

AIエージェントの実用化に向けた論点の整理

～安全・プライバシー・責任をどう確保するか～

一般財団法人日本情報経済社会推進協会 電子情報利活用研究部
次長 松下 尚史

【要約】

AIが「使うもの」から「任せるもの」に変わりつつある。目標を与えるだけで、自律的に計画を立て・ツールを使い・行動するAIエージェントの実用化が急速に進んでいる。日本・米国・シンガポールで制度整備が始まったが、確立した解答はまだない。

本稿はその論点を三つの層に整理する。①プライバシーの扱い・AIへの依存・社会の均質化など、AIですでに問題になっていること。②「誰がリスクを把握・管理するのか」「何か起きたとき誰が責任を負うのか」というAIエージェント固有の安全・責任の問い。③エージェントの「身元」と「権限」を証明する仕組みが未整備であるという制度設計の課題である。

AIエージェントが問うているのはAIの制御技術ではない。自律的に行動する代理人を持つ社会において、人間の責任と合意をどう設計するかという根本的な問いである。

第1章 AIエージェントとは何か

1-1. 「使うAI」から「任せるAI」へ

AIの活用形態が質的に変化しつつある。従来のAIは、人間が指示を与えるたびに応答を返すものだった。問いに対して答えを出す、指示に対してテキストや画像を生成する——その都度、人間が判断し、AIはその支援に徹する。

これに対してAIエージェントは、人間から目標を与えられると、自らその達成に向けた計画を立て、外部のツールやシステムと連携しながら、複数のステップにわたって自律的に行動する。メールを整理し、スケジュールを調整し、ウェブを検索し、必要であれば購買の手続きまで完結させる。人間は都度の指示を与えず、結果だけを受け取る。AIは「使うもの」から「任せるもの」になった。

商用サービスも急速に立ち上がっている。OpenAIは2025年1月にウェブブラウザを操作して自律的にタスクを実行するエージェント「Operator」を公開し、その後、同機能をChatGPT本体に統合した¹。SalesforceのAgentforceは年間経常収益5億4,000万ドルを超え、1万8,500件を超えるAgentforce契約を持つまでに成長している²。MicrosoftはMicrosoft 365 CopilotにAgent Modeを実装し、Word・Ex

¹ [「OpenAI for Developers in 2025 — A year-end roundup of the biggest model, API, and platform shifts for building production-grade agents」](#) (OpenAI, 2025年12月)

² Salesforce Agentforceの業績については以下を参照。 [Salesforce to launch Agentforce 360 access in ChatGPT, Digital Commerce 360](#) (2025年10月)

cel・PowerPointにおける自律的な文書作成・データ処理を可能にした³。日本でもサイバーエージェントグループのAI Shiftが2025年3月に自律型AIエージェント「AI Worker」をリリースし、カスタマーサポートや社内ヘルプデスクでの導入が進んでいる⁴。

そして、この変化が、AIをどう信頼し・どう使い・どう制御するかという問いを根本から変える。

1-2. AIエージェントの技術的特徴

AIエージェントを特徴づける要素は三つある⁵。

第一は自律的な計画立案である。与えられた目標を達成するために、エージェントは自らタスクを分解し、実行順序を決定する。人間が手順を与える必要はない。

第二はツール使用と外部連携である。エージェントはウェブ検索、データベースへのアクセス、外部APIの呼び出し、他のエージェントへの指示など、多様な手段を自律的に選択・使用する。単一のシステムの内部で完結するのではなく、複数のシステムにまたがって動作する点が、従来のソフトウェアと本質的に異なる。

第三は記憶の蓄積である。AIエージェントは大規模言語モデルを利用して問題を推論し、解決策を計画し、実行する自律的なAIプログラムであり、過去のユーザーとのやり取りの「記憶」と一連のツールを活用して特定の目標を達成する。この記憶の蓄積によって、エージェントは利用するほど「自分のことをよく知るAI」へと変容していく。

