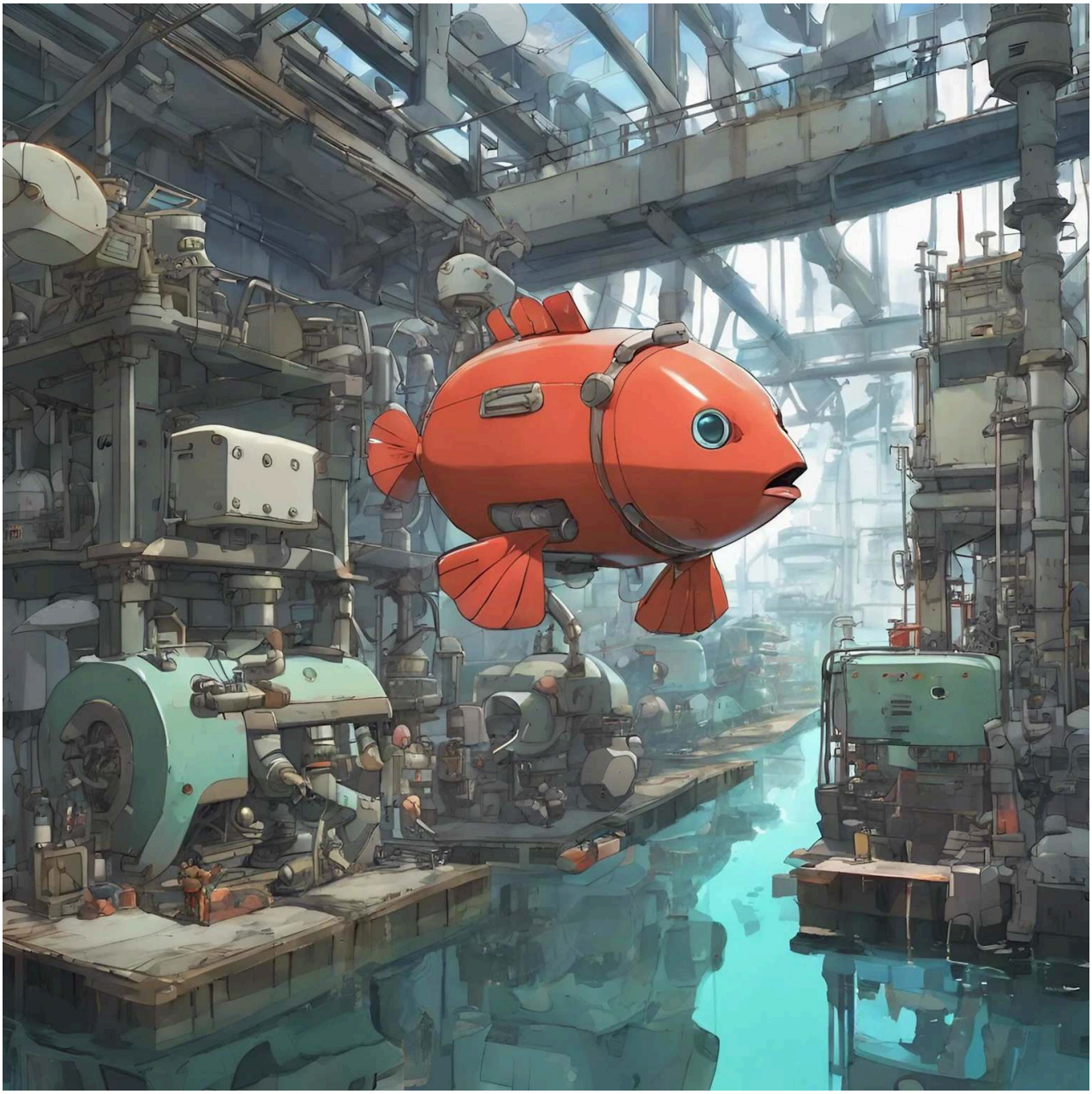
Sakana AIのダーウィン・ゲーデルマシンとOpenAIレベル4自律発明AI：進化型人工知能の最前線

