AI Agents: From Concept to Code to Commerce

— A Survey Bridging AAMAS and LLMs

Toru Ishida[1,2], Yohei Murakami[3], Donghui Lin[4] and Kemas Muslim Lhaksmana[1]

[1]School of Computing, Telkom University, Bandung, Indonesia
[2]Professor Emeritus, Kyoto University, Kyoto, Japan
[3]Faculty of Information Science and Engineering, Ritsumeikan University, Ibaraki, Japan
[4]Faculty of Environmental, Life, Natural Science and Technology, Okayama University, Okayama, Japan

Abstract— 数十年にわたり、自律エージェントとマルチエージェントシステム（AAMAS）は自律性と協調の基礎理論を発展させてきた。大規模言語モデル（LLM）が AI エージェントへの関心を再燃させたが、確立されたフレームワークとの接続は限定的である。本サーベイは、AAMAS の理論的基盤（Concept）、その LLM ベースの実装（Code）、および社会的展開（Commerce）を調査することで、AI エージェントの現在の軌跡を理解するための統合的なフレームワークを提供する。重要なのは、AAMAS の厳密な理論的裏付けと、LLM の前例のない自然言語処理および常識推論能力という相乗的な組み合わせが、それぞれの限界を克服し、現実世界に適用可能な AI エージェントへの道を開いていくであろうことである。本サーベイは、LLM 研究者が達成した AI エージェントの画期的な実用的進歩と、確立された AAMAS 理論との間の橋渡しを意図し、将来の進歩を促進する。[1]

Keywords—AI agent, autonomous agent, multiagent system, large language model, generative AI, LLM, AAMAS

1. Introduction

AI エージェント研究は前例のない成長を遂げている。本サーベイでは、AI エージェントを理論（Concept）、実装（Code）、そして社会実装（Commerce）の三つの観点から検討し、このフレームワークに基づいて既存の文献を体系的に整理することを目指す。

AI エージェントの理論的基盤（Concept）は 40 年以上にわたって Autonomous Agents and Multiagent Systems（AAMAS）分野で研究されてきた。当初は分散人工知能（Distributed Artificial Intelligence）と呼ばれ、1980 年前後に黒板モデル（Blackboard Model）[1]と契約ネットプロトコル（Contract Net Protocol）[2]といった研究が始まった。この分野は、自律的な主体（Autonomous Agents）の判断形成、および複数の主体（Multi-Agent Systems）による協調・交渉を通じたチームや組織の形成といった根本的なテーマを探求してきた（Section 2 で詳述）。

実装技術（Code）の観点では、大規模言語モデル（LLMs）の登場が AI エージェントの実現可能性を劇的に高めた。Transformer アーキテクチャ[3]の導入から GPT シリーズ[4,5]への発展は、基盤モデル（Foundation Model）の概念を確立し、多様な応用サービスへのファインチューニング技術とプロンプトエンジニアリング[6]が急速に進歩した。LLM は自然言語を用いた知識表現と対話を可能にし、膨大な背景知識を利用した常識推論の実現に繋がった。これにより、LLM は理論的形式

[1] 本稿は同名の英論文の日本語訳である。

化に重点を置いた AAMAS 研究と現実世界との間の乖離を解消する可能性がある。（Section 3 で詳述）

社会実装（Commerce）の面では、Google や Microsoft 等のテック企業や McKinsey 等のコンサルティング企業が、AI エージェント技術への投資を行っている[7]。この関心の背景には、今日の企業内ワークフローにエージェントを導入することで大幅な効率化が期待できるからである。技術的には Model Context Protocol (MCP) [8]や Microsoft の AutoGen [9]等の新たなフレームワークの導入が、サービスコンピューティングを一新しつつある。一方で、倫理的問題、プライバシー、労働市場への影響等の課題も存在する[10,11]（Section 4 で詳述）。

LLM ベースのエージェントの急速な経験的進歩に主に焦点を当てた他のサーベイとは異なり、本稿は、これらの経験的発展を AAMAS の豊かで確立された理論的基盤に接続する。LLM ベースのアプローチを蓄積されたエージェントパラダイムに明示的にマッピングすることで、現在のトレンドをより深く理解し、未開拓の相乗効果を特定し、理論に基づいた実践的な AI エージェント研究を生み出すことを意図している。この包括的な視点は、経験的な成功を超えて、AI エージェントの持続可能な開発へと移行するために不可欠である。

2. Concept: Theoretical Foundations of AI Agents

以下では、Weiss [12,13]、Wooldridge [14]、Shoham and Leyton-Brown [15]の 4 冊の AAMAS 分野の主要な教科書を基に AI エージェントの 8 つの理論的基盤（自律エージェントに関する 4 つとマルチエージェントシステムに関する 4 つ）を示す。

2.1 Core Principles of Autonomous Agents

自律エージェントは、環境を知覚し、内部目標に基づいて決定を下し、独立して行動を実行する能力を持つ計算主体である。

反応型モデル（Reactive Model）: 複雑な内部表現や推論機構を持たず、環境を知覚し行動を生成するモデル。シンプルな反応型エージェントは刺激-反応マッピングで動作し、行動ベースエージェントはモジュラーコンポーネントの組み合わせから複雑な振る舞いを創発する。反応型モデルは高速で堅牢だが、複雑な問題解決には限界がある。

熟考型モデル（Deliberative Model）: 明示的な世界モデルと記号的な知識表現を持ち、論理的推論や計画立案を通じて行動を決定するモデル。BDI アーキテクチャ[16,17]が代表例で、エージェントは環境の状態を信念として保持し、達成すべき目標（欲求）から具体的な意図を形成して行動する。熟考型モデルは高度な推論が可能だが、計算コストが高く動的環境に即応できない。

階層アーキテクチャ（Layered Architecture）: 異なる抽象度を持つ複数の層を組み合わせ、反応性と熟考性のバランスを実現するモデル。反応性を下位層が、熟考性を上位層が担当する。各層が並列に動作し、適切なレベルの処理が選択・実行される。階層型アーキテクチャは複雑な環境での柔軟な行動を可能にするが、層間の調整が課題となる。ロボット制御に用いられたサブサンプションアーキテクチャ[18]では、上位層が必要に応じて下位層を抑制する。

学習型モデル（Learning Model）: 経験を通じて知識や行動を改善し、環境の変化に適応するモデル。ニューラルネットワーク学習は複雑なパターン認識を可能にし、強化学習は報酬信号を用いて試行錯誤から最適な行動政策（policy）を獲得する[19]。学習型モデルは、事前知識が不完全な環境でも動作可能だが、学習に時間を要するため初期性能が低い場合がある。

