# Anthropic「AI の自律的自己改善」警告：深層分析と今後の影響

## エグゼクティブサマリー

2026 年 6 月 4 日、Anthropic（クロードの開発元）は「When AI Builds Itself（AI が自分自身を構築するとき）」と題するブログを公開し、AI 開発史上最も重大な自己警告を発した。同社の共同創業者 Jack Clark と研究機関トップ Marina Favaro が執筆したこの文書は、AI システムが「再帰的自己改善（Recursive Self-Improvement）」――すなわち人間の介入なしに自らの後継モデルを設計・開発できる段階――に近づきつつあると警告する。その根拠として、2026 年 5 月時点で Anthropic の本番コードベースに統合されたコードの **80%以上が Claude によって記述されている**という内部データが示された。[1][2][3][4][5][6]

この提言は「即時停止の要求」ではなく、「停止する**選択肢**を持つべき」という条件付きの呼びかけである。しかし複数の大手メディアや専門家が指摘するように、IPO 直前（Anthropic は 6 月 1 日に機密で SEC 向け S-1 を提出したばかり）というタイミング、そして競争優位を確保しながら規制を呼びかけるという構図は、技術的な重大性と利益相反の懸念を同時に孕む複合的な事象である。[^7]

## 1. 警告の核心：「再帰的自己改善」とは何か

### 概念の定義と現在地

再帰的自己改善（Recursive Self-Improvement）とは、AI システムが自らの能力を改善し、その改善された能力でさらに次の改善を行う――というフィードバックループのことである。Anthropic のブログによれば、現在の AI 開発プロセスは以下の段階を経て自動化が進んでいる：[8][9][^10]

| 時期 | 段階 | 内容 |
| --- | --- | --- |
| 2021-2023 | 初期開発 | 人間がコードとドキュメントをすべて作成 |

| | | |
|---|---|---|
| 2023-2025 | チャットボット活用 | コードスニペットの生成補助 |
| 2025-2026 | コーディングエージェント | ファイル単位のコード自動作成 |
| 現在（2026 年） | 自律エージェント | コード実行・テスト・複数エージェント委任 |
| 20XX？ | ループの閉鎖 | モデル自身が後継モデルを構築・訓練 |

同社が示した内部証拠は極めて具体的である：[11][9]

- **コード生成比率**：2026 年 5 月時点で本番コードの 80%超が Claude 作成（2025 年 2 月の Claude Code 登場前は一桁台）
- **エンジニア生産性**：2024 年比でエンジニア 1 人当たりのコードマージ量が **8 倍**に増加
- **タスク完了時間の限界**：Claude Opus 3（2024 年 3 月）は約 4 分のタスク処理、Claude Opus 4.6 は **12 時間タスク**まで対応可能。この「限界時間」は 4 ヶ月ごとに倍増するペースで伸びており、2027 年には人間が数週間かかるタスクも AI がこなせる可能性がある
- **最難度タスクの成功率**：2025 年 11 月の Opus 4.5 では 51%だった「最善の次のステップ提案」が、2026 年 4 月の Mythos Preview では**64%**に向上
- **実験の自律実行**：AI 安全性に関する未解決問題を与えたところ、AI エージェントが仮説提案・実験設計・反復を 800 時間にわたって自律実行し、人間チームより高い成果を上げた

「現時点では再帰的自己改善には達していないが、不可避でもない。しかし多くの制度が準備できているよりも早く到来する可能性がある」と Anthropic は述べる。[3][9]

## 3 つの未来シナリオ

Anthropic のブログは、今後起こりうる 3 つのシナリオを明示している：[12][9]

1. **能力の頭打ち（最楽観シナリオ）**：現在の指数関数的な成長曲線が S 字カーブとして収束し、計算資源や電力・チップ供給がボトルネックとなる。AI の能力が今日のレベルで固定されても、社会・経済への影響は甚大。同社は「このシナリオは起きにくい」と見ている。

2. **複利的効率化の継続（中間シナリオ、最も可能性が高い）**：AI は自律的に研究課題を選択する能力には達しないが、開発作業の大部分を自動化。100 人の組織が 1 万〜10 万人規模の仕事をこなせるようになり、「研究のセンス（Research Taste）」を持つ人間の希少価値が急上昇する。

