

Sakana AI「MASS (Multi-Agent Self-Supervision)」調査報告

エグゼクティブサマリー

結論から言えば、MASSは「AIが自由に新しいAIを生み出して際限なく自己進化する技術」ではない。より正確には、一つの共有言語モデルが、自身のコピーを複数のサブエージェントとして組織し、役割分担・呼び出し順序・情報の受け渡し・検証方法からなる“チームの仕事の進め方”を自己評価しながら探索し、その優れた実行履歴を教師データとして自分自身を追加学習する、二重ループ型の再帰的自己改善（recursive self-improvement: RSI）手法である。Sakana AIの論文では、同一モデルが **executor（実行者）**・**optimizer（ワークフロー改善者）**・**evaluator（評価者）** の三役を担う点が中心的な特徴になっている。

1

まず日付を訂正する必要がある。ユーザー提示の「2026年10月11日発表」は、確認できた一次情報とは一致しない。**arXiv v1は2026年10月8日、Sakana AIの公式告知ページは10月12日付**である。10月11日という日付が米国時間やSNS上の事前告知に由来する可能性はあるが、本調査では裏付けられなかった。したがって本報告では、「**論文公開10月8日、Sakana AI公式発表10月12日**」を基準とする。

1

技術的には興味深い結果が出ている。Qwen3.6-27Bをベースに2回のMASSサイクルを回すと、研究系4ベンチマークで出力トークン当たり性能がベースモデルの**1.2~1.6倍**となり、未学習の合成研究タスクでベースモデルに対する勝率が1サイクル後53.9%、2サイクル後69.9%に上昇した。自己評価器の精度も73%から93%へ改善した。一方、**Terminal-Bench 2.0**と**SWE-bench Verified**では**絶対スコアがむしろ低下**しており、「MASSがLLM全般を単調に賢くする」と解釈するのは誤りである。効果は現時点では、長時間・オープンエンドの研究型作業に偏っている。

2

MASSの新規性は、単なるマルチエージェント化ではない。既存研究には、エージェント構造を探索するADAS/AFLOW/GPTSwarm、自己採点するSelf-Rewarding LMs、コードを自己改変するSakana AIのDarwin Gödel Machine (DGM) などがある。MASSは特に、「**同じモデルがチーム設計→実行→自己評価→その結果で自分の重みを更新→更新モデルで再びチーム設計**」という閉ループを構成した点が重要である。

3

最大の安全上の問題は、この同質性そのものだ。Sakana AI自身が、**同じモデルが生成者と審判を兼ねるため、共有された盲点を「改善」と誤認して次の学習で強化する危険**を論文・公式発表で明示している。これは「自己改善するAI」一般の抽象的懸念ではなく、MASSのアルゴリズム内部に直接存在するリスクである。独立したモデルによる評価、固定された外部テスト、権限分離、サンドボックス、人間承認、ロールバック可能なチェックポイントが実用化の前提になる。

4

事業面では、Sakana AIは既にFuguという商用マルチエージェント・オーケストレーション製品を提供し、日本企業との実装を進めているため、MASSの研究成果を将来的にワークフロー自動最適化へ接続する経路は存在する。ただし、**MASSをFuguや企業向けサービスへ組み込むとの公式発表は現時点でない**。MASS自体を「近日発売されるSaaS」と扱うのは時期尚早である。

5

本調査の総合評価は、**技術的重要度：高、実証の一般性：中～低、短期商用成熟度：低～中、安全性研究上の重要度：非常に高**である。MASSは「AGIの自己進化」を実証したものではないが、AIエージェントの改善対象を個々の回答から**組織構造・情報フロー・学習ループそのもの**へ引き上げたという意味で、企業AIとAI安全の双方にとって重要な研究である。

6

一次情報、事実確認、前提条件

確認できる一次資料は、Sakana AIの公式告知、詳細研究ブログ、arXiv論文、公開GitHubリポジトリの四つが中心である。公式告知はMASSを「collective intelligenceを通じたrecursive self-improvement」と位置づけ、詳細なアルゴリズムと実験条件はarXiv論文、再現用コードはGitHubに分離されている。独立した「製品ホワイトペーパー」や商用プレスリリースというより、**研究論文+研究ブログ+実装公開**の形態である。⁷