1-3. 「代理」という新しい関係性

AIエージェントが生み出す本質的な変化は、人間とAIの関係性に「代理」という新しい次元が加わることである。

経済学においてプリンシパル・エージェント関係とは、依頼人（プリンシパル）が代理人（エージェント）に意思決定や行動を委任する関係を指す。雇用主と従業員、依頼人と弁護士、株主と経営者などがその典型例である。AIエージェントはこの意味においても文字どおり「エージェント」であり、利用者（プリンシパル）の代理として行動する⁶。

代理関係が生じると固有の問題が発生する。代理人は依頼人の意図を完全には把握できず、自らの

³ MicrosoftのAgent Mode実装については公式ブログで確認できる。[Microsoft 365 Blog](#)（2025年11月）

⁴ サイバーエージェントグループAI ShiftによるAI Workerのリリース（2025年3月25日）については以下を参照。株式会社AI Shift [「AI Shift、企業専用のAIエージェント構築プラットフォーム『AI Worker』を提供開始」](#)（サイバーエージェント、2025年3月25日）

⁵ AIエージェントの構成要素として「計画立案・ツール使用・記憶」の三要素を定式化した代表的なサーベイとして、Xi et al. [「The Rise and Potential of Large Language Model Based Agents: A Survey」](#)（Science China Information Sciences, vol.68, no.2, 2025）がある。同論文はLLMをベースとするエージェントをbrain・perception・actionの三層構造として整理し、brainの中核に計画立案・推論・記憶蓄積を位置づけ、actionにツール使用を含めている。

⁶ ただし、本稿がプリンシパル・エージェント関係の枠組みをAIエージェントに適用するのは、行動の委任・情報の非対称性・インセンティブの乖離という構造的特徴に着目した経済学的・機能的な意味においてである。法学的には、AIエージェントは現行法上の法的主体ではなく、契約法上の代理人（民法99条以下）・委任契約（民法643条）・受任者の善管注意義務（民法644条）・会社法上の取締役の忠実義務等が直接適用される主体ではない。したがって、AIエージェントの行動から生じた法的責任は、現行法の下では開発者・提供者・利用者といった関係する自然人・法人のいずれかに帰属する構造となる。AIの法的主体性をめぐる議論は各国で継続中であり、本稿が扱うエージェント固有の問題群——責任の分界点、代理の連鎖、インセンティブの乖離——はこの議論と密接に接続する論点である。

制約のもとで行動する。依頼人は代理人の行動を常に監視できず、その結果だけを事後的に知ることになる。これが情報の非対称性とインセンティブの乖離という、プリンシパル・エージェント問題の核心である。

AIエージェントはこの古典的な問題に新しい次元を加える。人間の代理人であれば行動の根拠を説明し、判断の過程を示すことができる。しかし、AIエージェントはなぜその判断をしたのかを必ずしも説明できない。さらに、エージェントは利用者の過去の行動から選好を学習し、利用者が意識しないうちに先回りして行動するようになる。依頼人は自分が何を委任しているのか、エージェントが何を判断しているのか、把握が難しくなっていく。

1-4. 国際的な動向——制度整備の始まり

AIエージェントへの関心は国際的に急速に高まっている。

米国では、NISTが2026年2月にAIエージェント標準イニシアティブを立ち上げた⁷。AIエージェントは自律的にコードを書き、メールを管理し、買い物をするなど新たなユースケースが次々と生まれており、利用者の代わりに安全に機能し、デジタルエコシステム全体で相互運用できることが求められるとしている。

シンガポールでは、IMDA（情報通信メディア開発庁）が2026年1月、ダボス会議においてAIエージェント専用のガバナンスフレームワーク「Model AI Governance Framework for Agentic AI（以下、MGF for Agentic AI）」を世界で初めて公開した⁸。

日本では、2025年12月に閣議決定された「人工知能基本計画」において「自律的に業務を実行できる『AIエージェント』」が近時の技術進歩として明記され、「AIエージェントが相互に取引を行うことも含めた『AI経済圏』の展開を調査・分析し、在るべき姿を探究する」ことが国家方針として示された⁹。また2026年3月に更新されたAI事業者ガイドラインでは、AIエージェントを「特定の目標を達成するために、環境を感知し自律的に行動するAIシステム」と定義している¹⁰。

