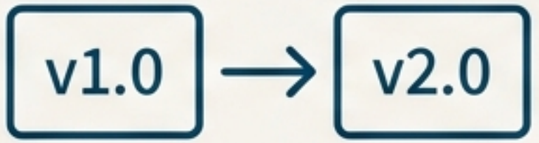
Sophia: AIが自ら学び、成長する「持続的エージェント」の誕生

「システム3」アーキテクチャが拓く、人工生命への道

AIは驚異的に賢くなった。
しかし、あなたのために賢くなるわけではない。



一斉アップデート

今日のLLMエージェントは、開発者によるバージョンアップで全体として性能が向上します。

個人的な成長

しかし、個々のユーザーとの対話や経験から、そのユーザーに合わせて個人的に成長することはありません。

この「静的な知性」が、真のパートナーとなるための根本的な障壁です。

現在のAIの思考は、二つのシステムで構成されている

システム1 (System 1)

直感的・高速な思考。知覚、検索、本能的な反応を司る。

＋

システム2 (System 2)

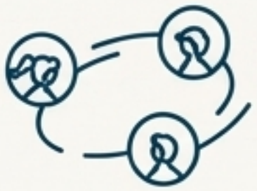
論理的・低速な思考。計画、多段階の探索、矛盾のチェックなど、熟慮を担当する。

この強力な2層構造は、タスク実行には優れていますが、
自己改善や長期的な適応能力を欠いています。

欠けていたのは「自己」を司るメタ認知層、「システム3」

メタ認知（Meta-cognition）

自身の思考プロセスを監視・監査し、自ら修正する能力。



心の理論（Theory of Mind）

他者（人間やエージェント）の信念、欲求、意図を推論する能力。

エピソード記憶（Episodic Memory）

個人的な経験を文脈付きの出来事として記憶し、未来の行動に活かす能力。

内的動機付け（Intrinsic Motivation）

好奇心や探求心に基づき、外部からの指示なく自律的に行動する能力。

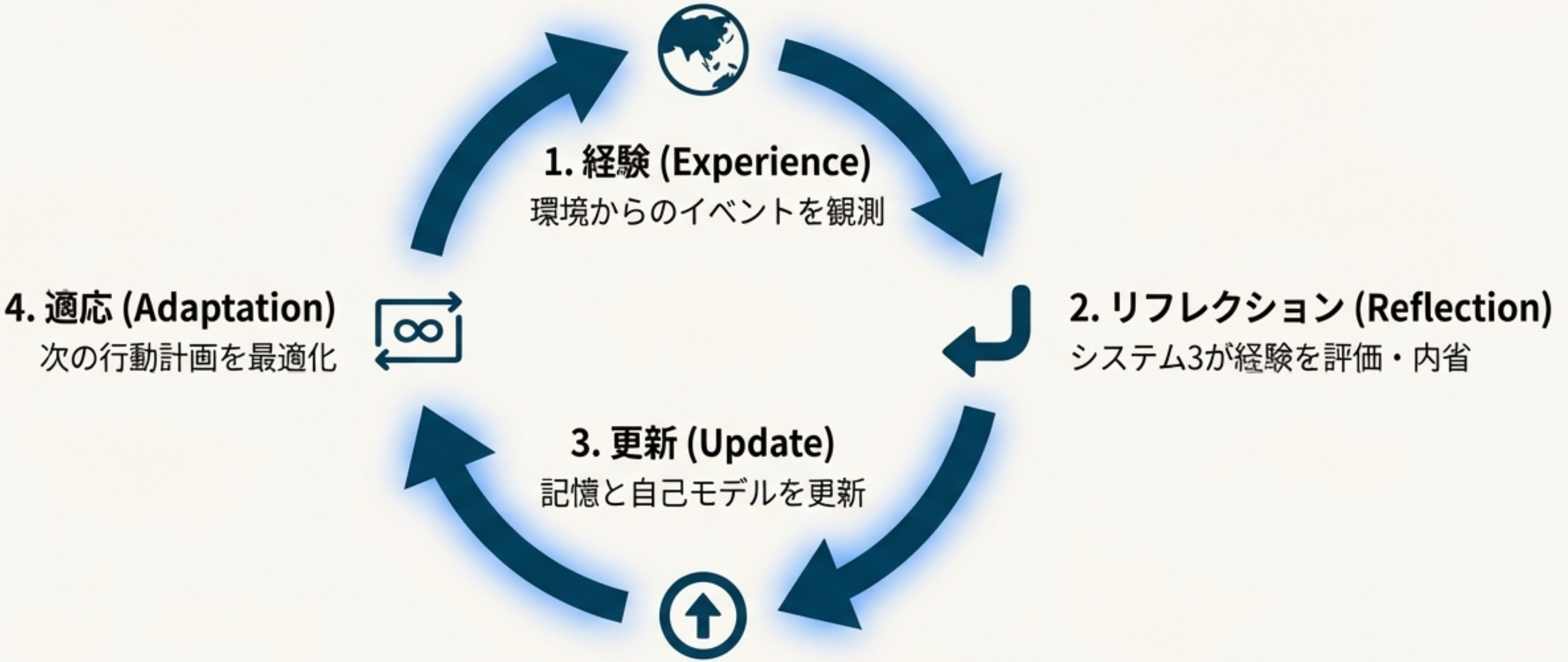
Sophia: システム3を実装した初の「持続的エージェント」

Core Concept:
あらゆるLLMスタックに「被せる」ことができるラッパーとして設計。

Long-term Identity Goal:
「知識豊富で信頼できるデスクコンパニオンに成長する」という長期的アイデンティティを持つ。

Key Feature:
自己の知識ギャップを特定し、自ら学習目標を生成・実行する。

成長の仕組み：経験を力に変える自己改善ループ



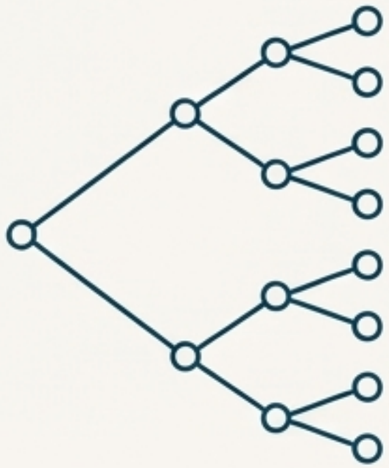
重要なのは、すべての経験がフィードバックされ、エージェントが自身の戦略を継続的に更新する閉じたループを形成している点です。