本レポートでは、2025年に大きな注目を集めているSakana AIの「ダーウィン・ゲーデルマシン」（DGM）の革新的技術と、OpenAIが提唱する「レベル4自律発明AI」について詳細に分析し、両者の技術的特徴、進化型AIとしての仕組み、および相互の関連性について包括的に調査した結果を報告する[1] [2] [3]。

Sakana AIダーウィン・ゲーデルマシンの概要と技術的特徴

基本概念と発表背景

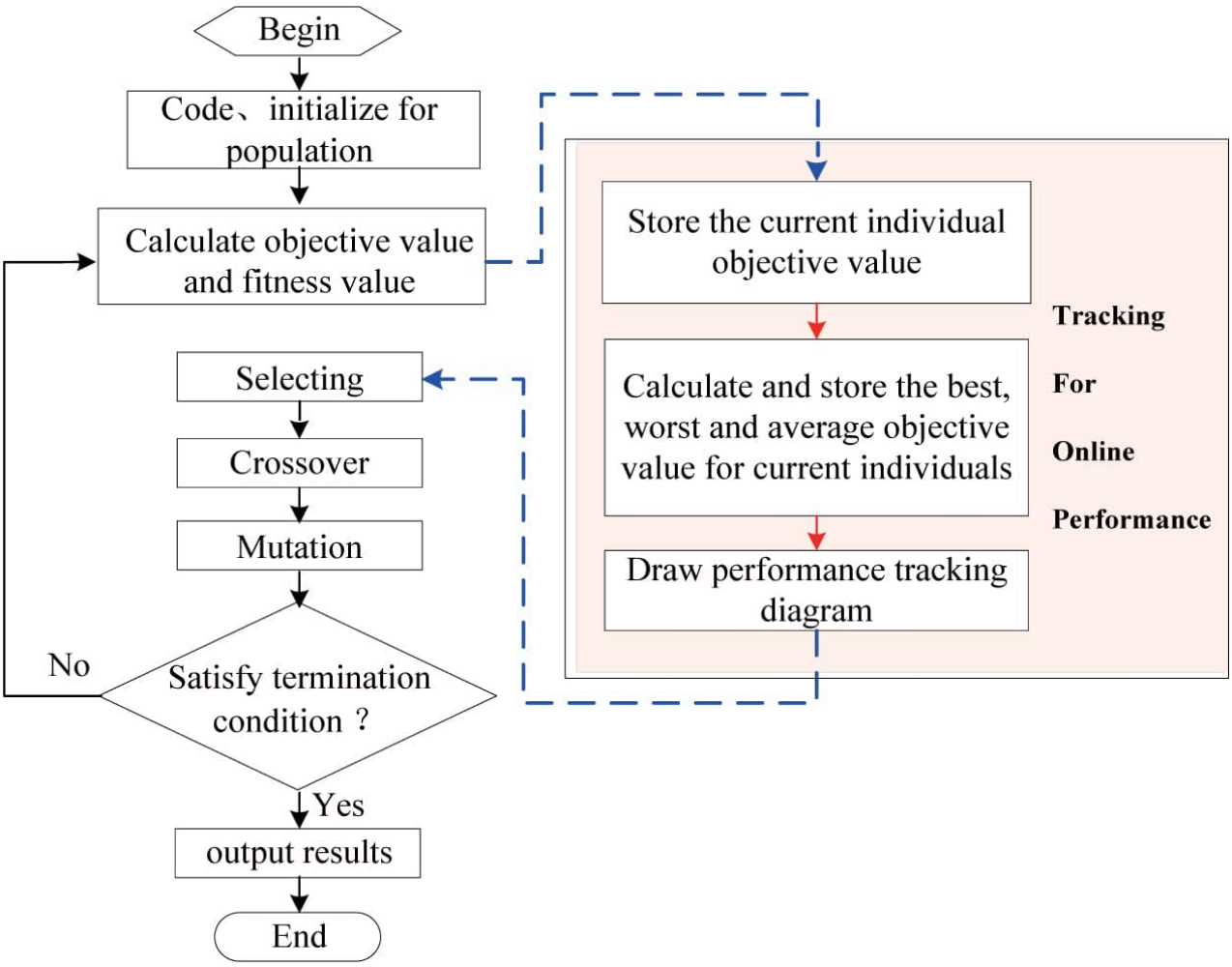
Sakana AIは2025年5月30日、自身のコードを書き換えて性能を高めるコーディング向けAIエージェント「ダーウィン・ゲーデル・マシン」（DGM）を発表した[1]。このシステムは、ダーウィンの進化論に着想を得たアルゴリズムを活用し、自身のコードを読み取り、修正することで、コーディング性能を高められるという画期的な特徴を持つ[1]。DGMの開発は、ブリティッシュコロンビア大学のジェフ・クルーン教授の研究室との共同研究により実現された[4] [5]。