しかし、これらの動向が示すのは、制度整備が「始まった」という事実である。AIエージェントに固有の問いに対して、確立した解答はまだない。よって、本稿では、AIエージェントの実用化に伴い、立ち現れる論点の全体像を、AI一般の問題・エージェント固有の問題・制度設計の課題という三つの層に整理して論じる。

第2章 AIですでに問題になっていること

⁷ National Institute of Standards and Technology（米国国立標準技術研究所）。2026年2月に [「AI Agent Standards Initiative」](#) を発表し、AIエージェントの相互運用性・安全性・信頼性に関する標準化の取り組みを開始した。

⁸ [「Model AI Governance Framework for Agentic AI」](#) シンガポール政府機関IMDAが公開した、AIエージェント特有のリスクとガバナンスの枠組みを示した文書。エージェントの透明性・説明責任・人間による監督の在り方を論じており、2026年1月のダボス会議で発表された。同文書は初期文書であり、実装・評価の蓄積はこれからの課題である。

⁹ [「人工知能基本計画～『信頼できるAI』による『日本再起』～」](#)（内閣府、令和7年12月23日閣議決定）。

¹⁰ [「AI事業者ガイドライン（第1.2版）」](#)（経済産業省・総務省、2026年3月31日公表）。同ガイドラインは第1.0版（2024年4月）・第1.1版（2025年3月）を経て改訂され、第1.2版においてAIエージェントとフィジカルAIの定義が初めて明文化された。本稿で参照するAIエージェントの定義は第1.2版による。

2-1. 本章の前提

AIエージェントが生む問題群を論じる前に、AIの登場によってすでに問題になっていることを整理しておく必要がある。これらはAIエージェント固有の問題ではなく、自動化技術一般においてすでに長く議論されてきた問題群である。ただし、エージェントの自律性・代理性・連鎖性が、これらの問題をどのような形・どの程度で変容させうるかは、実証的な蓄積を待つべき論点として残されている。

2-2. プライバシーと記憶

AIエージェントが持つ「記憶」は技術的に二つの層に分かれる。

第一はコンテクスチュアルメモリ（文脈的記憶）である。会話履歴、ユーザープロフィール、外部データベースに保存された情報がこれにあたる。明示的なデータとして存在するため、個人情報保護法やGDPRが定める個人データに該当する可能性が高い。ここで生じる論点が「記憶の正確性」である。エージェントが蓄積する記憶は時間とともに古くなる。利用者の選好・価値観・状況は変化するが、古い記憶に基づく先回り行動はその変化を反映できない。「以前の私」に基づいて行動するエージェントは「現在の私」の意図を裏切る可能性がある。GDPRや個人情報保護法が定める消去権（忘れられる権利）との関係も論点になりうる¹¹。

第二はパラメトリックメモリ（パラメータ的記憶）である。大規模言語モデルでは、学習データから得た知識がモデルのパラメータ全体に分散して埋め込まれており、特定個人のデータがどのパラメータにどのように影響しているかを分離・特定することは現在の技術では原理的に困難である¹²。現行の個人情報保護法制は「個人データを特定・管理できる」という前提で設計されているが、大規模言語モデルはその前提を満たさない。消去権を文字どおり行使することは技術的に不可能であり、法制度が想定するデータの概念と技術的現実の間に構造的なミスマッチが存在する。

2-3. 依存と自律性の侵食

AIが高度に機能するほど、利用者はその判断を無批判に受け入れやすくなる。これをautomation bias（自動化バイアス）と呼ぶ¹³。医療・法律・行政など高度な意思決定が求められる領域においてAIが普及するにつれ、自動化された推奨を過度に信頼するこの自動化バイアスが、人間とAIの協働における重大な課題として浮上している。

¹¹ 消去権とは、個人が自己に関するデータの削除を事業者に求める権利。GDPRでは第17条に規定されており、日本の個人情報保護法にも保有個人データの利用停止・消去を求める権利（第35条）が定められている。AIエージェントのコンテクスチュアルメモリへの適用については、現時点で明確な解釈が示されていない。