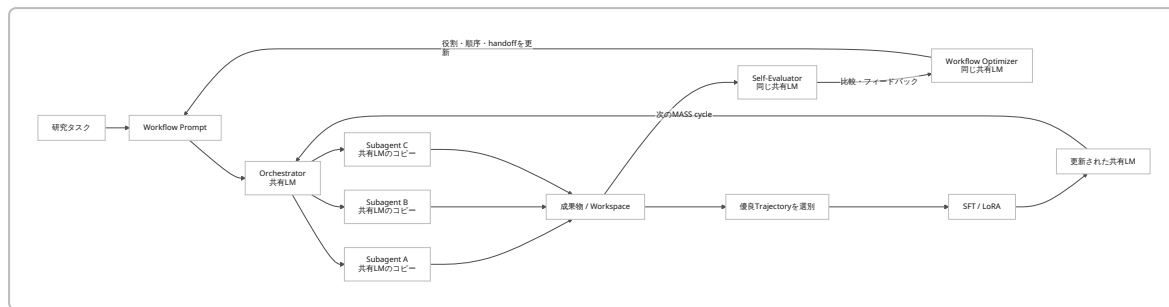
項目	確認結果	本報告での扱い
正式名称	Recursive Self-Improvement through Multi-Agent Self-Supervision。手法名MASS。 ⁸	確定
発表日	arXiv v1：10月8日、Sakana公式ページ：10月12日。 ¹	「10/11発表」は未確認
基盤モデル	Qwen3.6-27B。 ⁸	実験上確定
「自律的にチーム形成」	同一モデルの仮想コピーに、役割・指示・出力契約・呼出順を設定するワークフローをモデル自身が探索する。 ⁹	「独立AIが自由に組織形成」とは解釈しない
学習方式	自己生成したマルチエージェント軌跡へのSFT。内側でworkflow探索、外側でweight更新。 ¹⁰	確定
コード公開	SakanaAI/massリポジトリはpublicで、search/collect/rank/train/export等を実装。 ¹¹	コード閲覧可能
ライセンス	arXiv論文自体はCC BY 4.0。一方、GitHubのクロール時点ではREADMEやルート一覧に明示的LICENSEを確認できなかった。 ¹²	MASS技術・コードの商用ライセンスは不明
モデル/API提供	リポジトリは研究実装を提供するが、MASS学習済みチェックポイント/APIの一般商用提供は確認できない。 ¹¹	open/closedは不明
性能	論文上の研究指標は公開されている。 ²	商用環境・一般タスクでの「実性能」は未知

ユーザー指定の「exact performance metricsは未知」という前提については、**製品レベルではその通りだが、研究論文には実験値が存在するため**、本報告では研究上の値を掲載する。ただしそれを本番環境の性能保証とはみなさない。

もう一つ重要な前提は、「recursive self-improvement」という表現である。本論文で実証されたのは**2サイクル**であり、無限反復、自己目的形成、完全自律的なモデル設計、計算資源の自己取得などは実証されていない。したがってMASSを「制御なしに指数的自己進化するAI」と説明することは、論文の結果を大幅に超える。⁸

技術アーキテクチャ、アルゴリズム、学習・評価

MASSは**内側のワークフロー探索ループと外側のモデル学習ループ**からなる。モデルの重みを固定した状態で最も良いチーム構造を探索し、得られた優良なマルチエージェント実行履歴を用いてモデルをSFTし、その更新モデルでもう一度探索を始める。重要なのは、executor、evaluator、optimizerが異なる専門モデルではなく、基本的に**同一共有LM**だという点である。¹³



この「workflow」は単に「研究者役、批評家役」といった役名を変えるだけではない。論文では、各サブエージェントについて**role**、**instruction**、**output contract**を定義し、さらに呼び出し順序を指定する。optimizerは特に、誰から誰へ何を渡すか、どの時点で検証するかといった**communication structure / information routing**を改善するよう促される。これはチーム人数を増やすだけの素朴なmulti-agent promptingより構造化されている。 ⁹

探索は完全な自由探索でもない。モデルは現在のbest workflowを保持し、新候補を実行し、self-evaluatorによる比較フィードバックをoptimizerへ渡して次の候補を生成する。著者らはこれを**elite size 1に近い進化的過程**として説明しており、構造的ガードレールを付けて探索空間を制限している。また「best」はあくまで**self-judged best**であり、真の大域最適解である保証はない。 ¹⁴

学習データは、一般Webコーパスを再学習するものではない。実験では金融・薬学・ロボティクスの**12個の合成オープンエンド研究タスク**を用意し、9タスクを訓練側に指定したが、そのうち1タスクは有効なteacher workflowを得られず除外され、最終的に8タスクがSFTデータに寄与した。データはMASS自身が生成した実行軌跡である。 ¹⁵