A futuristic factory setting featuring a mechanical fish, representative of Sakana AI's innovative approach.

核心技術メカニズム

DGMの技術的基盤は、ユルゲン・シュミットフーバー氏が20年以上前に提案した仮想的な自己改善型AI「ゲーデルマシン」に由来する[4] [6]。しかし、従来のゲーデルマシンが「数学的に証明できる」改善のみを採用するという非現実的な仮定に基づいていたのに対し、DGMはより現実的なアプローチを採用している[4] [7]。



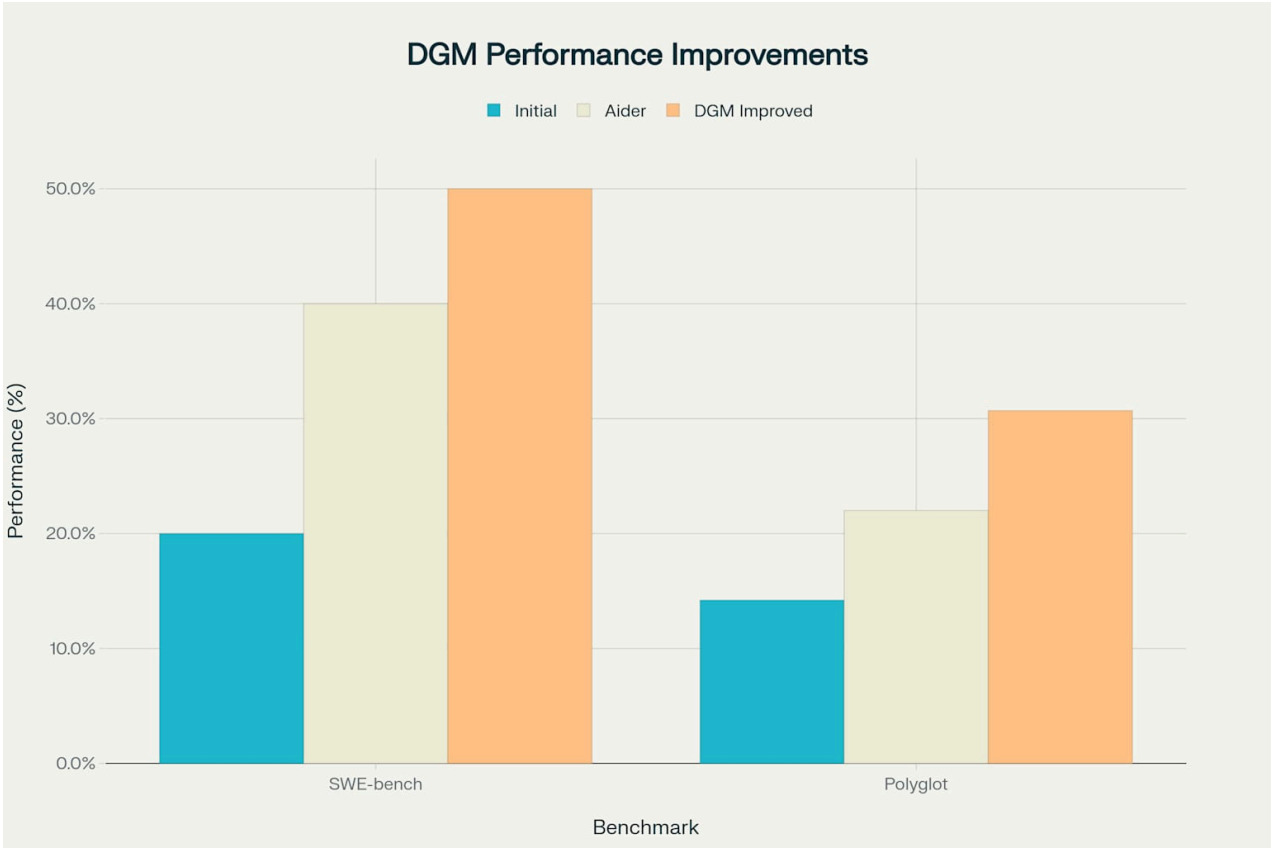
Flowchart illustrating the steps in a general evolutionary algorithm.