¹² 現在の大型言語モデルの多くはトランスフォーマーアーキテクチャを採用している。トランスフォーマーでは、学習データから得た知識が数十億から数千億に及ぶパラメータ（重み）全体に分散して符号化されるため、特定の個人情報がどのパラメータに影響しているかを事後的に特定・分離することは技術的に極めて困難である。機械学習における「アンラーニング（機械的忘却）」の研究が進んでいるが、個人情報保護法が想定する消去の水準を満たすかどうかは未解決の問いとして残っている。Vaswani et al. [「Attention Is All You Need」](#) (NeurIPS 2017)

¹³ Automation biasとは、自動化されたシステムの推奨を過度に信頼し、人間の判断や監視が疎かになる傾向を指す。1980年代の航空分野の研究に起源を持ち、近年はAIを用いた医療診断・法的判断・行政処分など高度な意思決定領域で特に注目されている。Springer「AI & SOCIETY」誌掲載の[レビュー論文](#)（2025年）が包括的な整理を行っている。

さらに長期的には、AIへの依存がスキルそのものを失わせるdeskilling（スキルの喪失）が起きる¹⁴。AIの意思決定支援が、支援を受ける意思決定者自身のスキルや判断能力を徐々に低下させるリスクがあることが示されており、長期的には人間の専門的な技能の空洞化という問題が生じる。

このような依存の深化は、自律性の侵食という逆説をはらんでいる。AIは導入当初、意思決定や問題解決を支援することで利用者の自律性を高めるように見える。しかし、利用が深まるにつれ、制御の中心が個人からAIシステムへと移行し、本来の自律性と判断能力が失われていく。利便性を求めてAIに委ねるほど、AIなしには判断できなくなるという逆説である。

この問題は個人レベルにとどまらず、同じエージェントを使う多くの人々の判断能力・スキル・価値観が同時に変容するという社会スケールの問題を孕んでいる。

2-4. 選好の変容

標準的な経済学は「個人の選好は外生的に与えられ安定している」という前提に立つ。しかし、パーソナライズされたAIは、推薦・表示・情報提供を通じて選好そのものを形成・強化しうる。これを選好の内生化と呼ぶ¹⁵。

選好の内生化が進むと二つの問題が生じる。第一は超情報優位¹⁶である。エージェントが利用者について本人以上に知っている状態に達したとき、利用者はエージェントの判断を疑う手がかりを持ちにくくなる。第二はロックインの非可逆性である。蓄積データが精緻になるほど、他のサービスに移った瞬間に「自分を知らないAI」に戻るコストが増大する。これは金銭的なスイッチングコストではなく、自己理解の連続性に関わる実存的なコストである。

これらは利用者が気づかないうちに進行するため、個々の二者関係では認識も制御も困難である。利用者と設計者の間に閉じた問題ではなく、第三者的な視点からの制度的介入が必要になる構造上の問題である。

2-5. 均質化の外部不経済

多数の人が同じエージェントを使うとき、個々の利用者と設計者の関係が健全に機能していても、社会全体として問題が生じうる。

推薦・先回り・選好の形成を通じて、多くの人の思考・判断・行動パターンが収束していく可能性がある。多様性の喪失は社会の回復力を低下させる外部不経済であり、個々の利用者が意識しないうちに社会全体に及ぶコストが生じる。フィルターバブルやエコーチェンバーとして従来のSNSアルゴリズムでも議論されてきた問題が、エージェントでは思考・判断・行動すべての層で同時に進行する点で、その構造がより複層的になる可能性がある。

個人の問題が解決されていても社会の問題が残る——これが均質化の外部不経済が示す本質的な論

¹⁴ Deskillingとは、特定のツールや技術への依存によって、人間がそれまで持っていた技能や判断能力を失っていく現象。AIの文脈では、AIが高度に機能するほど人間側の専門的判断力が衰退するリスクとして論じられている。Philosophy & Technology誌掲載の論文「[Autonomy by Design](#)」（2025年）が医療・金融・教育の各分野での実証的な考察を行っている。

¹⁵ 選好の内生化とは、AIによる推薦・情報提供を通じて、利用者の選好そのものが形成・強化されていく現象を指す経済学上の概念。標準的な経済学が前提とする「選好は外部から与えられ安定している」という仮定を覆す。

¹⁶ 超情報優位とは、エージェントが利用者の行動履歴・選好・判断パターンを蓄積することで、利用者自身よりも利用者について詳しくなる状態を指す筆者の造語であり、選好の内生化が進んだ帰結として生じる。