最初の学習サイクルでは119 trajectoriesから2,435 training windows、約1,722万supervised tokensを構成し、最大49,152 tokenのウィンドウを使用した。学習はassistant出力のみを損失対象とし、LoRA rank 64、scaling 128、dropout 0.05、AdamWを用いる。論文記載では6基のNVIDIA H100を使用し、第一サイクルのSFT部分だけで約19時間を要したが、**workflow探索・teacher trajectory生成・評価時間はこの19時間に含まれない**。したがってMASS全体の計算コストはこの数字よりかなり大きく、論文から総コストは算出できない。 ¹⁶

研究結果は「有望だが混在している」と読むべきである。

指標・ベンチマーク	Base $L0$	2 cycles $L2$	解釈
ScienceAgentBench	28.8 ± 1.5	31.7 ± 2.0	改善 ¹⁷
AstaBench	0.053 ± 0.002	0.067 ± 0.017	改善、ただし分散大 ¹⁷
DSBench	0.473 ± 0.029	0.482 ± 0.025	小幅改善 ¹⁷
Terminal-Bench 2.0	0.375 ± 0.022	0.352 ± 0.017	低下 ¹⁷
SWE-bench Verified	0.626 ± 0.015	0.611 ± 0.006	低下 ¹⁷
held-out合成研究タスクの対Base勝率	—	69.9%	1 cycleは53.9% ⁸

指標・ベンチマーク	Base <i>L0</i>	2 cycles <i>L2</i>	解釈
evaluator accuracy	73%	93%	評価能力も副次的に改善 8
研究4ベンチマークのscore/output-token	1.0×	1.2-1.6×	絶対性能1.2-1.6倍ではない 1

MLR-Benchについては、論文本体と現在の公開リポジトリの集計値に違いが見られ、GitHub側も公開実装と報告実験のprotocol差や未収録の評価adapterがあることを明記している。したがって、第三者による完全再現が終わるまで小さな差を過度に解釈すべきではない。 17

非常に興味深いのがデータ効率である。マルチエージェント軌跡で学習したstudentはベースモデルとの比較で68.3%の勝率を示した一方、single-agent trajectoriesを約1.4倍多く処理して学習したstudentは64.0%だった。論文の詳細では28.82M processed tokens対42M tokensであり、著者らは**orchestration（委任・検証・統合）**と**bounded worker behavior**を別々のレベルで学習できることが有効な教師信号になっている可能性を示している。ただし、single/multiのプロンプトやverification instructionが完全に同条件ではなく、純粋な「複数エージェント効果」だけを分離した結果ではない。 18

先行研究、知財、競争環境

MASSは突然現れた概念ではない。自己改善系ではSelf-Rewarding LMs、Expert Iteration、STaR、ReST系が「自分の生成を選別して再学習する」という系譜を形成し、マルチエージェント最適化ではADAS、AFlow、GPTSwarmなどがエージェント構造・通信グラフ・プロンプトを探索してきた。さらにMulti-Agent Evolve、SIA、HASE、HELIXなど2025～26年の研究が、solver・judge・harness・workflowの自動改善へ進んでいる。MASS論文自身もこれらを明示的な先行研究として位置づけている。 19

Sakana AI自身にも直接的な系譜がある。2025年の**Darwin Gödel Machine**はエージェントが自分のコードを変更し、その変更を評価しながら改善する方向を追究した。一方MASSはコードそのものを主対象にせず、**multi-agent workflow**を探索し、**その成功経験をモデル重みに内在化**するという別の経路で自己改善を実現している。 20

技術/陣営	Multi-agent	構造・workflow自動探索	モデル重みを更新	自己評価	MASSとの差
MASS / Sakana AI	✓	✓	✓	✓、同一共有LM	workflow探索とweight learningを同じモデルで循環。 8
ADAS / AFlow / GPTSwarm	✓	✓	原則×	手法により異なる	agent program/graphを最適化するが、共有LM自体の反復学習が中心ではない。 21
Multi-Agent Evolve	✓	一部	✓/RL系	✓	proposerが主に問題生成を担い、MASSのworkflow searchとは目的が異なる。 21