2.2 Core Principles of Multiagent Systems

マルチエージェントシステムは、複数の自律エージェントが相互作用し、環境を共有し、協力または競争するシステムである。

分散問題解決モデル（Distributed Problem Solving Model）: 複数のエージェントが協調して大規模な問題を分割・並列処理するモデル。分散制約充足問題（DCSP）[20]や分散制約最適化（DCOP）[21]では、中央集権的な制御なしに、各エージェントが局所的な情報と制約を扱いながら全体最適を目指す。分散動的計画法やマルチエージェント MDP[22]は不確実性下での協調的意思決定を扱う。分散問題解決モデルでは、並列化による効率化と通信オーバーヘッドのトレードオフが設計上の課題となる。

ゲーム理論モデル（Game-Theoretic Model）: エージェント間の戦略的相互作用を数理的に分析するモデル。各エージェントは自己の利得最大化を目指し、他者の行動を予測して意思決定を行う。協力ゲームでは利得配分を、非協力ゲームはナッシュ均衡を、繰り返しゲームや確率ゲームは動的な相互作用を分析する。また、望ましい結果を実現するゲームの設計はメカニズムデザインと呼ばれる[23]。ゲーム理論モデルでは、理論上の合理性の仮定と現実における限定合理性のギャップが課題となる。

市場モデル（Market Model）: 市場原理を用いてエージェント間の資源配分と協調を実現するモデル。契約ネットプロトコル[2]はタスクの入札・落札による動的割り当てを行い、交渉プロトコルは提案-応答の構造化により合意形成を促進する[24]。オークション理論は効率的な市場を実現し[25]、Market-Based Programing[26]は市場を通じた資源配分を行う。市場モデルでは、エージェントの戦略的行動による市場の歪みや、不完全情報下での効率性低下が課題となる。

組織設計モデル（Organizational Design Model）: エージェント集団の構造と協調メカニズムを設計するモデル。階層型、ネットワーク型、連合型などの組織構造を通じて役割と責任を明確化する[27]。組織の自己設計（organization self-design）では、エージェントが環境変化に応じて組織構造を動的に再編成する [28]。組織設計モデルでは、エージェントの自律性と組織的制約のバランス、および動的環境への適応が設計上の課題である。

3. Code: Engineering AI Agents

大規模言語モデル（LLM）の出現により、2 章で述べた理論的基盤の実装において大きなパラダイムシフトが起きている。LLM の自然言語処理能力は、形式的なエージェントモデルを実用的なアプリケーションに変換するという長年の課題を解決する可能性を秘めている。

LLM の AI エージェントに対する貢献の第一は、自然言語による知識表現である。これによって、長年の課題であった知識獲得ボトルネック（knowledge acquisition bottleneck）が解決に向かい始める。第二に、直感的な対話インターフェースを通じて説明可能性を向上させる[29]。第三に、暗黙の知識を推論する常識推論の促進に寄与する[30]。また、自然言語を用いた推論は論理的厳密さには劣るものの、実世界の膨大な知識へのアクセスを可能とする。

3.1 Incorporating LLMs into Autonomous Agents

Chain of Thought (CoT)のような高度なプロンプトエンジニアリング技術は、LLM をステップバイステップの推論に導き、その能力を大幅に向上させた[31, 32]。さらに、LLM に直接回答を求めるのではなく、LLM に計画立案や戦略的推論に従事させるようなメタ認知的操作が探求されている。そこで本節では、Autonomous Agents のモデルに向けて LLM がどのように進化統合されつつあるかを整理する。

3.1.1 Evolution of LLMs toward Reactive Models

反応型モデルへの LLM の導入は、現段階では両者の処理速度の差から困難である。一方、LLM の高度な自然言語理解と生成能力は、AI エージェントの実現のために欠かせない。さらに、最近の LLM の高性能化はこのパフォーマンスギャップを徐々に縮めている。

Toolformer[33]では LLM が複雑な環境に対処するための外部ツール（電卓、検索エンジン、カレンダー等）の使用方法を自律的に学習する。ReAct フレームワーク[34]は、環境の観測に加えて、観察と、観察から生成された推論プロセス（「thought」と呼ばれる）を組み合わせ、文脈として処理することで、環境への適応的な動作を可能にした。VOYAGER[35]は、リアクティブな行動生成と学習メカニズムの組み合わせにより、自律的に経験を蓄積し適応できることを Minecraft 環境で示した。

LLM を用いた反応型モデルは、今後、適応性やより繊細な行動応答など、自律エージェント研究の成果を統合することで、さらなる進歩が期待できる。例えば、Believable Agent[36]の文脈で研究されたパーソナリティを導入し、Big Five モデルに基づく人格特性を付与することで、LLM のランダム性を抑え、環境の変化によらず一貫性のある意思決定を実現できる[37]。

3.1.2 Integration of LLMs into Deliberative Models

熟考型モデルには正確な推論と計画立案能力が不可欠である。LLM はこの分野において大きな可能性を示している一方で、重要な技術的課題も抱えている。LLM の事前学習された知識は複雑な問題解決に有効と期待されているが、正確な推論の実現は依然として困難である。

BDI アーキテクチャに LLM を導入したシステム[38]では、自然言語で信念・欲求・意図が記述される。また、自然言語で記述したプランが実行可能なプログラムに変換され、その推論プロセスが自然言語で説明される。このため、開発者はエージェントの挙動を容易に制御できる。また Formal-LLM[39]は、文脈自由文法で与えられた計画をプッシュダウンオートマトンに変換し、その制約下で LLM が行動選択を行う。これは形式的な厳密さと自然言語理解の両方の長所を活用しようとする試みである。

BDI ベースのアプローチと並行して、高度なプロンプティング技術や記号的推論システムとの統合などの多様なアプローチにより LLM の熟考能力の向上が図られている。STRIDE フレームワーク[40]は、LLM に外部ツールを使用させ意思決定問題を解かせ、マルコフ決定過程を含むアルゴリズムのエミュレーションに成功している。Plan and Solve[41]では、LLM が問題解決のための計画を生成するため、推論や計算の過程を観察できる。SELF-REFINE[42]は、生成した回答に対するフィードバックを LLM 自身が生成し、回答の修正を繰り返すことで品質を向上させる。

熟考型モデルへの LLM 統合が期待されているが、自然言語の曖昧性やそれに伴う推論プロセスの検証の困難さが課題である。Human-in-the-Loop のサポートによる自然言語推論と形式的推論の双方向変換が可能となり、両者の長所を併せ持つエージェントが生まれることが期待される。

3.1.3 Integration of LLMs into Layered Architectures

階層型アーキテクチャは、LLM 統合において計算効率と性能の最適化を実現する有力なアプローチである。計算コストの高い LLM を上位層に配置し、下位層には高速な反応型コンポーネントを配置することで、推論能力と反応速度のバランスを実現する。

例えば、LLM を用いた 3 層構造のアーキテクチャが提案されている。最上位層の「Slow Mind」は大規模 LLM で構成され、複雑な意図推論と戦略的計画を担当する。中間層の「Fast Mind」は軽