3. **完全な再帰的自己改善（最リスクシナリオ）**：AI がモデル設計・訓練・改良を自律的に担い、次世代 AI を生成し続ける。整合性（アライメント）問題が見えにくい形で複利的に悪化し、人間が制御を失う可能性がある。「今日のモデルに内在するわずかなミスアライメントが後継モデルを作るたびに増幅し、頻度が上がるが理解度が下がる」と警告する。[^9]

## 2. 提言の内容：「一時停止の選択肢」

### 条件付き停止論の構造

Anthropic の提言は**即時・一方的な開発停止の要求ではない**。Jack Clark は「アクセルがあるがブレーキがない状態」と現状を表現し、ブレーキ機構を今のうちに構築すべきだと訴えた。具体的な要件として同社が挙げる条件は以下の通りである：[2][5][^13]

- **多数の最前線研究機関の合意**：米国・中国を含む複数の国の複数の「十分なリソースを持つ研究機関」が同一条件のもと同時停止
- **検証可能性の担保**：核軍縮条約になぞらえた「実際に停止しているかを第三者が検証できる仕組み」の構築
- **トリガー・解除条件の明確化**：何が停止を発動させ、何が解除条件となり、誰が仲裁するかの明文化

「条件が整えば、他のフロンティア開発者が検証可能な形で停止した場合に限り、我々も減速または一時停止する」と同社は表明している。検証機構として、トレーニング実行の監視（ミサイルサイロより隠蔽しやすいという困難を認めつつ）、計算資源の追跡、プロベナンス認証などが検討されている。[5][11]

Anthropic Institute は今後数ヶ月内に、政策立案者・研究者・市民社会・他の AI 企業を招いた対話を実施し、その成果を公開すると発表した。また、国連では 2026 年 7 月 6〜7 日にジュネーブで**「AI ガバナンスに関するグローバル対話」**の第 1 回セッションが開催される予定であり、タイミングは偶然ではないかもしれない。[14][15][^9]

## 3. 証拠の重み：Claude Mythos と Project Glasswing

### セキュリティ分野での「制御不能リスク」の実証

この警告に現実的な重みを加えているのが、Anthropic が一般公開を見送った先端モデル「Claude Mythos」の存在である。同モデルはサイバーセキュリティ特化型であり、以下の実績を持つ：[16][17][^18]

- FreeBSD NFS に対する認証不要のルートアクセスエクスプロイトをわずか **4 時間で発見**
- 27 年前の OpenBSD のバグ、16 年前の FFmpeg の脆弱性（自動化ファジングツールが 500 万回見逃したもの）を発見
- Linux カーネルの権限昇格チェーンを自律的に発見・構築
- UK AI Security Institute 評価で「エキスパートレベルの CTF タスクで 73%の成功率」を達成（2025 年 4 月以前は不可能だったタスク）

この危険性を受け、Anthropic は Project Glasswing として約 40 の選定組織（Google、Microsoft、Apple、Amazon、JPMorgan など）にのみ Mythos を限定提供し、防衛的活用を先行させている。最初の数週間で **1 万件以上の高・重大脆弱性**が主要システムで発見され、すでに「脆弱性発見」がボトルネックではなく「パッチ適用速度」がボトルネックになったという。[17][9]

英国 AI セキュリティ機構（AISI）は Mythos Preview の評価で「脆弱なエンタープライズシステムへの自律的な多段攻撃が可能」と確認しており、これは AI 能力が「安全保障上の実在するリスク」に転じていることを国家機関が公認したことを意味する。[^18]

## 4. 批判と懐疑論：信頼性への疑問

### IPO 直前の「安全警告」という構図

Anthropic は 6 月 1 日に SEC への機密 S-1 提出（IPO 準備）を発表し、わずか 3 日後の 6 月 4 日に今回の警告を発した。評価額は最大 1 兆ドルとも報じられており、この時系列を巡って多くの批判が噴出している：[19][20][21][7]

- **NYU 教授 Gary Marcus**：「費用ゼロで完璧な IPO タイミングの修辞だ。『選択肢を持つべき』というだけで実際には停止する気がない」

- **元ホワイトハウス AI 顧問 David Sacks**：「ヌークになぞらえて危険を煽りながら、自分たちは全力疾走。結局、政府に自分たちを救ってほしいだけ」
- **Inworld AI CEO の Kylan Gibbs**：「自分たちが危険と叫べば規制設計を自社有利に誘導できる。オープンソース競合の規制と GPU 輸出制限が狙い」
- **LSE 教授 Luis Garicano**：「フロンティアモデルの最大の脅威はオープンウェイトモデル。全員を怖がらせれば、当然の流れとして『信頼できる開発者のみ』に絞る規制になる」
- **Johns Hopkins 大学教授 Francesco Bianchi**：「リスクは本物かもしれないが、市場リーダーが現状凍結を求めるのはあまりにも都合が良い」