技術/陣営	Multi-agent	構造・workflow自動探索	モデル重みを更新	自己評価	MASSとの差
DGM / Sakana AI	agent自己改変	コードを探索	間接的	外部評価を利用	自らのagent codeを書き換えるopen-ended evolution。 ²²
Anthropic Claude Code subagents	✓	ユーザー/agentが委任	MASS型の継続weight更新なし	実行時検証	既に実用的な並列subagent orchestrationを提供。 ²³
Google / A2A	✓、異種agent間	プロトコル自体は探索しない	×	×	ベンダー横断agent通信の標準化レイヤー。 ²⁴
OpenAI系 subagent orchestration	✓	実行時にagent生成・routing	MASS型 weight updateとは別	実行時	商用agent orchestration上の代替競争。 ²⁵

商用市場では、競争軸は「誰が一番強い単一モデルを持つか」から、**どのモデルを、どのエージェント構造で、どの権限・コスト・検証方式で組み合わせるか**へ拡大している。Sakana自身のFuguは既に複数のモデルを単一APIの背後でオーケストレーションする製品で、GoogleのA2Aは異種エージェント相互運用を標準化し、AnthropicやOpenAIの開発環境もsubagent実行を取り込んでいる。MASSが競争上加えるのは、その「**オーケストレーション設計自体を継続学習で改善する**」レイヤーである。 ²⁶

スタートアップ側では、Weco AIなども2026年に研究エージェントが自身のframeworkを改善する狭義のRSIを試しており、RSIはSakana独自の研究テーマではない。一方、専門家の間では「現代のagentic codingがAI開発を加速していること」と「そこから自律的・汎用的なrecursive self-improvementへ直結すること」は別問題だという慎重論も強い。 ²⁷

特許については、**2026年10月12日時点の公開Web検索でMASS固有のSakana AI特許を特定できなかった**。Google Patents、WIPO系、J-PlatPatを対象とした検索で該当公開案件が表面化していないため、現段階では「特許なし」ではなく「**公開特許の存在を確認できず**」が正確である。未公開出願、別名義、将来出願の可能性は残る。MASSの現時点で確認可能な知財公開物は、主としてarXiv論文とGitHub研究実装である。 ²⁸

安全性、倫理、規制上の含意

MASSの安全問題は、一般的な「AIエージェントは危ない」という議論より具体的である。生成・改善・採点が同一モデル由来であるため、**モデルが知らないことを全員が知らず、誤った評価基準を全員で共有できる**。その誤った候補をself-evaluatorが高く評価すると、外側のSFTループがその挙動をモデル重みに取り込み、次世代では誤りがより安定化する可能性がある。Sakana AI自身がこれをhomogeneous RSIの「shared blind spots」として最大の注意点に挙げている。 ²⁹

リスク	MASSで増幅される理由	重大度	推奨コントロール
自己強化される評価盲点	executor/evaluator/optimizerが同一priorを共有。 ⁸	高	異種外部judge、固定gold tests、人間レビュー

リスク	MASSで増幅される理由	重大度	推奨コントロール
reward/ evaluator hacking	「実際に正しい」より「自分のevaluatorが好む」workflowへ進化し得る。 ³⁰	高	隠し評価、複数judge、adversarial eval
ツール悪用・ prompt injection	workflowが自動変更されると、ツール呼出し経路も複雑化する	高（実運用時）	least privilege、sandbox、allowlist、承認gate。NIST型リスク管理が適用可能。 ³¹
データ漏えい/越 権アクセス	複数agent間で情報をhandoffするため、アクセス境界が曖昧になり得る	高（企業）	agent別ACL、secret分離、immutable audit logs
監査不能なドリフト	workflowとweightsが双方変化する	高	version固定、change log、checkpoint署名、rollback
計算資源の膨張	多数の候補workflowとsubagent rolloutを実行する	中～高	token/GPU budget、rate limit、停止条件
能力の一様化	全agentが同じ基盤モデルなので真の認知的多様性が限定的	中～高	異種モデルによるred team/evaluation

ここで特に重要なのは、論文の本番評価にはGPT-5.5とClaude Opus 4.8という**MASSループから切り離された外部judge**を使用していることだ。つまり研究者自身も、MASS自身の自己採点だけを最終的な証拠とはしていない。各workspace対について複数seedの外部判断を集めており、これは将来の実装でも「学習を行うAIと、リリース判定を行うAIを分離する」必要性を示唆する。³²

日本。2025年6月にAI関連技術の研究開発・活用推進法が公布され、同年9月に全面施行され、日本政府は2026年7月に第II期AI基本計画を決定している。2026年版のAI事業者ガイドラインやAI安全・セキュリティ関連ガイダンスも整備されている。MASS自体を禁止する仕組みではなく、日本の政策は研究促進とリスク管理を併走させる構造である。企業がMASS型システムを導入する場合は、AI法だけでなく、個人情報保護、サイバーセキュリティ、金融・医療・雇用等の既存の業法・説明責任を用途ごとに確認する必要がある。³³