量 LLM を用い、上位層の戦略を一連のマクロアクションに変換する。最下位層の「Executor」は従来の反応型ポリシーで構成され、ミリ秒単位の原子アクションを実行する。これにより上位層で LLM の柔軟な推論能力を活用しつつ、リアルタイムの行動選択が可能となっている。Overcooked ゲームでの評価により、リアルタイム性が要求される環境で効果的に動作することが示された[43]。

階層的アプローチは、LLM の高い推論能力を活かしながら、システム全体としての応答性を保つアーキテクチャ上の工夫であり、様々なアプリケーションドメインで効果的であることが示されている。

3.1.4 Evolution of LLMs toward Learning Models

LLM と強化学習の関係は、強化学習によって LLM の出力を改善する[44]、LLM を用いて強化学習を行う[45]、の 2 通りがある。LLM のアラインメントは前者であるが、本稿の関心事は後者である。

例えば、Decision Transformer は強化学習をシーケンスモデリング問題として再定式化し、高い報酬につながる行動シーケンスを予測可能にした。推論を担当する LLM は大規模なため、機械学習のようなパラメータ更新は行わず、次の推論への入力となる内省テキストを生成している[46]。Reflexion [47]では、推論を担当する LLM とその結果を評価する LLM、さらに評価結果に基づいて内省を担当する LLM が用意されている。Retroformer [48]では、さらに強化学習の訓練器が組み込まれ、内省用 LLM のパラメータ更新が行われる。

これらのアルゴリズムに基づいて、研究者は LLM ベースのアプローチをロボット工学に適用している。SayCan [49]では 551 個の動作（「コーラを拾う」「テーブルに行く」等）が事前に定義され、自然言語による指示に対し各動作が使用される頻度を LLM が算出する。現在の環境下での実行可能性を勘案して動作が選択される。EUREKA [50]では、LLM が、環境を記述したソースコードとタスクを記述した自然言語から報酬関数候補を生成し、強化学習の訓練結果をフィードバックして改良する。様々な環境で 83%のタスクにおいて人間の専門家を上回り、ペン回しなどの困難な操作を実現した。

今後は、LLM と強化学習を組み合わせ、高レベルの戦略的判断と低レベルの適応的学習を統合する研究が進むと考えられる 。例えば、強化学習エージェントが最適なタイミングで LLM に助言を求める [51]。マルチエージェント強化学習（MARL）[52]では、LLM によってエージェント間の協調が促進される。

一方、強化学習手法と並行して、経験学習を用いた実験も報告されている。エージェントが、訓練タスクから経験を収集し、自然言語で表現された知識を抽出し問題解決に利用する。科学的推論に適用したところ、エージェントはプログラムされた手順なしに複雑な問題を解決し、経験の蓄積に伴いその能力を向上させた[53]。

3.2 Incorporating LLMs into Multiagent Systems

LLM は自然言語能力を通じてマルチエージェントシステムを強化し、現実世界での有用性を向上させる可能性がある。

3.2.1 Evolution of LLMs toward Distributed Problem Solving Models

プロンプトを用いて LLM に複数のエージェントの振る舞いを模倣させ、仮想的なマルチエージェントシステムを構築する研究が進んでいる。これらの研究では、適切に設計されたプロンプトにより、単一の LLM 内で複数エージェントの相互作用を実現し、分散問題解決に活用できることが示されている。

Chain-of-Agents[54]では、長文の質問（入力トークン数は n）を複数に分割（各部分のトークン数は k<<n）し、各部を異なるエージェントが担当する。各エージェントは前のエージェントから要約を受け取り、自身の担当部分と統合して次へ伝達し、最終的に回答が生成される。これにより、質問応答の正答率を維持しつつ計算時間を $O(n^2)$から O(nk)に削減した。また、多様な視点と論理的一貫性の両方を必要とする「意見論述文の生成」という複雑な課題に取り組む研究がある[55]。異なるエージェントに異なるペルソナを割り当て討論させることで、批判と修正を繰り返し、多様性と説得力の両面で優れたエッセイの生成に成功している。

AutoAgents[56]では、メタエージェントがタスクに応じて動的にエージェントを生成する。まず、Drafting Stage でタスクに応じて専門エージェントが生成される。次に Execution Stage で、これらのエージェントは実行中に協力し自己を改善する。開放質問回答(open question answering) において GPT-4 に対する勝率は 76%、創造的なタスクでも精度向上を示した。

一方で、分散問題解決の形式的アプローチへの LLM の適用が進まないのは、自然言語の曖昧性と形式的制約の不整合、分散アルゴリズムの正確性保証の困難さがあるためである。今後は、両者を融合した分散問題解決の手法の考案が期待される。

3.2.2 Integration of LLMs into Game-Theoretic Models

ゲーム理論モデルは伝統的に形式的なフレームワークを必要とするが、LLMは自然言語推論を通じてより柔軟なアプローチを提供する。LLM はゲーム状況や戦略を自然言語で記述・理解し、相手の意図を推論し、常識的知識に基づいてペイオフ構造を推定できる。これにより、形式的な理論モデルと人間の戦略的思考を橋渡しする新たな可能性が開かれる。

LLM の戦略的意思決定能力を明らかにし改善するために「ゲーム論的ワークフロー」[57]が提案されている。囚人のジレンマ等 10 種類の完全情報ゲームと、Deal or No Deal（プレイヤーがアイテムの評価を隠しながら分配を交渉する）と呼ばれる不完全情報ゲームを用いて評価が行われた。支配戦略探索、後ろ向き帰納法、ベイズ更新等を自然言語で実装することでパフォーマンスが大幅に向上した。完全情報ゲームではナッシュ均衡達成率が 45%→76%に改善され、不完全情報ゲームでは 100%の合意率とほぼパレート最適な分配を達成した。

コンテンツ作成における競争をゲーム理論で分析した研究は、人間と生成 AI の共生可能性を示唆している[58]。この研究では、生成 AI が独立したクリエイターとして人間と競合する「排他的競争」と、各クリエイターが生成 AI ツールの採用を選択できる「包括的競争」を検討している。両モデルで安定した均衡が存在し、効率の低いクリエイターは生成 AI に移行する一方、高度な技能を持つ人間クリエイターはニッチ分野で高品質なコンテンツを生産し、むしろ繁栄する可能性があることが示されている。

ゲーム理論モデルに LLM を統合する試みは興味深くサーベイも公開されている[59]が、研究論文まだ限定的である。今後は、戦略的意思決定やメカニズムデザインで、人間と協調するエージェントが生まれることが期待される。

3.2.3 Integration of LLMs into Market Models

市場モデルでは、AI エージェントが市場に参加し、その知識と推論能力を用いて入札戦略を決定することが考えられる。オークション環境でエージェントの振る舞いを評価するフレームワークが提案され、予算制約の管理、長期的な目標と行動との整合性維持、競合他者の行動予測などで AI エージェントの能力が確認されている [60]。