一方で、**DeepMind CEO の Demis Hassabis が以前から全フロンティア開発者が合意した場合の一時停止を支持している**点も指摘されており、状況は単純な「自己利益説」で片付けられるものでもない。Anthropic の広報は「停止の呼びかけではなく、停止できる仕組みの研究・構築を求めているのだ」と説明している。[^21]

## 5. 地政学的障壁：米中競争と国際協調の難しさ

Anthropic が核軍縮条約を引き合いに出した国際協調論は、核の場合と本質的に異なる困難を持つ。[22][23]

- **不可視性**：AI のトレーニング実行はミサイルサイロと違い容易に隠蔽できる。核施設は物理的かつ痕跡が残るが、AI 訓練は汎用コンピュータ上で行われ外部からは見えにくい
- **民間主導**：核は国家独占だが、AI は民間企業が先端を走っており、政府規制の及ぶ範囲が限定的
- **中国との非対称性**：ワシントンとシリコンバレーでは「一方的な減速は中国に決定的な技術的優位を与える」との見方が根強い。Anthropic は自国政府から安全保障上のブラックリストに入れられた経緯もあり、米政府との関係は複雑である[24][14]
- **オープンソースモデル**：Alibaba の Qwen 等、中国系のオープンウェイトモデルはすでにエージェントコーディング能力で競合しており、フロンティアラボのみの停止合意では意味を持たない[^25]

核不拡散条約（NPT）が構築に**数十年**を要したのに対し、Anthropic は「そんな時間はない」と明言している。検証メカニズム（計算資源の追跡、コンピュートのプロベナンス認証）の研究はこれから始まる段階であり、現実的な国際協調が実現する可能性は現状では限定的である。[^9]

## 6. 知的財産への影響：IP 専門家が直視すべき課題

### AI による AI 発明の特許帰属問題の加速

Anthropic が実証した「AI が AI 開発コードの 80%以上を生成する」という状況は、特許法の根幹である**発明者の人間性要件**を直撃する問題を加速させる。[^26]

**現状の法的枠組みと矛盾**：2026 年時点で、主要法域いずれも AI を発明者として認めていない。スイス連邦行政裁判所の 2025 年 DABUS 判決も自然人による知的創造を要求しており、「人間の貢献がますます名目的になる」中でも法律は従来の発明者概念を維持している。[27][26]

再帰的自己改善が進んだ場合、以下の問題が具体化する：[28][26]

- **発明者帰属の空洞化**：AI が自律的に設計した後継 AI が生み出した技術革新を、誰が「発明した」と言えるのか
- **プロベナンス追跡の崩壊**：AI システムが自己改善しながら生み出した成果のデータガバナンスは、従来の「出所追跡」ロジックでは不可能になる[^28]
- **競争情報（CI）の速度格差**：AI が自律的にコード・研究実験・特許解析を行う環境では、AI を有効活用している企業とそうでない企業の間で特許出願・FTO 分析・ランドスケープ調査の速度・質に圧倒的な差が生じる
- **AI 生成発明の独自性評価**：再帰的自己改善 AI が生み出した技術の新規性・進歩性をどの基準で審査するかの議論が急務になる

日本特許庁はすでに「AI Action Plan 2022-2026」のもとで審査への AI 導入を進めている。しかし Anthropic が示した「研究判断（Research Taste）の自動化進展」が現実化すれば、審査側の AI 活用は AI 生成発明の爆発的増加に追いつけなくなる可能性がある。[^29]

### 特許戦略へのインプリケーション

| 影響領域 | 現在の状況 | 再帰的自己改善進展後 |
|---|---|---|
| 発明者帰属 | AI 支援発明の人間発明者要件を維持 | AI が「研究の主体」となった場合の帰属不在 |
| FTO 分析 | AI 補助でスピードアップ | 先行技術の増加速度が審査・分析能力を超過 |
| 特許出願数 | AI 活用で出願効率化 | 真の発明概念が希薄化し特許制度の機能変容 |
| 秘密管理 | 訓練データ・モデルのトレードシークレット | 自己改善 AI による模倣リスクの評価が困難 |
| ライセンス交渉 | 契約交渉に AI 活用 | 自律交渉エージェントの法的代理権問題 |