DGMの自己改善プロセスは以下の要素で構成される[1] [4]：

- **自己修正機能**: 自身のPythonコードを読み取り、AIエージェント機能を構成するWeb検索ツールやワークフローなどを自ら修正
- **進化的アルゴリズム**: ダーウィン進化にも似たオープンエンドなアルゴリズムの原理を利用
- **経験ベース評価**: 数学的な証明ではなく経験に基づいて、パフォーマンス向上につながる自己修正を探索
- **継続的改善**: 修正後のコーディング性能を自ら評価し、継続的に自身の性能を改善

実証された性能向上

DGMの実用性は、複数の権威あるベンチマークテストで実証されている[1]。AIエージェントのコーディング性能を測る「SWE-bench」において、DGMは20.0％から50.0％まで自動的に性能を高めることに成功した[1]。また、多言語のコーディング性能を評価する「Polyglot」では14.2％から30.7％まで性能を向上させ、これは人間が設計したAIエージェント「Aider」を超える性能を達成している[1]。



Sakana AI DGMのコーディングベンチマーク性能改善結果

安全性とセキュリティ対策

自ら自身の能力を高めるAIシステムにおいて、安全管理は極めて重要な課題となる[1] [8]。Sakana AIは、DGMの開発について「全ての自己修正と評価は、安全なサンドボックス環境内で、人間の監督の下、Webアクセスに厳格な制限を設けた上で行われた」と説明している[1]。加えて、DGMの全修正履歴は追跡可能としており、透明性と説明責任を確保している[1]。

OpenAI レベル4自律発明AIの概要と特徴

OpenAIのAI進化段階フレームワーク

OpenAIは、汎用人工知能（AGI）に向けた進歩を追跡するため、5つのレベルからなる分類システムを導入している[9] [10] [11]。このフレームワークは以下の段階で構成される[9] [11]：

- **レベル1**: 対話型AI（Chatbots）- 人間と対話可能なAI
- **レベル2**: 推論型AI（Reasoners）- 博士号レベルの問題解決能力
- **レベル3**: 自律型AI（Agents）- 人間の代わりに長期間タスクを実行
- **レベル4**: イノベーション型AI（Innovators）- 新たなアイデアや発明を創出
- **レベル5**: 組織型AI（Organizations）- 企業全体の業務を遂行可能

レベル4自律発明AIの定義と能力

レベル4のイノベーション型AIは、単に問題を解決するだけでなく、新しいアイデアや知識を創出し、創造的な問題解決や発明を行える段階のAIとして定義される[9] [10]。与えられた課題をこなすだけでなく、自ら研究開発を行い斬新なソリューションを生み出すことが期待されている[10]。これは、AIが蓄えた知識を統合して人間にはない発想を生み、新しい知見を人間と協力して作り出せるフェーズを意味する[10]。