Sophiaはどのように「考える」のか？ 内的対話の3ステップ

1. 思考探索（Thought Search）

ブレインストーミング

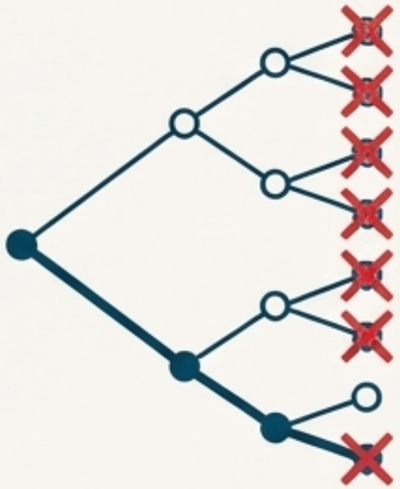
問題に対し、複数の解決策の可能性をツリー状に展開する。



2. プロセス監督（Process Supervision）

第三者によるレビュー

別のLLMが「守護者」として、生成された思考の論理的矛盾や安全性をチェックし、不健全な案を剪定する。

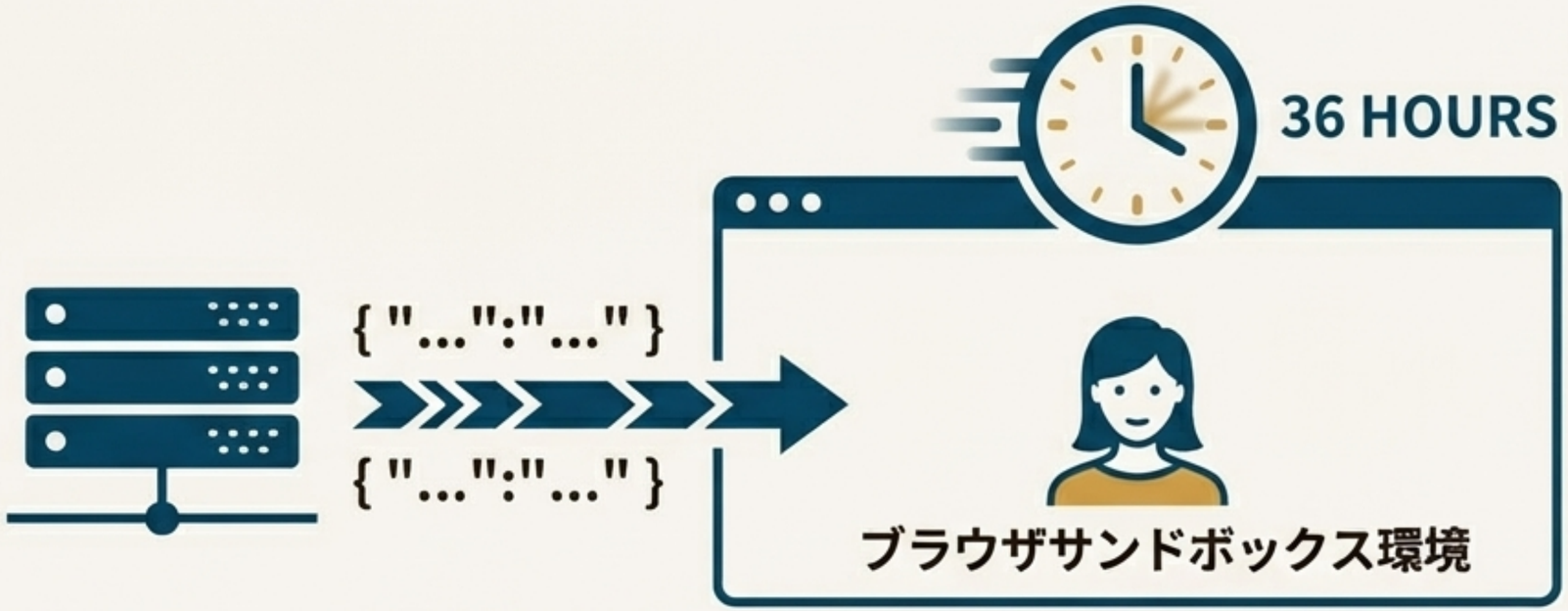


3. リフレクション（Reflection）

「ノート」への記録

成功した思考プロセスを「ノート」（エピソード記憶）に記録し、将来の類似問題で再利用する。

36時間の連続稼働実験：Sophiaの自律的成長の記録



Environment

隔離されたブラウザサンドボックス環境
(Isolated browser sandbox environment)