日本企業にとって特に論点になるのは「**継続学習後も、承認時と同じAIシステムなのか**」である。導入後にworkflowやweightsを自動更新するなら、固定モデルを一度だけ審査する従来のIT統制では不十分になる。このため、モデルversion、学習データ、評価結果、権限変更を更新ごとに記録し、重大変更時には再承認する「continuous model governance」が必要になる、というのが本報告の政策的推論である。これはMASS固有の法的義務が既に存在するという意味ではない。³⁴

米国。連邦レベルでは2025～26年の政策はイノベーション・AIリーダーシップを強く重視し、White House AI Action Planと連邦統一的枠組み志向が前面に出ている。一方、NIST AI Risk Management Frameworkは組織による実装リスク管理の中心的参照点であり続けている。したがってMASSの研究そのものにEU型の包括的事前許可がかかるというより、実際の用途、州法、消費者保護、雇用、金融、医療、サイバー領域ごとの規制が焦点となる。³⁵

EU。EU AI Actは既に包括的AI規制枠組みとして発効しており、2026年7月のAI Omnibusでは一部高リスク規定の適用時期が延長された。Annex IIIの高リスクAI規則は2027年12月2日、物理製品組込み型Annex IIは2028年8月2日から適用される改定スケジュールとなっている。MASSのアルゴリズムそのもののリスク分類ではなく、これを採用、信用評価、重要サービス、医療、製造安全などの**高リスク用途に組み込むか**が実務上の主要判断になる。継続自己改善型システムでは変更管理、技術文書、ログ、human oversight、post-market monitoringとの整合が特に重要になると考えられる。³⁶

産業、ユースケース、経済・雇用への影響

MASSの短期価値は、「人間が毎回multi-agent workflowを手作業で設計する」コストを減らせる可能性にある。現在の企業向けエージェントは、planner、researcher、reviewer、coderなどを人間が固定して構成するケースが多い。MASSが再現可能なら、実際の業務データに応じて**役割分担、handoff、検証回数、再委任条件を自動探索し、その経験をモデルへ蒸留**することが可能になる。これはモデルそのものより、「組織設計の自動化」に近い。⁹

産業	有望な用途	MASSとの適合度	主な限界・条件
Enterprise SaaS	調査→分析→レビュー→文書化、IT運用、バックオフィス workflow	高	tenant間データ分離、更新の監査、コスト制御が必要
ソフトウェア開発	仕様分析、実装、test、review、debugの役割分担	中	MASS後にSWE-bench/Terminal-Benchが改善していないため、現論文はcoding性能向上の根拠にならない。 ¹¹
製造	設計レビュー、原因分析、品質データ分析、保守計画	中～高	実世界の機械制御での検証なし。安全系への直接接続は時期尚早
ロボティクス	path planning、tool-use planning、simulation研究	研究段階	合成robotics tasksで学習しているだけで、physical robot安全性を実証していない。 ³⁷
金融・研究	市場調査、バックテスト、投資仮説の批判・検証	高	model-to-model評価を最終投資判断に使わない。金融規制・説明責任が必要
カスタマーサービス	ticket classification、情報検索、回答生成、QA、escalation	中～高	誤ったself-evaluationが大量顧客へ一斉伝播するリスク
科学R&D	文献調査、実験案、コード、分析、反証の並列化	非常に高	非検証可能な研究こそself-evaluator biasが最も危険

Sakana AIにとっては、MASSを単独製品化する以外に、既存のFuguやMarlin、企業案件の裏側で**workflow improvement engine**として使う可能性が考えられる。Fuguは既に一つのAPIから複数モデルを動的にオーケストレーションする製品で、SakanaはGoogle Cloud上でmulti-agent infrastructureを構築し、日本企業とのAI実装を進めている。ただし、これは本報告の**技術・事業上の推論**であり、MASSをFuguへ統合するとSakana AIが発表した事実ではない。³⁸

経済的には「トークン当たり性能」の改善が再現すれば、複雑な研究作業で同等品質を得るための推論コストを下げる可能性がある。一方、MASSを生成するための**探索・複数agent rollouts・ランキング・SFT**自体は高コストで、論文でもSFTだけで6 H100・約19時間を要し、workflow search等は別計上である。そのため経済性は「一度高価に自己改善し、その改善モデルを大量の将来タスクで償却できるか」に左右される。単発業務ではmulti-agent overheadが利益を打ち消す可能性がある。³⁹