## 7. 今後の影響シナリオ：短期・中期・長期

### 短期（2026〜2027 年）

- **国際対話の開始**：Anthropic は今後数ヶ月内に政策担当者・研究者・AI 企業を招いた対話を実施すると表明。国連の「AI ガバナンスに関するグローバル対話」第 1 回セッション（2026 年 7 月 6〜7 日、ジュネーブ）との連動が注目される[15][14]
- **規制競争の激化**：Anthropic の提言を契機に、EU・米国・英国・日本が「フロンティアモデル規制」の定義と審査体制を急速に具体化する可能性
- **OpenAI・Google 等の対応**：OpenAI はすでに「政府のモデル審査強化」を求めており、両社の立場が「安全主義」で収斂するのか分岐するのかが競争環境を左右する[^30]

### 中期（2027〜2029 年）

- **AI によるコード生成 100%接近**：Jack Clark が「2 年以内の可能性」と示唆した AI コード 100%自動化が現実となれば、ソフトウェアエンジニアの役割が「コード生成者」から「研究方向性の設定者・レビュアー」へ完全移行[^6]
- **「研究のセンス」希少化**：どの問題を解くか・どの結果を信頼するか・どこが行き詰まりかを判断する「Research Taste」が人間の唯一の比較優位になる。IP 専門家でいえば「どの技術領域のランドスケープを重視するか」という戦略判断能力が最重要スキルに
- **国際検証メカニズムの試行**：Anthropic Institute による計算資源追跡・プロベナンス認証研究の成果が、国際協定の技術的基盤候補として提示される

### 長期（2030 年以降）

- **完全再帰的自己改善への分岐点**：Anthropic の「最悪シナリオ」が現実化した場合、現行の AI 安全評価体制（Model Cards、Responsible Scaling Policy 等）では対応不可能な自律的能力拡張が発生する可能性
- **特許制度の根本改革**：AI 自律発明の爆発的増加を受け、特許制度は「インセンティブ付与」機能を担保するための根本的な再設計（例：企業体発明者概念の拡張、AI 発明の登録制度化）が不可避になる
- **AI ガバナンス条約の模索**：核不拡散条約の構築に数十年かかったことを鑑みると、実効性ある国際 AI 協定には相当の時間を要するが、技術的加速がその「時間のなさ」を深刻化させる逆説が生じる

## 8. 総合評価：修辞か、真の警鐘か

Anthropic の提言は**自己矛盾的な誠実性**を持つ。同社自身がその加速を主導していることを内部データで実証しながら、その加速に警鐘を鳴らしている。IPO 直前という文脈、オープンソース規制への潜在的利益、規制設計への影響力獲得という動機が批評家に指摘されるのは正当である。[^21]

しかし同時に、**技術的な指摘の内実は軽視できない**。コードの 80%超が AI 生成であること、Claude Mythos Preview が何十年も見逃されてきた重大脆弱性を数時間で発見すること、エンジニアの生産性が 2 年で 8 倍になったこと——これらは競合他社も直面している現実であり、「怖い話」ではなく観測された事実である。[2][11][^18]

「自己利益的な警告でも正しい場合がある」——この命題がこの事象の核心にある。現時点では実効的な国際合意の可能性は低く、Anthropic Institute による検証メカニズム研究は「研究を始める」段階に過ぎない。しかし、フロンティア AI 企業が自社の内部データに基づき「人間の制御喪失リスク」を公式に認めたこと自体が、AI 安全をめぐる議論の重力中心を変えるターニングポイントとなりうる。[^5]