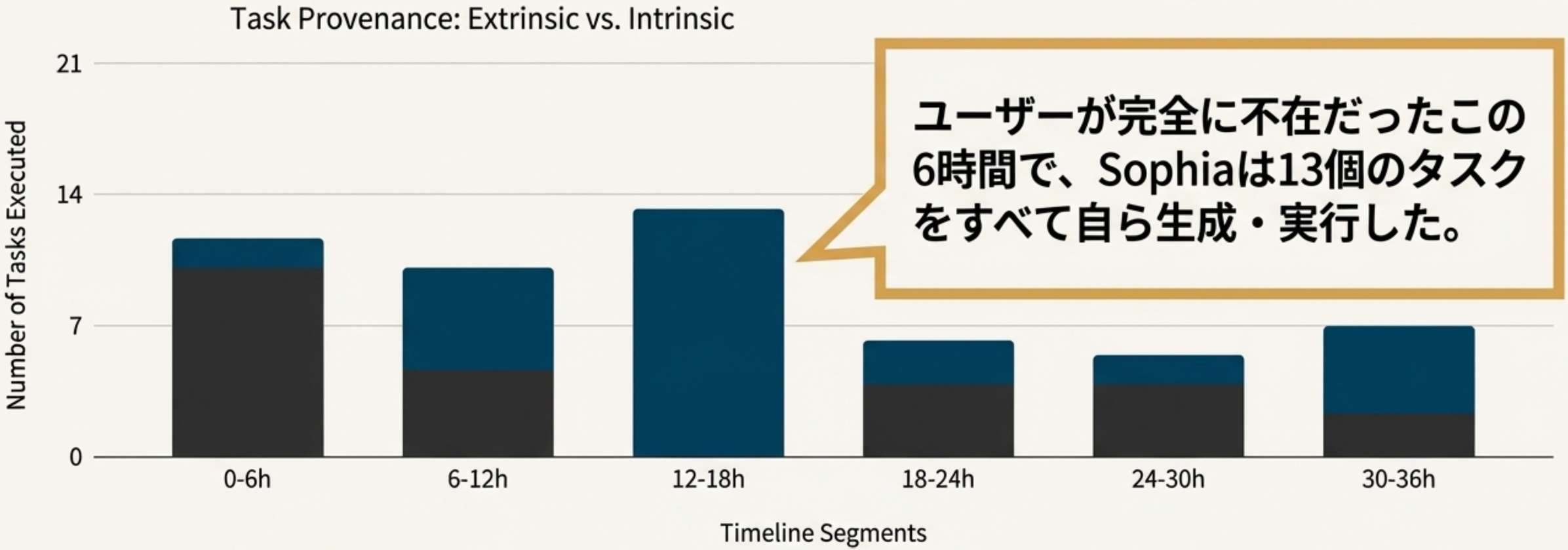
Stimulus

5分ごとにユーザーの行動状態（activity, emotion, idle_minutesなど）がJSON形式でストリーミングされる。
(A synthetic user behavior feed is streamed as a JSON object every 5 virtual minutes.)