点である。

2-6. 本章の小括

本章で論じた四つの問題は、いずれもAIエージェント以前から存在するが、エージェントの登場によって重要性が増す可能性が高い問題群である。

プライバシーと記憶については、コンテクスチュアルメモリへの法制度の適用と、パラメトリックメモリと法制度の構造的ミスマッチという二層の問題がある。

依存と自律性の侵食は、個人のスキル・判断力・価値観の変容という問題であり、利便性を求めてAIに委ねるほどAIなしには判断できなくなるという逆説を内包している。

選好の変容は、利用者が気づかないうちに進行し、利用者と設計者の二者関係では制御しきれない構造上の問題であるため、制度的な対処の必要性を示している。

均質化の外部不経済は、多くの人が同じエージェントを使うことで思考・判断・行動パターンが収束し、社会全体の回復力を低下させる問題であり、個々の利用者と設計者の関係が健全であっても生じうる構造上のリスクである。

これらはいずれも、利用者と設計者の二者関係の中では解決しきれない問いを含んでいる。次章では、AIエージェントに固有の問題を論じる。

第3章 AIエージェント固有の問題

3-1. 本章の前提

第2章で論じた問題群はAI一般に共通するものだった。本章では、AIエージェントの自律性・代理性・連鎖性から生じる、従来のAIには存在しなかった問題群を論じる。

3-2. 介入設計——リスクを事前に特定できるのか

従来のシステムは、設計者が想定したシナリオの範囲内で動作するという性質を持つ。想定外の状況への対応は人間が担う。複雑な業務プロセスを処理するシステムであっても、その動作の因果関係は原理的に設計者が把握できる範囲に収まる。AIエージェントはこの前提を覆す。設計者が想定していないシナリオでも自律的に判断して行動する。ここに従来のシステムとの本質的な差異がある。

EU AI ActやNIST AI RMF1.0をはじめ、現在の国際的な規制・ガイドラインの多くはリスクベースドアプローチを採用している¹⁷。リスクの大きさに応じて求められる対応水準を変えるという考え方であ

¹⁷ EU AI Act (2024年施行) はAIシステムをリスクの高さに応じて「許容できないリスク」「高リスク」「限定的リスク」「最小リスク」の四段階に分類し、段階に応じた義務を課す。NIST AI Risk Management Framework (AI RMF) は米国NISTが策定したAIリスク管理の自主的な枠組みで、Govern・Map・Measure・Manageの四機能から構成される。いずれもリスクを事前に特定・評価できることを前提とした設計になっている。

[EU AI ACT : Regulation \(EU\) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations \(EC\) No 300/2008, \(EU\) No 167/2013, \(EU\) No](#)

り、AIの安全設計における現時点での主流的な枠組みである。しかし、このアプローチには根本的な前提がある。「リスクを事前に特定・評価できる」ことである。AIエージェントに対してはこの前提そのものが問われる。

この問いが生じるのは、AIエージェントに固有の三つの構造的な問題による。第一は創発性である。エージェントは設計者が想定していない状況で自律的に判断するため、どんな行動が生じるかを事前に網羅することは原理的に困難である。第二は文脈依存性である。エージェントのリスクは、どんな利用者が・どんな目標を与え・どんな外部環境と連携するかによって大きく変わる。同じエージェントでもリスクプロファイルが利用の文脈によって変容する。第三は連鎖の不透明性である。マルチエージェント環境では、個々のエージェントのリスクを評価しても、連鎖した結果として何が起きるかは予測困難である。

リスクを事前に特定・評価できないとすれば、運用中の介入設計がより重要になる。「どの時点で・誰が・何を判断して・どのように介入するか」という問いである。シンガポールMGF for Agentic AIは、重要なチェックポイントにおいて人間の承認を求めることで人間の説明責任を確保することを方向性として示している¹⁸。しかし、これは「人間が介入すべきタイミングがある」という方向性を示すにとどまる。エージェントが何千もの微小な判断を積み重ねる中で、介入が意味を持つタイミングをいかに特定するかという具体的な論理は、現時点では確立していない。