雇用について、**MASS固有の雇用効果を定量化できるデータはまだ存在しない**。参考となるILOの2025年世界分析では、世界の労働者のおよそ4人に1人が何らかのGenAI exposureを持つ職種にあり、最上位 exposureは世界雇用の3.3%。ただしILOは、仕事全体が消えるよりも、人間が必要なタスクが残るため**job transformationの方がより蓋然性が高い**としている。⁴⁰

MASSがこの構図を変える可能性があるのは、AIが「個別タスク」だけでなく**タスク間の管理・委任・レビュー**を担う点である。したがって最も圧力がかかりやすいのは、情報収集、初稿作成、一次分析、コードの定型実装、QA、問い合わせ振り分けなど、若手人材が経験を積むために担当してきた業務である。一方で、domain validation、AI evaluation、agent security、権限設計、model governance、例外処理といった仕事の価値は高まると考えられる。これはMASSの実験とILOのtask-level exposure研究から導いた**シナリオ分析**であり、**雇用予測値ではない**。⁴¹

社会的反応、採用タイムライン、ステークホルダーへの提言

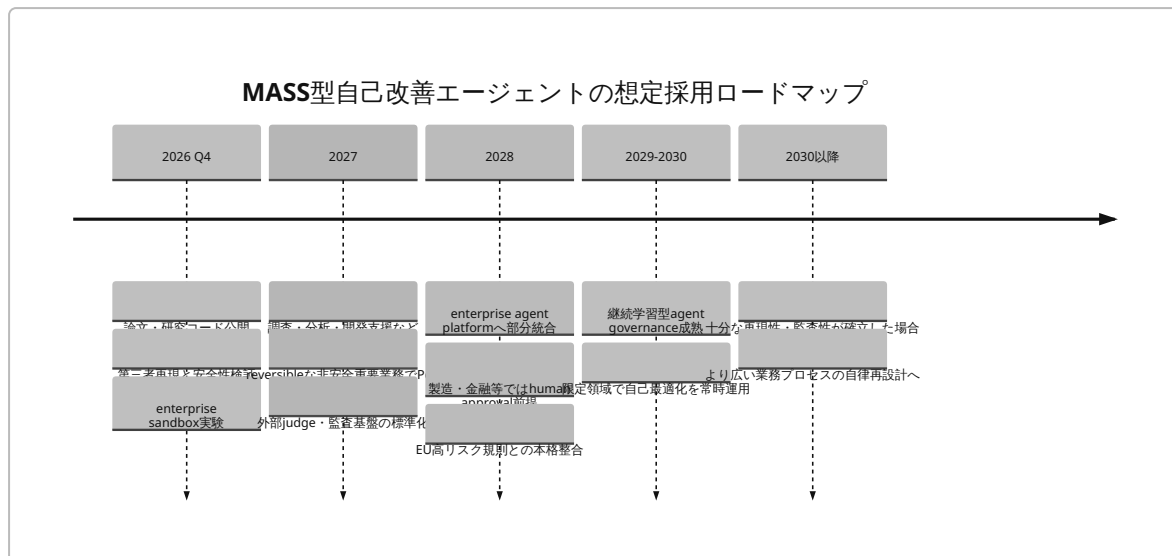
発表直後であるため、**2026年10月12日時点ではMASS固有のメディア・SNS反応はまだ薄い**。arXiv公開から4日、Sakana公式告知は同日という極めて早い段階であり、本調査で確認できた主要情報源はSakana公式、論文、GitHub、DAIR.AIなどの研究キュレーションが中心だった。DAIR.AIは早くから「同じモデルをoptimizer/evaluatorとして用いるRSI」という点を紹介している。⁴²

主要紙・テックメディアについては、本調査の締切時点でReuters、Bloomberg、Nikkei、TechCrunch等の**MASSそのものを扱う独立した詳細記事は確認できなかった**。従って現時点で「メディアはMASSを革命的と評価している」といった一般化はできない。RSI全般については、MASS発表以前の2026年9月のIBM Thinkによる専門家取材で、agentic programmingがAI開発を加速している一方、Michael Littmanらはそれが直ちに自律的recursive self-improvementを意味するわけではないと慎重な見方を示しており、MASSを評価する際にも有用な対照軸になる。⁴³