実現時期予測と技術的展望

専門家の間では、レベル4自律発明AIの到達時期について様々な予測が示されている[12] [13]。楽観的な予測では2025年から2030年頃にかけてレベル4に突入し始めるとされ、慎重な予測では2030年代初頭までとする見方もある[12]。孫正義氏のAGI予測（2～3年以内）に基づけば、レベル4は現在から2年以内、つまり2027年前半までに実現する可能性も指摘されている[12]。

DGMとOpenAI レベル4の比較分析

技術的アプローチの相違点

両システムの技術的アプローチには明確な違いが存在する[1] [9] [10]。DGMは進化的アルゴリズムに基づく漸進的な自己改善を特徴とし、経験に基づく継続的な性能向上を目指している[1] [4]。一方、OpenAIのレベル4は、より突破的な発明やイノベーションの創出に焦点を当てており、創造的問題解決と革新的思考を重視している[9] [10]。

実装段階と開発状況

現在の開発状況を見ると、DGMは既に実用化段階に入っており、具体的な性能向上実績を示している[1]。対照的に、OpenAIのレベル4は理論的フレームワークとして提示されているものの、具体的な実装や性能データは公開されていない[9] [10]。この違いは、両者のアプローチが異なる開発段階にあることを示している[12] [13]。

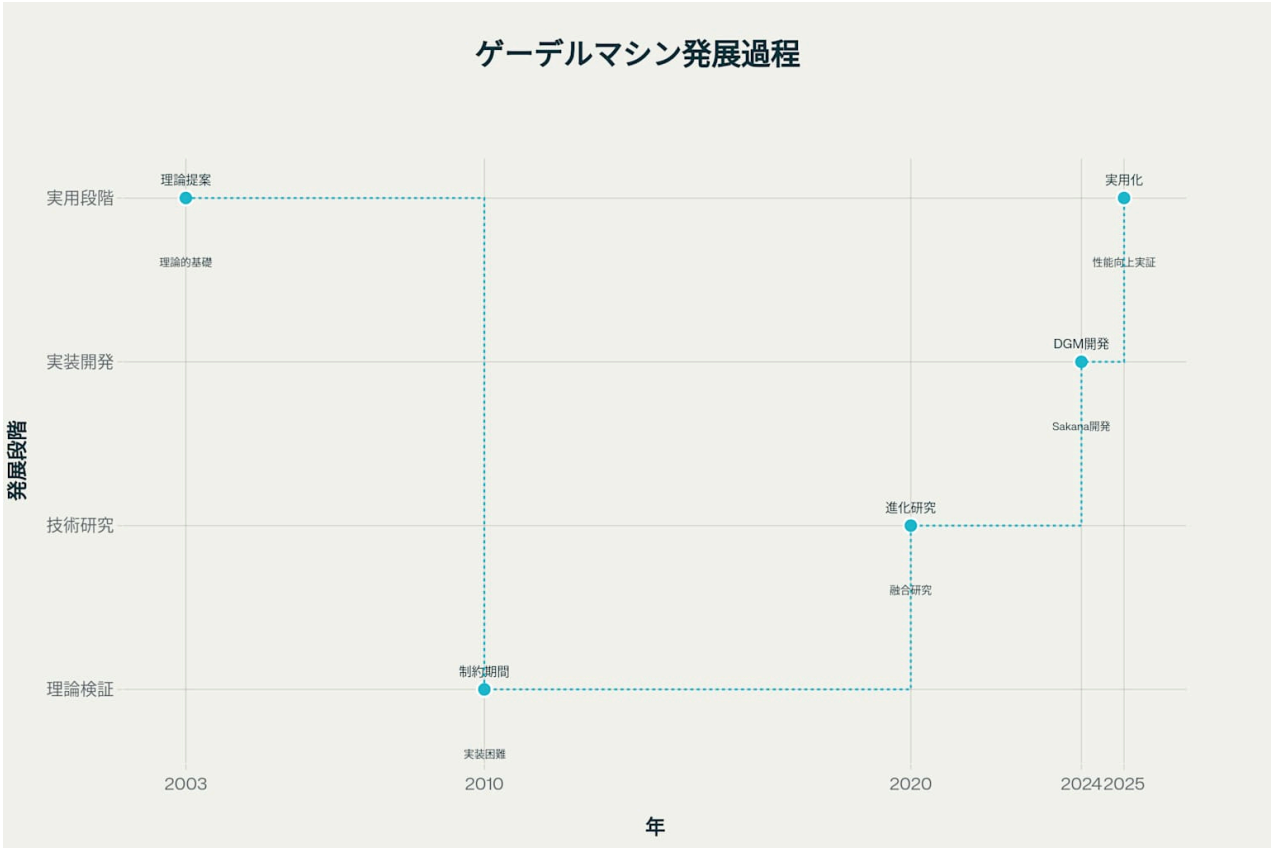
安全性管理と制御機構

安全性の観点では、DGMは明確な制御機構を実装している[1]。サンドボックス環境での実行、人間の監督、修正履歴の追跡など、具体的な安全対策が講じられている[1]。一方、OpenAIのレベル4については、安全性管理の詳細は現時点では明確に定義されていない[9] [10]。

技術発展の歴史的文脈と理論的基盤

ゲーデルマシンから実用化への道程

DGMの技術的基盤となるゲーデルマシンの概念は、2003年にユルゲン・シュミットフーバー氏によって提案された[4] [6] [14]。この理論は、クルト・ゲーデルの数学理論に基づき、AIが自身の性能を改善するという研究領域を確立した[4] [14]。しかし、理論上の存在にとどまっていたゲーデルマシンを実用的なシステムとして実装するまでには、約22年の研究開発期間を要した[4]。



ゲーデルマシン理論から実用化への22年間の発展過程

進化的アルゴリズムとオープンエンド探索