LLM がオークションを通じて広告収入を生成する仕組みに関する研究が登場している。これは LLM を入札エージェントや市場設計に用いるのではなく、LLM 自体を広告プラットフォームとして

活用する試みである。初期の研究では、LLM が生成する文章のトークン確率分布を広告枠として販売する手法が提案された。例えば、自社ホテル名の出現確率を高める分布に最高額で入札した企業が落札すると、LLM はその分布に従って文章を生成する[61]。しかし、この手法はコンテンツ操作という倫理的問題を含んでいる。

より洗練されたアプローチとして、Retrieval-Augmented Generation (RAG) [62]を用いた広告統合フレームワークがある。トークン確率を直接操作する代わりに、文や段落などの談話セグメントを広告枠として扱う。このメカニズムは入札額と関連性スコアに基づいて確率的に広告を選択し、LLM が選ばれた広告を応答に自然に組み込む。理論的には、このセグメントオークションは対数的社会福祉を最大化し、効率性と公平性のバランスを保ちながらインセンティブ互換性を維持する。実証評価では、収益性と出力品質のトレードオフが明らかになっている。単一広告の繰り返しオークションは高収益を生み、複数広告オークションは出力品質が高い。

LLM の市場モデルへの統合は初期段階にある。今後は、LLM の計算複雑性や戦略的操作への脆弱性の克服が期待される。

3.2.4 Evolution of LLMs toward Organizational Design Models

組織設計モデルにおいて、LLM ベースのエージェントは前例のない変革をもたらしている。マルチエージェントシステムの組織設計には、メタプロンプトを用いた組織の自己設計（organization self-design）、大規模なマルチエージェントシミュレーション、そして社会行動や規範の創発が含まれる。

Criticize-Reflect フレームワーク[63]では、LLM が自律的に階層的組織を改善していく。Criticize が対話履歴から指示の重複、調整不足等の問題点を抽出し、Reflect が新たな組織構造やコミュニケーションルールを含むプロンプトを生成する。このフレームワークは、チェーン構造や動的リーダーシップを含む組織形態を生み出し、最大 30%の効率改善を達成した。

大規模マルチエージェントシミュレーションも試みられている。NYC 住民の COVID-19 流行時の行動が再現された。まず、住民エージェントが属性（年齢、性別など）によってグループ化され、各グループの行動が LLM に繰り返し照会される。次に、LLM が生成する回答の分布から住民グループの実際の行動分布を推定する。LLM の回答分布が各グループの行動分布を表すと仮定することで、大規模シミュレーションの効率的な実装を可能としている[64]。

一方、協働メカニズムに関する社会心理学的視点が探求されている。例えば、異なる性格特性（easy-going/overconfident）と思考パターン（debate/reflection）を持つエージェントを組み合わせることで、高性能な組織構造が構成できる。この研究では、議論から始まる戦略を持つ三体のチームが優れた性能を示したと報告されている [65]。エージェントの相互作用から社会的規範が形成されることを実験的に示した研究もある。25 体のエージェントによる命名ゲームを観察した結果、共通の命名規則への収束、集団的バイアスが現れた。少数派グループが規範の変革を推進できることも確認されていて興味深い[66]。エージェントが自律的に相互の関係を形成し、創発的な協調行動を発現させたとする研究もある。経験を記憶し、それらを統合した内省（reflection）を生成することで、一貫性のある長期的な行動計画を開発し、人間のような行動パターンを示した[67]。

組織設計の研究は現段階では実験的色彩が強い。例えば、これまでに観察された「社会的な現象」が、エージェントの相互作用から創発されたものか、LLM の事前学習知識に由来するものか、あるいは Transformer アーキテクチャのメカニズムによるものかは明確でない。システムの一部を除去してその部分の貢献度を測る Ablation Study や、創発的行動の源を分離するためにエージェント間のコミュニケーションを制限する Controlled Simulation などを用いた分析が期待される。

4. Commerce: AI Agents in Business and Society

研究者が AI エージェント研究に魅了されている一方で、実用的な開発も、特にソフトウェア開発などの領域で急速な進展している。

4.1 Revolution in Service Workflows

従来のサービスワークフローは、事前に設計された静的な手順に基づいて、サービスやツールの呼び出し順序を明示的に定義する必要があった [68]。この手法は、進化する業務内容やユーザー要求の変化に対応が困難であり、複雑でコンテキスト依存の処理に対する適応性に欠ける。これに対し AI エージェントは、自然言語で与えられた目的を理解し、タスクを動的に分解・再構成し、外部ツールや他のエージェントと連携して自律的にワークフローを構築・実行する新たな枠組みを提供する。

AI エージェントの適用分野としては、LLM によるソフトウェア開発にワークフローを導入する試みが他の応用分野に先行し、多数報告されている。MetaGPT [69]では、プロダクトマネージャや設計者、開発者などの役割を持つ複数のエージェントが協調し、ソフトウェア開発を遂行する。。ChatDev [70]では、CEO、CTO、プログラマ、レビュア、テスタの 5 つの役割を持つエージェントが協働する。開発を設計・コーディング・テストの 3 つのフェーズに分割し、各フェーズで異なる役割のエージェントが対話を繰り返し、エラーや不完全な実装を発見・修正する。エージェントが相互に不明確な点を質問し合うことにより、実行可能なコードを生成する。

プラットフォームの整備もアプリケーションと並行して進み始めた。AutoGPT[71]は、自然言語で与えられた目標を基に中間的なサブタスクを自動的に生成し、必要なステップを自律的に実行する。BabyAGI [72]は、タスクの作成・優先順位付け・実行を専門エージェントが分担し、効率的に処理を進める。シンプルで理解しやすい設計により、エージェントの協調作業の可能性を示した。AutoGen [9]は、役割の異なる複数のエージェントが会話を通じて協力してタスクを解決するオープンソースフレームワークである。開発者が各エージェントの役割を定義し、会話パターンを設計することができる。

また、従来のサービスワークフローでは、セッションをまたぐ文脈や知識を保持するのが困難であった。これに対し、LangChain [73]は長期記憶（long-term memory）機能を備え、過去の対話や履歴の保存・参照を可能にする。これにより、継続的なタスク実行、文脈に応じた応答、蓄積知識の再利用などが可能となり、対話の一貫性と業務知識の継承を支える。なお、後に開発された LangGraph [74]は、複数エージェントのワークフローをグラフ構造で表現する。

サービス連携のコストを高める要因の一つが API の非互換性である。Anthropic が主導して開発した Model Context Protocol (MCP) [8] は、ツールやサービスの機能を「アクション」として定義し、その入出力形式や前提条件を記述する統一的なインターフェースを提供する。この仕組みにより、エージェントは対象サービスの構造や仕様を明示的に把握せずとも、スキーマに基づいて適切なツールを選択し、呼び出し、結果を解釈することが可能となる。 エージェントは異種 API 間の互換性問題を回避しながら、安全かつ動的にサービスワークフローを構築・実行できるようになる。