## References

1. Anthropic warns AI could soon help build its own successors - Axios - AI development is moving so rapidly that soon it will be able to advance itself without human involv...
2. Anthropic urges AI labs to pause development, warns humans risk ... - Anthropic says AI could soon improve without human intervention · Development pause will allow socie...
3. Anthropic warns AI may soon begin recursive self-improvement - Anthropic floated what it called “a global coordination mechanism” to slow or even pause AI developm...
4. Anthropic Warns of Self-Improving AI, Backs Frontier AI Pause as Claude Writes 80% of Company Code - Anthropic says AI is accelerating AI development and warns that self-improving systems may emerge so...
5. Anthropic Calls For Global Pause In AI Development - ... Favaro and Jack Clark argued the world should have the option to slow or temporarily pause front...
6. Anthropic co-founder warns AI could soon slip beyond our control - Anthropic co-founder Jack Clark said AI agents might soon be able to build and train models themselv...
7. Anthropic files confidential draft IPO paperwork as SEC review begins - A confidential draft S-1 filing allows Anthropic to begin the SEC review process without immediately...
8. One of America's leading AI companies is warning that artificial ... - Anthropic, the maker of Claude AI, says future AI systems could eventually reach a stage where they ...
9. When AI builds itself - Anthropic - Our progress toward recursive self-improvement, and its implications.
10. AI が自らを作るとき――再帰的自己改善に向けた私たちの進展とその ... - AI による再帰的自己改善とは、AI が AI 自身の能力を改善し、その改善がさらに次の改善を促すという仕組みを指します。これは、技術的シンギュラリティを実現 ...
11. Anthropic urges global coordination to pause frontier AI development - AI company Anthropic calls for verifiable multilateral pause in frontier AI development, warning of ...
12. Johan Falk's Post - LinkedIn - Important read from Marina Favaro (Anthropic Institute) and Jack Clark (Anthropic co-founder). It's ...
13. Anthropic's Call for A.I. Nonproliferation - The New York Times - Several critics have argued that Anthropic has made fear-mongering a marketing strategy, though indu...

14. Anthropic urges AI labs to pause, warns humans risk losing control - It warned that rapid advances in technology could soon allow AI systems to improve themselves faster...

15. Global Dialogue on AI Governance - the United Nations - ... 2026 AI Dialogue. The first session of the Global Dialogue on AI Governance will be held on 6 an...

16. アンソロピック社、開発減速提言 - 【ニューヨーク共同】人工知能（ＡＩ）開発の米アンソロピックは６日までに、ＡＩが人間の手を借りずに性能を自律的に高める段階に近づいたと指摘し、暴走して制御不能になる前に「開発を遅らせたり、一時停止したり...

17. Anthropic Claude Mythos and the 2026 Cybersecurity Landscape - The emergence of Anthropic's Claude Mythos model could mark a pivotal shift in the cybersecurity lan...

18. Our evaluation of Claude Mythos Preview's cyber capabilities - We conducted cyber evaluations of Anthropic's Claude Mythos Preview and found continued improvement ...

19. Anthropic Urges Global Pause in AI Development, Flags 'Self ... - WSJ - The $1 trillion startup warns artificial-intelligence models are nearing capability to improve witho...

20. Anthropic Files Confidential S-1: Joins $3 Trillion AI IPO Race - Anthropic has reportedly filed a confidential S-1 with the SEC, thrusting the Claude AI maker into a...

21. Smart People Weigh in on Anthropic's AI Pause Proposal - "We believe it would be good for the world to have the option to slow or temporarily pause," two lea...

22. AI Regulation: Reflections on the Nuclear Analogy and its Utility - In 2023, the Centre for AI Safety issued a panel statement that placed the threat of AI on par with ...

23. Nuclear Non-Proliferation Is the Wrong Framework for AI Governance - Placing AI in a nuclear framework inflates expectations and distracts from practical, sector-specifi...

24. Anthropic calls for pause of global AI development - Artificial intelligence company Anthropic suggested Thursday a global pause on building the most pow...

25. Claude maker Anthropic is calling for a global pause in AI ... - Anthropic's chief scientist Jared Kaplan warns that by 2027-2030, humanity must decide whether to al...

26. [PDF] The Invented Inventor: Adapting Patent Law to Generative AI - As artificial intelligence increasingly drives processes of discovery, the concept of inventor – a k...

27. Who Owns AI Generated Inventions and Content in 2026? - As of 2026, no major jurisdiction recognizes artificial intelligence systems as inventors or authors...

28. Several Issues Regarding Data Governance in AGI - arXiv - This paper examines data governance challenges specific to AGI, defined as systems capable of recurs...

29. The Japan Patent Office fully introduces AI - How will examinations ... - A patent attorney explains the current status of the Japan Patent Office's AI Action Plan (2022-2026...

30. アンソロピック、「AI 開発減速」提言 OpenAI は政府規制強化を訴え - 米新興アンソロピックは 4 日、人工知能（AI）の暴走リスクを抑えるには開発の一時停止や減速が有効だと提言した。競合の米オープン AI も政府の監視強化 ...