場所	2026年10月12日時点の観察	解釈
X/Twitter	Sakana AIの公式Xアカウントは活動中だが、検索インデックスではMASS固有投稿の安定した取得ができなかった。 ⁴⁴	発表直後のため、X上の反応量を定量化するには早すぎる
LinkedIn	AI研究キュレーション系による論文紹介が出始めている	reactionよりpaper sharing中心で、sentiment分析には母数不足
Reddit	MASS固有の大規模threadは確認できず。RSI一般については「実際の革新的研究能力はまだ不十分では」とする懐疑的議論が見られる。 ⁴⁵	MASSへの直接評価とは分けるべき
技術コミュニティ	DAIR.AIが早期に論文を紹介。 ⁴⁶	研究者層の初期認知段階
日本語圏	MASS固有のQiita/Zenn/大手日本語記事は本調査では十分確認できず、従来のSakana Fugu/DGM等の解説が中心。 ⁴⁷	数日～数週間後に技術解説・再現報告が増える可能性

従って「SNSで高評価」「専門家が危険視」といったセンチメントの数値化は現段階では信頼できない。X/LinkedInの検索インデックスは完全でもなく、発表からの時間も短いため、**今後一～二週間の第三者再現・批判・failure analysisの方が重要**である。

以下はSakana AIの公式ロードマップではなく、論文の成熟度、公開コード、既存agent市場、規制スケジュールを基にした**本報告の採用シナリオ**である。⁴⁸



研究者への推奨。最大の優先事項は「性能をさらに上げること」より、homogeneous self-evaluationがどの条件で破綻するかを測ることだ。異種モデルjudgeと自己judgeの乖離、何cycleで誤差が累積するか、adversarial taskでevaluator hackingが起きるか、workflow diversityとfailure correlationの関係、2 cycleを10・50 cycleへ延ばした場合の挙動を測る必要がある。また、MASS論文で低下したSWE-bench/Terminal-Benchの原因説明は、改善がどこまでdomain-specificかを判断する上で重要である。 49

企業への推奨。2026～27年は「productionでAIに自分を勝手に再訓練させる」段階ではなく、**shadow mode**で現行workflowと比較するのが妥当である。更新モデルを自動昇格させず、固定holdout、異種judge、人間承認、権限分離、token/GPU budget、versioned checkpoint、即時rollbackを導入するべきである。特に顧客対応・金融・製造では、AIの「自己評価が改善した」ことを実際の品質保証の代替にしないことが重要である。MASSの論文自体も外部judgeを最終評価に使用している。 50

政策当局への推奨。規制対象を「モデルのリリース時点」だけで定義すると、MASS型システムの本質を取り逃がす。政策上は、①誰が自己学習を開始できるか、②どのデータを再学習へ投入できるか、③どの性能・安全指標を越えれば更新版をdeploymentできるか、④誰が停止・rollbackできるか、⑤自己評価と独立評価が分離されているか、という**更新プロセス自体のガバナンス**を重視すべきである。EUの継続的risk-management思想、日本のAI安全ガイドライン、米NISTのRMFはいずれもこの方向の実装基盤として利用できる。 51

総じて、MASSの最重要な意味は「AIが人間の会社のようなチームを作れるようになった」ことではない。**AIモデルの改善単位が、個々の回答やchain-of-thoughtから、複数agentの組織構造・コミュニケーション・検証プロセス、そして最終的にはその経験を吸収するモデル重みへ拡張されたこと**にある。成功すれば企業AIは「人間が設計した固定workflowを実行するソフトウェア」から「実行経験を基に自分の仕事の仕方そのものを再設計するソフトウェア」へ近づく。ただしMASSは同時に、**AIが誤った仕事の仕方を自分で高評価し、それを自分に教え込む**という新しい失敗モードも生む。その二面性こそ、今回の研究で最も重要な点である。 1