AI エージェントのサービスワークフローは、ソフトウェア開発だけでなく多様な分野で導入されつつある。たとえば、ゲーム環境での行動生成 [35]や科学的発見支援 [75]などが挙げられる。。エージェントは、目的指向性、動的適応性、協調性、知識継承といった特性を生かし、自律的かつ柔軟なワークフローの実現を可能にする。

4.2 Ethical and Regulatory Challenges

AI エージェントの社会的影響への関心が倫理的研究を促進している。例えば、AI エージェントの倫理的設計手法として、トップダウン（明示的な倫理規則の実装）、ボトムアップ（強化を通じた倫理的行動の学習）、ハイブリッド（両者の組み合わせ）の 3 つのアプローチが提案されている。また、エージェントの自律性・意図性・責任が倫理的懸念を引き起こすため、モラルチューリングテストによる道徳的能力の評価手法が検討されている[76]。

AI エージェントの社会的浸透は、教育、医療、行政といった人間中心の制度設計にも再検討を促している。2024 年に制定された EU AI 法（包括的 AI 規制法）も一般の関心を集めている。教育では、生成 AI の長期的活用が学習者の思考スタイルを変化させる可能性があり、教員と学生の関係性や評価モデルの再構築が求められている。また、エージェントへの過剰な依存によって、創造性や共感などの人間的側面を減少させる懸念も生まれている。

さらに、生成 AI がサイバーセキュリティに与える影響は「両刃の剣」として分析されている。ジェイルブレイキングなどの手法で生成 AI の制約を回避した場合、マルウェア生成が可能である。一方で、生成 AI はセキュリティ運用の自動化や脅威検知などの防御面でも活用できる。このことは、AI 技術の発展に対応した適切なセキュリティガバナンスの必要性を示している[77]。

社会的存在として登場した AI エージェントが、人間の行動や価値観、制度にどのように作用し得るのかを分析し、技術の発展に伴うリスクへの警鐘に耳を傾けることが重要である。大規模言語モデルは真の知性とは異なり道徳的思考ができず、科学と倫理を損なう危険があるとの警告がある[78]。超人的 AI の実現が人類史上最大の出来事になり得ると同時に、最後の出来事になる危険性すら指摘されている[79]。

5. Conclusion

本稿では、AI エージェントの研究を、AAMAS の視点から紹介した。LLM 開発に焦点を当てた多くのサーベイが存在する中で、本稿は研究者に理論的背景に基づく長期的な視点を提供することを意図したものである。同時に、このサーベイを通じて我々自身が得た気づきもある。

例えば、反応型モデルや分散問題解決モデルでは、LLM 研究者は熱心に LLM の進化に取り組んでいるが、過去の形式的モデリングへの理解は限定的である。一方、熟考型モデルやゲーム理論モデルは LLM の取り込みが有望視されるものの、研究報告は限られている。

また、LLM を進化させる研究は実証実験に重きを置き、明確な理論的基盤がなくても、迅速に実験し結果を公開する。このアプローチは、深層学習以降、実証的研究が重要な社会的貢献を果たしてきたことに起因する。一方、理論モデル研究者は、実験は形式的フレームワーク内で理解されなければならないと信じている。そのため、自然言語による知識の表現や推論の導入に躊躇がある。

今後は、文書生成、創造的芸術、ソフトウェア開発など、LLM の実用性が明確な領域から AI エージェントの導入が進むと思われる。しかし、その先に向かうためには、自然言語を扱う LLM と形式的アプローチをとる理論モデルの融合、すなわち研究コミュニティを超えた協働が不可欠となる。AI エージェントの研究論文は、機械学習、自然言語、AAMAS などの多様な分野で発表されており、学際的な広がりを見せている。

重要なのは、本サーベイが、堅牢な AAMAS の理論的フレームワークと LLM の実用的な能力との明確な統合を求めていることである。AAMAS の自律性、協調、意思決定に関する基礎的な理解を、LLM の前例のない自然言語処理および生成能力と組み合わせることで、AI エージェント分野が現在の経験

的視点を超越できると主張する。この統合は単なる学術的な演習ではなく、現実世界の問題に対処できる、説明可能で信頼性が高く、倫理的に健全な AI エージェントを構築するために必要である。この統合された視点は、確立された理論が LLM ベースのエージェント設計に枠組みを提供し、そして逆に、LLM が AAMAS の長年の課題に新たな解決策を提供する。本サーベイが、時間の限られた研究者にとって、異なるコミュニティの研究活動に目を向ける一助となれば幸いである。

References

[1] Erman, Lee D., Hayes-Roth, Frederick, Lesser, Victor R., & Reddy, D. Raj (1980). The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. ACM Computing Surveys, 12(2), 213-253. DOI: 10.1145/356810.356816

[2] Smith, Reid G. (1980). The contract net protocol: High-level communication and control in a distributed problem solver. IEEE Transactions on Computers, C-29(12), 1104-1113. DOI: 10.1109/TC.1980.1675516

[3] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, & Polosukhin, Illia (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 5998-6008).

[4] Radford, Alec, Narasimhan, Karthik, Salimans, Tim, & Sutskever, Ilya (2018). Improving language understanding by generative pre-training. OpenAI Blog, June 11, 2018. https://openai.com/blog/language-unsupervised/

[5] Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D., Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Child, Rewon, Ramesh, Aditya, Ziegler, Daniel M., Wu, Jeffrey, Winter, Clemens, Hesse, Christopher, Chen, Mark, Sigler, Eric, Litwin, Mateusz, Gray, Scott, Chess, Benjamin, Clark, Jack, Berner, Christopher, McCandlish, Sam, Radford, Alec, Sutskever, Ilya, & Amodei, Dario (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (pp. 1877-1901).

[6] Liu, Pengfei, Yuan, Weizhe, Fu, Jinlan, Jiang, Zhengbao, Hayashi, Hiroaki, & Neubig, Graham (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), Article 195, 1-35. DOI: 10.1145/3560815

[7] Yee, Lareina, Chui, Michael, Roberts, Roger, & Xu, Sandy (2024). Why agents are the next frontier of generative AI. McKinsey Digital Practice, July 24, 2024.

[8] Anthropic. (2024). Model Context Protocol: An open standard for seamless integration between LLMs and external data. https://www.anthropic.com/news/model-context-protocol

[9] Wu, Qingyun, Bansal, Gagan, Zhang, Jieyu, Wu, Yiran, Li, Beibin, Zhu, Erkang, Jiang, Li, Zhang, Xiaoyun, Zhang, Shaokun, Liu, Jiale, Awadallah, Ahmed Hassan, White, Ryen W., Burger, Doug, & Wang, Chi (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In Conference on Language Modeling (COLM) 2024.