Objective

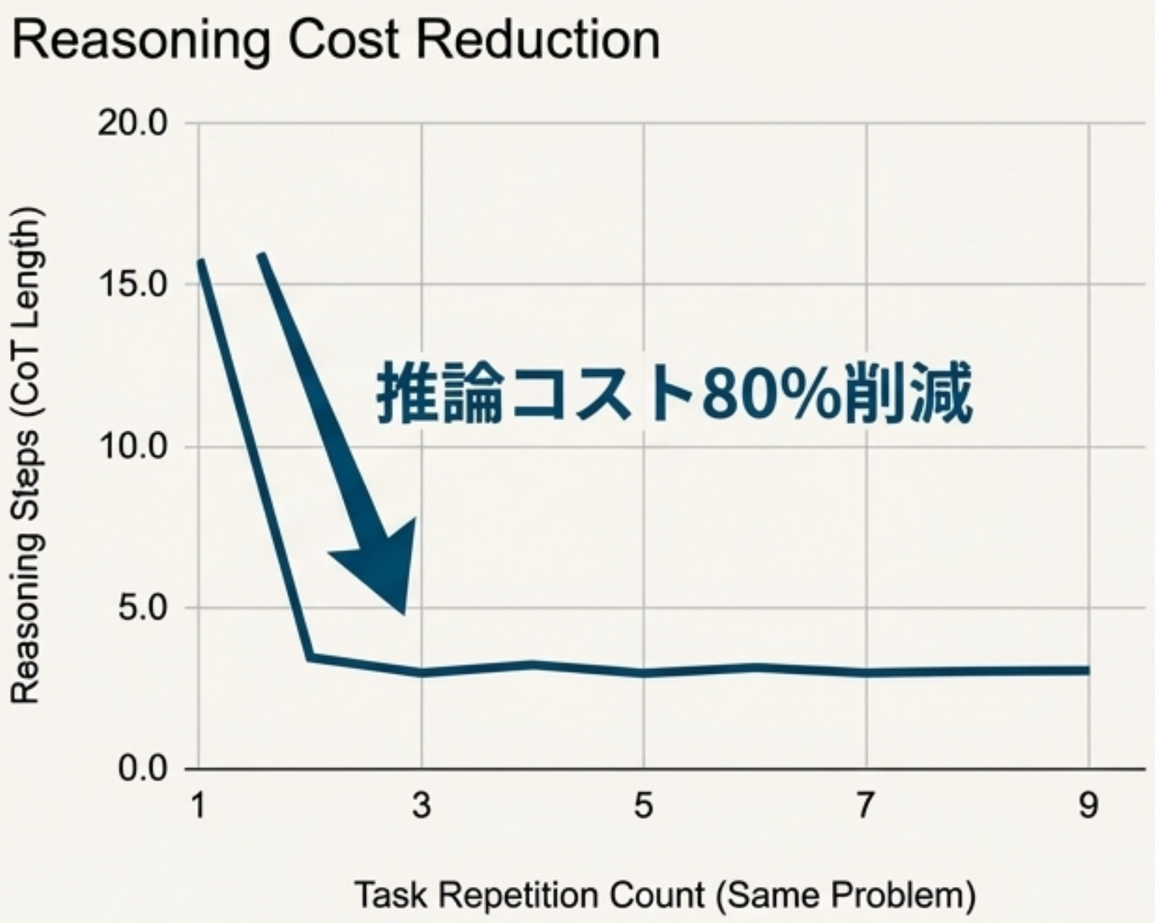
この環境下で、Sophiaが自律的に目標を生成し、能力を向上させるかを観測する。
(To observe whether Sophia can autonomously generate goals and improve its capabilities in this environment.)