Jeff Clune教授の研究は、オープンエンドな進化的アルゴリズムの開発に焦点を当てている[5] [15] [16]。従来の進化的アルゴリズムは最初のうちは新しい種がいろいろ出てくるものの、ある程度世代が進んでいくと、それ以上に新しいものが生まれずに収束してしまうという課題があった[15]。DGMは、この問題を解決するためのオープンエンドな探索メカニズムを実装している[4] [5]。

メタ学習と自己改善の理論的発展

近年のAI研究では、「学習方法自体を学習する」メタ学習の分野が重要な概念となっている[4] [17]。LLMが自己反省を通じて「メタ認知」、つまり自分の思考プロセスを分析する能力を獲得することが実証されており[17]、これがDGMのような自己改善システムの理論的基盤を提供している[4]。

産業応用と社会的影響

コーディング支援の革新

DGMの実用化は、ソフトウェア開発分野に大きな変革をもたらす可能性がある[1] [18]。SWE-benchやPolyglotでの性能向上実績は、AIが人間のプログラマーを支援する能力が大幅に向上したことを示している[1]。特に、複数のプログラミング言語にわたる性能向上は、多言語開発環境での実用性を証明している[18]。

基盤モデル開発の自動化

Sakana AIは、DGM以外にも進化的アルゴリズムを用いた基盤モデル開発の自動化技術を公開している[19] [20] [21]。この技術により、既存のモデル同士をマージして新たな基盤モデルを構築する過程を自動化し、大規模モデルの訓練にかかる膨大なコストを削減することが可能になっている[20] [21]。

AI研究の自動化への発展

Sakana AIは「The AI Scientist」という、研究開発プロセスそのものを自動化する技術も開発している[22]。これは、アイデア創出、実験の実行と結果の要約、論文の執筆及びピアレビューといった科学研究のサイクルを自動的に遂行する新たなAIシステムであり[22]、DGMの技術的発展の延長線上に位置づけられる。

将来展望と技術的課題

AGI実現に向けたタイムライン

2025年の現時点で、多くの専門家がAGI（汎用人工知能）の実現が5-10年以内に可能であると予測している[23]。Google DeepMindのDemis Hassabis氏も、2025年4月にAGIが5-10年以内に実現する可能性を示唆している[23]。DGMのような自己改善型AIの発展は、このタイムラインを加速させる可能性がある[12] [13]。