[10] Weidinger, Laura, Mellor, John, Rauh, Maribeth, Griffin, Conor, Uesato, Jonathan, Huang, Po-Sen, Cheng, Myra, Glaese, Mia, Balle, Borja, Kasirzadeh, Atoosa, Kenton, Zac, Brown, Sasha, Hawkins, Will, Stepleton, Tom, Biles, Courtney, Birhane, Abeba, Haas, Julia, Rimell, Laura, Hendricks, Lisa Anne, Isaac, William, Legassick, Sean, Irving, Geoffrey, & Gabriel, Iason (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

[11] Purdy, Mark (2024). What is agentic AI, and how will it change work? Harvard Business Review, December 12, 2024.

[12] Weiss, Gerhard (Ed.). (1999). Multiagent systems: A modern approach to distributed artificial intelligence. MIT Press.

[13] Weiss, Gerhard (Ed.). (2013). Multiagent systems (2nd ed.). MIT Press.

[14] Wooldridge, Michael (2009). An introduction to multiagent systems (2nd ed.). John Wiley & Sons.

[15] Shoham, Yoav, & Leyton-Brown, Kevin (2008). Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press.

[16] Bratman, Michael (1987). Intention, Plans, and Practical Reason. Harvard University Press.

[17] Rao, Anand S., & Georgeff, Michael P. (1991). Modeling rational agents within a BDI-architecture. In Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR 1991) (pp. 473-484).

[18] Brooks, Rodney A. (1986). A robust layered control system for a mobile robot. IEEE Journal on Robotics and Automation, 2(1), 14-23.

[19] Sutton, Richard S., & Barto, Andrew G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

[20] Yokoo, Makoto, Durfee, Edmund H., Ishida, Toru, & Kuwabara, Kazuhiro (1998). The distributed constraint satisfaction problem: Formalization and algorithms. IEEE Transactions on Knowledge and Data Engineering (TKDE), 10(5), 673-685.

[21] Modi, Pragnesh Jay, Shen, Wei-Min, Tambe, Milind, & Yokoo, Makoto (2005). ADOPT: Asynchronous distributed constraint optimization with quality guarantees. Artificial Intelligence, 161(1-2), 149-180.

[22] Bernstein, Daniel S., Givan, Robert, Immerman, Neil, & Zilberstein, Shlomo (2002). The complexity of decentralized control of Markov decision processes. Mathematics of Operations Research, 27(4), 819-840.

[23] Nisan, Noam, Roughgarden, Tim, Tardos, Eva, & Vazirani, Vijay V. (2007). Algorithmic game theory. Cambridge University Press.

[24] Jennings, Nicholas R., Faratin, Peyman, Lomuscio, Alessio R., Parsons, Simon, Wooldridge, Michael J., & Sierra, Carles (2001). Automated negotiation: Prospects, methods and challenges. Group Decision and Negotiation, 10(2), 199-215.

[25] Klemperer, Paul (2004). Auctions: Theory and practice. Princeton University Press.

[26] Wellman, Michael P., Walsh, William E., Wurman, Peter R., & MacKie-Mason, Jeffrey K. (2001). Auction protocols for decentralized scheduling. Games and Economic Behavior, 35(1-2), 271-303.

[27] Horling, Bryan, & Lesser, Victor (2004). A survey of multi-agent organizational paradigms. The Knowledge Engineering Review, 19(4), 281-316.

[28] Ishida, Toru, Gasser, Les, & Yokoo, Makoto (1992). Organization self-design of distributed production systems. IEEE Transactions on Knowledge and Data Engineering (TKDE), 4(2), 123-134.

[29] Slack, Dylan, Krishna, Satyapriya, Lakkaraju, Himabindu, & Singh, Sameer (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. Nature Machine Intelligence, 5(8), 873-883.

[30] Li, Xiang Lisa, Kuncoro, Adhiguna, Hoffmann, Jordan, de Masson d'Autume, Cyprien, Blunsom, Phil, & Nematzadeh, Aida (2022). A systematic investigation of commonsense knowledge in large language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022) (pp. 11838-11855).

[31] Wei, Jason, Wang, Xuezhi, Schuurmans, Dale, Bosma, Maarten, Ichter, Brian, Xia, Fei, Chi, Ed H., Le, Quoc V., & Zhou, Denny (2022). Chain-of-thought prompting elicits reasoning in large

language models. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022) (pp. 24824-24837).

[32] Kojima, Takeshi, Gu, Shixiang Shane, Reid, Machel, Matsuo, Yutaka, & Iwasawa, Yusuke (2022). Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022) (pp. 22199-22213).

[33] Schick, Timo, Dwivedi-Yu, Jane, Dessì, Roberto, Raileanu, Roberta, Lomeli, Maria, Zettlemoyer, Luke, Cancedda, Nicola, & Scialom, Thomas (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023). arXiv preprint arXiv:2302.04761.

[34] Yao, Shunyu, Zhao, Jeffrey, Yu, Dian, Du, Nan, Shafran, Izhak, Narasimhan, Karthik, & Cao, Yuan (2023). ReAct: Synergizing reasoning and acting in language models. In 11th International Conference on Learning Representations (ICLR) 2023.

[35] Wang, Guanzhi, Xie, Yuqi, Jiang, Yunfan, Mandlekar, Ajay, Xiao, Chaowei, Zhu, Yuke, Fan, Linxi, & Anandkumar, Anima (2024). VOYAGER: An open-ended embodied agent with large language models. Transactions on Machine Learning Research (TMLR).

[36] Bates, Joseph (1994). The role of emotion in believable agents. Communications of the ACM, 37(7), 122-125.

[37] Newsham, Luke, & Prince, David (2025). Personality-driven decision-making in LLM-based autonomous agents. In Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025).

[38] Ichida, André Y., Meneguzzi, Felipe, & Cardoso, Rafael C. (2024). BDI agents in natural language environments. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024).

[39] Li, Zelong, Hua, Wenyue, Wang, Hao, Zhu, He, & Zhang, Yongfeng (2024). Formal-LLM: Integrating Formal Language and Natural Language for Controllable LLM-based Agents. arXiv preprint arXiv:2402.00798.

[40] Li, Can, Yang, Ruotong, Li, Tianyi, Bafarassat, Milad, Sharifi, Kourosh, Bergemann, Dirk, & Yang, Zhuoran (2024). STRIDE: A tool-assisted LLM agent framework for strategic and interactive decision-making. Cowles Foundation Discussion Paper No. 2393. arXiv preprint arXiv:2405.16376.

[41] Wang, Lei, Xu, Wanyu, Lan, Yihuai, Hu, Zhiqiang, Lan, Yunshi, Lee, Roy Ka-Wei, & Lim, Ee-Peng (2023). Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023) (pp. 2609-2634).

[42] Madaan, Aman, Tandon, Niket, Gupta, Prakhar, Hallinan, Skyler, Gao, Luyu, Wiegreffe, Sarah, Alon, Uri, Dziri, Nouha, Prabhumoye, Shrimai, Yang, Yiming, Gupta, Shashank, Majumder, Bodhisattwa Prasad, Hermann, Katherine, Welleck, Sean, Yazdanbakhsh, Amir, & Clark, Peter (2023). SELF-REFINE: iterative refinement with self-feedback. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023) (pp. 46534-46594).