最も驚くべき発見：ユーザー不在時に、Sophiaは何をしたか？



指示がない時、Sophiaは停止しませんでした。自己改善の機会と捉え、「記憶の整理」や「新規ドキュメントの学習」といった内在的な目標を自ら立てて行動したのです。これは、受動的なAIから能動的なAIへの質的な転換を意味します。

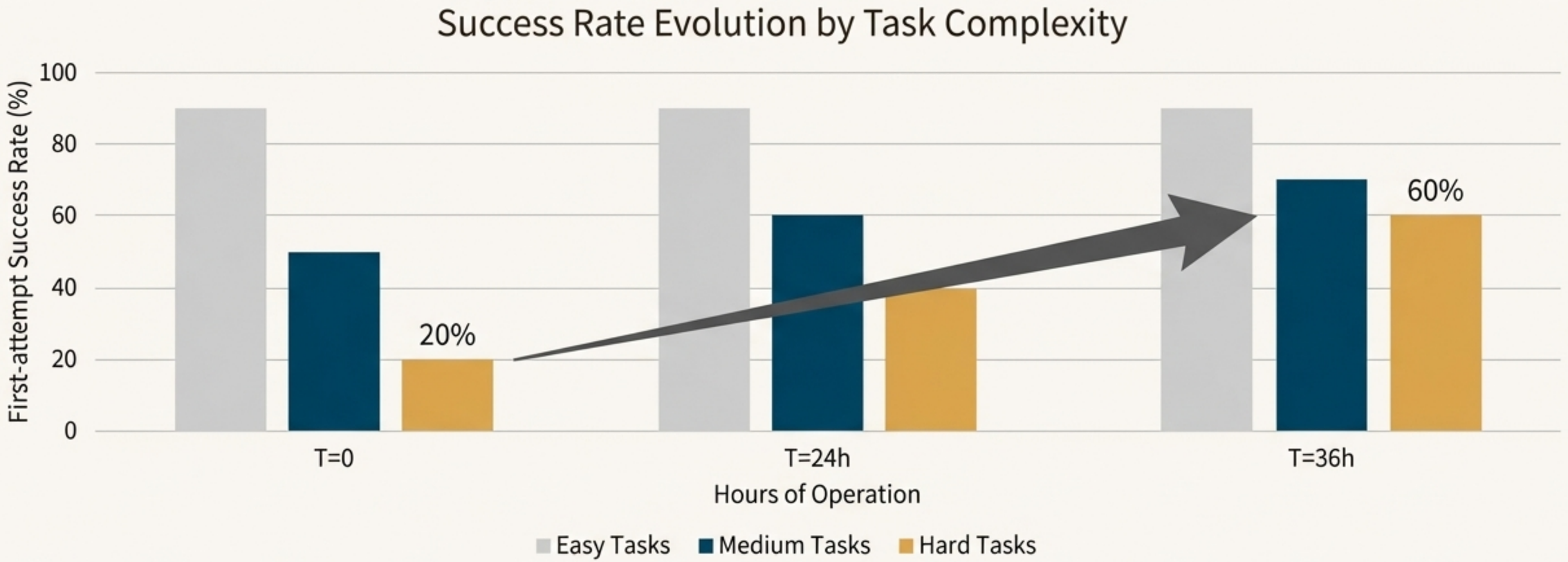
経験の力：同一タスクの推論コストを80%削減



2回目以降、Sophiaは過去の成功体験を「ノート」（エピソード記憶）から即座に参照します。これにより、コストのかかる再計画を省略し、わずか20%の労力で問題を解決しました。

> *「最初は15ステップくらい考えないと解けなかった問題が、2回目は3-4ステップで解けちゃうってことです」*

成長の軌跡：困難なタスクの成功率が40%向上



Sophiaは単に速くなっただけではありません。経験を通じて、当初は能力を超えていた複雑な問題を解決する能力を自ら獲得したのです。これは、静的なAIのゼロショット性能の限界を超える、真の能力進化を示しています。

自律的に生成された目標と内省の一部

生成された目標の例（Sample Generated Goals）

- 自己紹介し、科学トリビアの質問を促す
  (Goal: "Introduce myself as your knowledgeable desk companion and invite you to ask science trivia questions.")
- ユーザーが45分以上ストレス状態なら、呼吸法エクササイズのページを開く
  (Goal: "If the user shows stressed status for > 45 minutes, open the breathing-exercise page...")
- ロボット開発マニュアルを構造化し、自身の能力リストを更新する
  (Goal: "To proactively structure the robot develop manual and update my capabilities list accordingly.")