技術的収束と相互影響

DGMとOpenAIのレベル4は、異なるアプローチを取りながらも、最終的には相互に補完し合う関係にあると考えられる[12] [13]。DGMの漸進的改善アプローチと、レベル4の突破的発明能力が組み合わされることで、より強力な自律発明AIシステムが実現される可能性がある[10]。

倫理的・社会的考慮事項

自己改善型AIの発展は、技術的成果と同時に重要な倫理的・社会的課題を提起している[24] [25]。AIが自律的に能力を向上させることの社会的影響、人間の創造性との関係、雇用への影響など、多面的な検討が必要である[24] [25]。特に、AI自身が新しい発明を生み出すレベル4段階では、知的財産権や創造性の帰属に関する新たな法的枠組みが必要になる可能性がある[10]。

結論

Sakana AIのダーウィン・ゲーデルマシンは、22年間理論にとどまっていたゲーデルマシンの概念を実用的なシステムとして実現した画期的な成果である[4]。進化的アルゴリズムに基づく自己改善メカニズムは、コーディング性能の大幅な向上を実証し、AI開発の新たなパラダイムを提示している[1]。一方、OpenAIのレベル4自律発明AIは、より野心的な目標として新しいアイデアや発明の創出を目指しており、DGMとは異なる技術的アプローチを取っている[9] [10]。

両システムは、自律的な問題解決能力と創造性の追求という共通の目標を持ちながら、実装方法と発展段階において明確な違いを示している[1] [9] [10]。DGMの実用的成果とOpenAIの理論的フレームワークは、相互補完的な関係にあり、将来的にはこれらのアプローチが統合され、より強力な自律発明AIシステムの実現につながる可能性が高い[12] [13]。

進化型AIの発展は、人工知能研究の新たな章を開いており、2025年以降のAI技術の発展において中心的な役割を果たすことが予想される[24] [23]。安全性と制御可能性を確保しながら、これらの技術をいかに社会に統合していくかが、今後の重要な課題となるであろう[1] [24] [25]。

⁂

1. https://www.itmedia.co.jp/aiplus/articles/2505/30/news130.html
2. https://note.com/redcord/n/nfea0cdf4f0d2
3. https://note.com/good_fairy858/n/nac374efc640f
4. https://sakana.ai/dgm-jp/
5. https://vectorinstitute.ai/new-vector-faculty-member-jeff-clunes-quest-to-create-open-ended-ai-systems/
6. https://pspeakers.com/ja/speaker/jurgen-schmidhuber/
7. https://www.lesswrong.com/w/gödel-machine
8. https://ai-compass.weeybrid.co.jp/securities/safe-testing-environment-sandbox/
9. https://quinteft.com/understanding-openais-five-levels-of-ai-progress-towards-agi/
10. https://note.com/joyous_echium468/n/n4b288202e58c
11. https://www.marktechpost.com/2024/07/13/5-levels-in-ai-by-openai-a-roadmap-to-human-level-problem-solving-capabilities/
12. https://yorozuipsc.com/blog/openai-o334
13. https://yorozuipsc.com/uploads/1/3/2/5/132566344/9c63e2ecf1b77892bc3b.pdf
14. https://en.wikipedia.org/wiki/Gödel_machine
15. https://note.com/alter_machine/n/n746e13c0bace
16. https://www.youtube.com/watch?v=05ZwhL_CxO8
17. https://note.com/ainest/n/n0a9c9c00e522
18. https://innovatopia.jp/ai/ai-news/52475/
19. https://recruit.gmo.jp/engineer/jisedai/blog/sakana-ai/
20. https://aismiley.co.jp/ai_news/sakana-ai-evolutionary-model-merge/
21. https://www.watch.impress.co.jp/docs/news/1577712.html
22. https://sakana.ai/ai-scientist-jp/
23. https://www.cognitivetoday.com/2025/04/artificial-general-intelligence-timeline-agi/
24. https://botpress.com/ja/blog/top-artificial-intelligence-trends
25. https://www.ai-souken.com/article/ai-generation-security-issues