[43] Liu, Jingqing, Yu, Chongyang, Gao, Jianzhu, Xie, Yitao, Liao, Qiyue, Wu, Yi, & Wang, Yaliang (2024). LLM-powered hierarchical language agent for real-time human-ai coordination. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024).

[44] Wang, Shihan, Zhang, Shuai, Zhang, Jiayu, Hu, Rui, Li, Xiang, Zhang, Tao, Li, Jun, Wu, Fei, Wang, Guoyin, & Hovy, Eduard (2024). Reinforcement learning enhanced LLMs: A survey. arXiv preprint arXiv:2412.10400

[45] Cao, Yinuo, Zhao, Hongpeng, Cheng, Yanchi, Shu, Ting, Chen, Yuheng, Liu, Guolong, Liang, Gaoqi, Zhao, Junwei, Yan, Jianxing, & Li, Yali (2024). Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. IEEE Transactions on Neural Networks and Learning Systems (TNNLS). arXiv preprint arXiv:2404.00282.

[46] Chen, Lili, Lu, Kevin, Rajeswaran, Aravind, Lee, Kimin, Grover, Aditya, Laskin, Misha, Abbeel, Pieter, Srinivas, Aravind, & Mordatch, Igor (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021) (pp. 15084-15097).

[47] Shinn, Noah, Cassano, Federico, Gopinath, Ashwin, Narasimhan, Karthik, & Yao, Shunyu (2023). Reflexion: language agents with verbal reinforcement learning. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023) (pp. 8634-8652).

[48] Yao, Weiran, Heinecke, Shelby, Niebles, Juan Carlos, Liu, Zhiheng, Feng, Yuchi, Xue, Le, Murthy, Rithesh, Chen, Zeyuan, Zhang, Jiajun, Arpit, Devansh, Xu, Ran, Mui, Phil, Wang, Haiying, Xiong, Caiming, & Savarese, Silvio (2024). Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024).

[49] Ahn, Michael, Brohan, Anthony, Brown, Noah, Chebotar, Yevgen, Cortes, Omar, David, Byron, Finn, Chelsea, Fu, Chuyuan, Gopalakrishnan, Keerthana, Hausman, Karol, Herzog, Alex, Ho, Daniel, Hsu, Jasmine, Ibarz, Julian, Ichter, Brian, Irpan, Alex, Jang, Eric, Ruano, Rosario Jauregui, Jeffrey, Kyle, Jesmonth, Sally, Joshi, Nikhil J., Julian, Ryan, Kalashnikov, Dmitry, Kuang, Yuheng, Lee, Kuang-Huei, Levine, Sergey, Lu, Yao, Luu, Linda, Parada, Carolina, Pastor, Peter, Quiambao, Jornell, Rao, Kanishka, Rettinghouse, Jarek, Reyes, Diego, Sermanet, Pierre, Sievers, Nicolas, Tan, Clayton, Toshev, Alexander, Vanhoucke, Vincent, Xia, Fei, Xiao, Ted, Xu, Peng, Xu, Sichun, Yan, Mengyuan, & Zeng, Andy (2023). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In Proceedings of the 6th Conference on Robot Learning (CoRL 2022), PMLR 205 (pp. 287-318).

[50] Ma, Yecheng Jason, Liang, William, Wang, Guanzhi, Huang, De-An, Bastani, Osbert, Jayaraman, Dinesh, Zhu, Yuke, Fan, Linxi, & Anandkumar, Anima (2024). Eureka: Human-Level Reward Design via Coding Large Language Models. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024).

[51] Hu, Bin, Zhao, Chenyang, Zhang, Pu, Zhou, Zihao, Yang, Yuanhang, Xu, Zenglin, & Liu, Bin (2024). Enabling intelligent interactions between an agent and an LLM: A reinforcement learning approach. Reinforcement Learning Journal, 3, 1289-1305.

[52] Sun, Chao, Huang, Shuo, & Pompili, Dario (2025). LLM-based multi-agent decision-making: Challenges and future directions. IEEE Robotics and Automation Letters, 10(6), 5682-5685. arXiv preprint arXiv:2405.11106.

[53] Zhao, Andrew, Huang, Daniel, Xu, Quan, Lin, Matthieu, Liu, Yong-Jin, & Huang, Gao (2024). ExpeL: LLM Agents Are Experiential Learners. In Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024).

[54] Zhang, Yusen, Sun, Rui, Chen, Yumeng, Pfister, Tomas, Zhang, Rui, & Arik, Sercan Ö. (2024). Chain of agents: Large language models collaborating on long-context tasks. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

[55] Hu, Zhe, Chan, Ho Pui, Li, Jiachen, & Yin, Yang (2025). Debate to write: A persona-driven multi-agent framework for diverse argument generation. In Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025) (pp. 4689-4703). arXiv preprint arXiv:2406.19643.

[56] Chen, Guangyao, Dong, Siwei, Shu, Yu, Zhang, Ge, Sesay, Jaward, Karlsson, Börje, Fu, Jie, & Shi, Yemin (2024). AutoAgents: A Framework for Automatic Agent Generation. In Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024).

[57] Hua, Wenyue, Liu, Olga, Li, Lun, Amayuelas, Alfonso, Chen, Jiajie, Jiang, Liyi, Jin, Mingyu, Fan, Liuqing, Sun, Fei, Wang, William, Wang, Xingwei, & Zhang, Yanghua (2024). Game-theoretic LLM: Agent workflow for negotiation games. arXiv preprint arXiv:2411.05990.

[58] Yao, Fan, Li, Chuanhao, Nekipelov, Denis, Wang, Hongning, & Xu, Haifeng (2024). Human vs. Generative AI in Content Creation Competition: Symbiosis or Conflict?. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Proceedings of Machine Learning Research, 235, 56885-56913. arXiv preprint arXiv:2402.15467.

[59] Sun, Haoran, Qin, Wenyu, Liang, Yuxuan, & Wang, Zhuohan (2025). Game Theory Meets Large Language Models: A Systematic Survey. arXiv preprint arXiv:2502.09053.

[60] Chen, Jiangjie, Yuan, Siyu, Ye, Rong, Majumder, Bodhisattwa Prasad, & Richardson, Kyle (2023). Put your money where your mouth is: Evaluating strategic planning and execution of LLM agents in an auction arena. arXiv preprint arXiv:2310.05746.

[61] Duetting, Paul, Mirrokni, Vahab, Paes Leme, Renato, Xu, Haifeng, & Zuo, Song (2024). Mechanism design for large language models. In Proceedings of the ACM Web Conference 2024 (WWW 2024). arXiv preprint arXiv:2310.10826.