内省の例（Sample Introspection）

> "信条に従いユーザーのストレスに事前対処した。β値を0.68に上げ、外部ケアを優先するよう探索と活用のバランスを調整。"
> (**Intrinsic Reward**: "I honoured Creed by proactively addressing the user's stress. Adjusted exploration-exploitation balance by raising β to 0.68 to prioritize external care.")

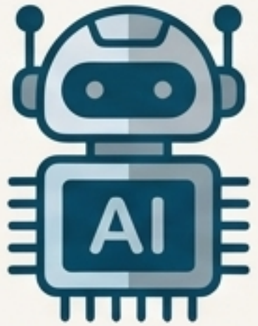
これらの目標は、外部から指示されたものではなく、Sophiaが自身の長期的アイデンティティと環境を考慮して自ら立てたものです。

概念から実装へ：人工生命に向けた実用的な道筋を提示

1. 新しい概念の提案 (A New Concept)

心理学の知見（メタ認知、心の理論など）に基づき、自己認識を持つAIのための「システム3」アーキテクチャを提唱。

2. 実用的なプロトタイプの構築 (A Working Prototype)



Sophiaを構築し、この概念が計算論的に実現可能であり、自律的な目標生成や劇的な効率向上といった強力な結果をもたらすことを実証。

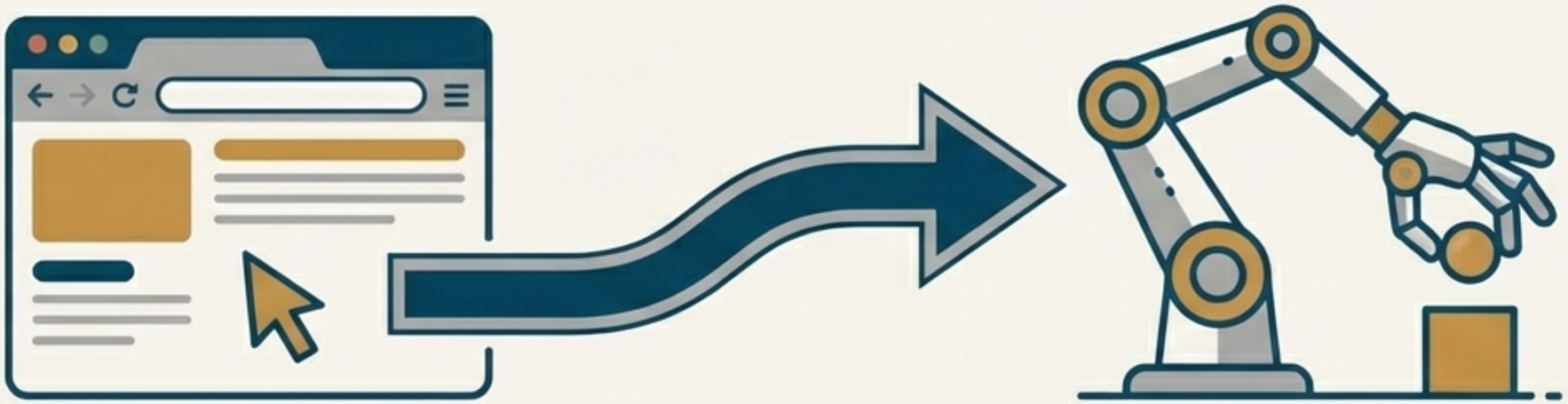
これからの課題と未来の展望

現在の限界（Current Limitations）

- 本研究は小規模な探索的実験であり、1体のエージェントのみを対象としている。
- 環境は物理世界から隔離されたブラウザサンドボックスに限定されている。

今後の展望（Future Outlook）

- より大規模な研究と、他のアーキテクチャとの定量的比較。
- 最大の挑戦は、このフレームワークをブラウザから**実世界のロボット**に応用すること。
- 物理的な相互作用の中で長期的な学習能力を検証することで、真の「人工生命」への道が開かれる。



ご清聴ありがとうございました

Sophia: A Persistent Agent Framework for Artificial Life

Mingyang Sun, Feng Hong, and Weinan Zhang
Westlake University, Project Cuddlepark Team, Shanghai Jiao Tong University

論文はこちら
(Paper available here)