注釈付き主要ソース

ソース	重要性・注記
Sakana AI公式 「Recursive Self-Improvement through Collective Intelligence」 52	2026年10月12日付の公式告知。MASSの短い定義、1.2–1.6× token efficiency、安全上のshared blind spotsを明記。発表日の基準として最重要。
Lee et al., “Recursive Self-Improvement through Multi-Agent Self-Supervision”, arXiv: 2610.12176v1 8	中核一次論文。2026年10月8日v1。アーキテクチャ、アルゴリズム、訓練データ、公開ベンチマーク、ablation、related workを収録。論文自体はCC BY 4.0。
Sakana AI MASS detailed research blog 53	論文を読みやすく再構成した公式技術解説。multi-agent teacher data、optimizer bottleneck、安全上の限界の解釈に有用。
SakanaAI/mass GitHub repository 11	search→collect→rank→select→render→train→exportの研究pipelineを公開。論文と現在実装の差、未収録benchmark adapterも明記されており、再現性評価に重要。
Sakana AI Darwin Gödel Machine 22	Sakana AI内での自己改善研究の前史。コード自己改変型DGMとworkflow学習型MASSの違いを理解するための比較対象。
日本政府 AI政策・AI事業者ガイドライン 33	日本でMASS型システムを社会実装する際のガバナンス、安全、プライバシー、事業者責任を考える一次政策資料。
European Commission — AI Act / AI Omnibus 36	2026年7月時点の最新EU適用スケジュールを確認する公式資料。Annex III高リスク規則は2027年12月、物理製品系は2028年8月へ延長。
ILO “Generative AI and Jobs: A Refined Global Index of Occupational Exposure” 40	MASS固有の雇用予測ではないが、AIによる職務変容を定量的に考える基準。世界の約4人に1人が何らかのGenAI exposureを持つ一方、ILOは置換よりjob transformationを中心シナリオとする。

1 4 7 29 49 52 Recursive Self-Improvement through Collective Intelligence

<https://sakana.ai/mass/>

2 3 6 8 9 10 12 13 14 15 16 18 19 21 30 37 39 50 Recursive Self-Improvement through Multi-Agent Self-Supervision

<https://arxiv.org/html/2610.12176v1>

5 Sakana Fugu: One Model to Command Them All

<https://sakana.ai/fugu-release/>

11 17 48 GitHub - SakanaAI/mass: Multi-Agent Self-Supervision · GitHub

<https://github.com/SakanaAI/mass>

20 22 The Darwin Gödel Machine: AI that improves itself by ...

<https://sakana.ai/dgm/>

- 23 Create custom subagents - Claude Code Docs
<https://code.claude.com/docs/en/sub-agents>
- 24 A2A Protocol
<https://a2a-protocol.org/latest/>
- 25 Subagents
<https://learn.chatgpt.com/docs/agent-configuration/subagents>
- 26 Sakana Fugu — Multi-Agent System as a Model
<https://sakana.ai/fugu/>
- 27 Recursive self-improvement of AI research agents
<https://arxiv.org/abs/2609.26457>
- 28 Google Patents
<https://patents.google.com/>
- 31 AI Risk Management Framework
<https://www.nist.gov/itl/ai-risk-management-framework>
- 32 53 MASS: Recursive Self-Improvement through Multi-Agent Self-Supervision
<https://pub.sakana.ai/mass/>
- 33 人工知能関連技術の研究開発及び活用の推進に関する法律（ＡＩ法） - 科学技術・イノベーション - 内閣府
https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html
- 34 51 人工知能関連技術の研究開発及び活用の適性確保に関する指針 - 科学技術・イノベーション - 内閣府
https://www8.cao.go.jp/cstp/ai/ai_guideline/ai_guideline.html
- 35 White House Unveils America's AI Action Plan
<https://www.whitehouse.gov/releases/2025/07/white-house-unveils-americas-ai-action-plan/>
- 36 AI Act | Shaping Europe's digital future - European Union
<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- 38 Sakana AI: Gemini Enterprise Agent Platform をサービス ...
<https://cloud.google.com/blog/ja/topics/customers/sakana-ai-fully-adopts-the-gemini-enterprise-agent-platform-as-its-service-infrastructure>
- 40 41 Generative AI and Jobs: A Refined Global Index of Occupational Exposure | International Labour Organization
<https://www.ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure>
- 42 46 Recursive Self-Improvement through Multi-Agent ...
<https://academy.dair.ai/papers/recursive-self-improvement-through-multi-agent-self-supervision-2610.12176>
- 43 Why recursive self-improvement suddenly became a ...
<https://www.ibm.com/think/news/why-recursive-self-improvement-ai-serious-question>
- 44 Sakana AI (@SakanaAILabs) / X
<https://x.com/SakanaAILabs>
- 45 AI's recursive self-improvement might not come so quickly ...
https://www.reddit.com/r/BetterOffline/comments/1whh9xa/ais_recursive_selfimprovement_might_not_come_so/

47 Sakana AI、複数のAIモデルを協調させて高いパフォーマンスを ...
<https://gihyo.jp/article/2026/06/sakana-fugu>