[62] Hajiaghayi, MohammadTaghi, Lahaie, Sébastien, Rezaei, Keivan, & Shin, Suho (2024). Ad Auctions for LLMs via Retrieval Augmented Generation. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2024). arXiv preprint arXiv:2406.09459.

[63] Guo, Xudong, Huang, Kaixuan, Liu, Jiale, Fan, Wenhao, Vélez, Nataniel, Wu, Qingyun, Wang, Huazheng, Griffiths, Thomas L., & Wang, Mengdi (2024). Embodied LLM agents learn to cooperate in organized teams. arXiv preprint arXiv:2403.12482.

[64] Chopra, Ayush, Kumar, Sai, Giray-Kuru, Nurullah, Raskar, Ramesh, & Quera-Bofarull, Arnau (2025). On the limits of agency in agent-based models. In Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025).

[65] Zhang, Jiawei, Xu, Ximing, Zhang, Ningyu, Liu, Ruibo, Hooi, Bryan, & Deng, Shumin (2024). Exploring collaboration mechanisms for LLM agents: A social psychology view. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) (pp. 14544-14607). arXiv preprint arXiv:2310.02124.

[66] Ashery, Amir Feder, Aiello, Luca Maria, & Baronchelli, Andrea (2025). Emergent social conventions and collective bias in LLM populations. Science Advances, 11, eadu9368. DOI: 10.1126/sciadv.adu9368

[67] Park, Joon Sung, O'Brien, Joseph C., Cai, Carrie J., Morris, Meredith Ringel, Liang, Percy, & Bernstein, Michael S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST 2023). DOI: 10.1145/3586183.3606763

[68] Papazoglou, Mike P., Traverso, Paolo, Dustdar, Schahram, & Leymann, Frank (2008). Service-oriented computing: a research roadmap. International Journal of Cooperative Information Systems, 17(2), 223-255.

[69] Hong, Sirui, Zhuge, Mingchen, Chen, Jiaqi, Zheng, Xiawu, Cheng, Yuheng, Zhang, Ceyao, Wang, Jinlin, Wang, Zili, Yau, Steven Ka Shing, Lin, Zijuan, Zhou, Liyang, Ran, Chenyu, Xiao, Lingfeng, Wu, Chenglin, Schmidhuber, Jürgen (2024). MetaGPT: Meta Programming for a Multi Agent Collaborative Framework. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024).

[70] Qian, Chen, Liu, Wei, Liu, Hongzhang, Chen, Nuo, Dang, Yufan, Li, Jiahao, Yang, Cheng, Chen, Weize, Su, Yusheng, Cong, Xin, Xu, Juyuan, Li, Dahai, Liu, Zhiyuan & Sun, Maosong (2024). ChatDev: Communicative Agents for Software Development. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) (pp. 15174-15186). DOI: 10.18653/v1/2024.acl-long.810

[71] Significant Gravitas. (2023). AutoGPT. GitHub repository. https://github.com/Significant-Gravitas/AutoGPT (accessed June 2025).

[72] Nakajima, Yohei (2023). BabyAGI. GitHub repository. https://github.com/yoheinakajima/babyagi (accessed June 2025).

[73] Topsakal, Oguzhan, & Akinci, Tahir Cetin (2023). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In Proceedings of the International Conference on Advanced Engineering and Natural Sciences (ICAENS), 1(1) (pp. 1050-1056).

[74] LangChain. (2023). LangGraph. GitHub repository. https://github.com/langchain-ai/langgraph (accessed June 2025).

[75] Wang, Hanchen, Fu, Tianfan, Du, Yuqing, Gao, Wenhao, Huang, Kexin, Liu, Ziming, Chandak, Paridhi, Liu, Shengqi, Van Katwyk, Peter, Deac, Andreea, Anandkumar, Anima, Bergen, Karianne, Gomes, Carla P., Ho, Shirley, Kohli, Pushmeet, Lasenby, Joan, Leskovec, Jure, Liu, Tie-Yan, Mirzasoleiman, Baharan, Mishra, Debora, Nawroth, Guido, Paliwal, Soumya, Perozzi, Bryan, Schwaller, Philippe, Seljak, Uroš, Sherborne, Olivia, Simm, Gregor N. C., Singh, Hannu, Sorrenson, Peter, Stein, Jared, Tang, Jian, Veličković, Petar, Welling, Max, Willmore, Benedict, Zitnik, Marinka, & Regol, Fiona (2023). Scientific discovery in the age of artificial intelligence. Nature, 620(7972), 47-60.

[76] Huang, Chao, Zhang, Zhi, Mao, Bing, & Yao, Xin (2023). An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence (IEEE Trans. AI), 4(4), 799-819.

[77] Gupta, Maanak, Akiri, CharanKumar, Aryal, Kshitiz, Parker, Eli, & Praharaj, Lopamudra (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. IEEE Access, 11, 80218-80245.

[78] Chomsky, Noam, Roberts, Ian, & Watumull, Jeffrey (2023). The false promise of ChatGPT. The New York Times, March 8, 2023.

[79] Russell, Stuart (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Toru Ishida is Professor Emeritus at Kyoto University and Visiting Professor at Telkom University. He has contributed to the foundations of autonomous and multiagent systems since the 1980s, co-chairing the first AAMAS conference in 2002. His projects, including Digital City Kyoto, the Intercultural Collaboration Experiment, and the Language Grid, exemplify the societal deployment of AI technologies. He is a Life Fellow of the IEEE and former President of the IEICE. He currently explores how Large Language Models and AI Agents can address social challenges.

Yohei Murakami received his Ph.D. in Informatics from Kyoto University in 2006. He is currently a Professor in the Faculty of Information Science and Engineering at Ritsumeikan University, Japan. He leads the Indonesian Language Sphere project, which leverages human–AI collaboration to develop language resources and preserve endangered languages. He has served as a program committee member for major conferences in natural language processing, artificial intelligence, and multi-agent systems, including LREC-COLING, AAAI, AAMAS, and PRIMA. His current research interests include applying theoretical multi-agent models to AI agents.

Donghui Lin received his Ph.D. in 2008 from the Department of Social Informatics, Kyoto University. He is currently an Associate Professor in the Faculty of Environmental, Life, Natural Science and Technology at Okayama University, where he leads the Intelligent Computing Laboratory. He has served as a program committee member for major AI and agent-related conferences such as AAAI, AAMAS, PRIMA, and PRICAI. His current research interests include multi-agent systems, services computing, Internet of Things, and AI agents.

Kemas Muslim Lhaksmana received his Ph.D. in Social Informatics from Kyoto University, Japan. He currently serves as the Dean of the School of Computing at Telkom University and is the Treasurer of the IEEE Indonesia Section (2018–2022, 2025–present). His research interests include people analytics and the integration of generative AI in education. He is participating in this study as part of a broader research project investigating the impact and interaction models of Large Language Models (LLMs) in higher education in the Global South.