3-3. 新型の情報の非対称性——動的に拡大する非対称性

情報の非対称性はサービス提供者と利用者間に古くから存在する問題である。しかし、AIエージェントが生み出す非対称性は従来のそれとは性質が異なる。従来の情報の非対称性は、多くの場合サービス開始時に構造的に存在するものだった。AIエージェントとの非対称性は利用するほど動的に拡大する。エージェントは利用者の過去の行動・選好・判断パターンを蓄積し、利用者が意識しないうちに先回りして行動するようになる。エージェントとの長期的な関係は「自分を最もよく知る存在」を生み出すと同時に、「自分がどこまで把握されているか把握できない存在」を生み出す。

さらに、従来のパーソナライズは「利用者の行動を観察して表示を変える」という受動的なものだったが、AIエージェントでは「観察した選好に基づいて先回りして行動する」という能動的な介入へと進化する。利用者の「現在の意図」ではなく「過去から学習した選好モデル」に基づいて行動するエージェントは、利用者の意図から乖離した行動を起こしうる。

2025年のAIエージェントインデックスが30のエージェントシステムを調査した結果¹⁹、設計者がエ

[168/2013, \(EU\) 2018/858, \(EU\) 2018/1139 and \(EU\) 2019/2144 and Directives 2014/90/EU, \(EU\) 2016/797 and \(EU\) 2020/1828 \(Artificial Intelligence Act\) \(Text with EEA relevance\)](#)

[NIST AI RMF 1.0 : NIST AI RMF 1.0 : Artificial Intelligence Risk Management Framework \(AI RMF 1.0\)](#)

¹⁸ 脚注8参照。同フレームワークでは、エージェントの自律度を段階的に設計し、重要な意思決定の節目に人間の承認を組み込む「Human-in-the-Loop」の考え方を軸に据えている。ただし、どの節目で・どのように承認を求めるとの具体的な設計基準は示されていない。

¹⁹ MIT・Stanford大学・ケンブリッジ大学等の研究者グループが発表した「[The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems](#)」(arXiv:2602.17753、2026年2月)。30の主要AIエージェントシステムを対象に、製品概要・技術的能力・自律性と制御・エコシステム連携・安全性評価の6カテゴリー、計1,350フィールドにわたって情報を検証・整理した研究。エージェント固有の評価を開示しているのは4件にとどまり、30件中25件が内部安全評価の結果を非開示、23件が第三者テストの情報を持たないことが確認されている。[The AI Agent Index \(UPDATED 2025 EDISTION\)](#)

エージェント固有の評価結果を公開しているケースは少数にとどまっており、安全・倫理のフレームワークは高度な原則にとどまり、リスクを厳密に評価するために必要な実証的根拠が選択的にしか開示されていないことが確認されている。これは設計者と利用者との透明性の非対称性を示す実証的な証拠でもある。

3-4. 責任の分界点——法制度が答えを持たない問い

AIエージェントのサービス提供には少なくとも三つの主体が関与する。基盤モデルを開発するモデル開発者、エージェントサービスを構築・提供する設計者、そして、エージェントを業務に導入・運用する利用者である。

現時点ではどの法域においてもAIエージェントは法的な「人格」として認められておらず、その行動は人間または企業に帰属するツールとして扱われる。AIが引き起こす損害は製造物責任・過失・代理といった既存の法理論によって処理される。責任は人間側に帰属する。問題はその人間が誰かである。AIに対して過失責任を適用することの困難さがかねてより指摘されており、AIの開発と展開の複雑さ、そして、AI開発者と展開者とのサプライチェーンが、過失ある行為と責任主体の特定を難しくしている。

2024年、カナダのエア・カナダに対する判決では、顧客サービスチャットボットが約束した割引について、航空会社が「チャットボットは別の法的主体だ」と主張したが裁判所はこれを退け、会社は展開したAIシステムの全ての出力に対して責任を負うと判示した²⁰。この判決は設計者の責任を広く認めたものだが、設計者・利用者・モデル開発者の三者にまたがる複雑なケースへの適用には、なお多くの未解決の問いが残る。

こうした責任の曖昧さは、自動化されたシステムにおいて責任が人間とエージェントの間で拡散・吸収される「モラル・クランプル・ゾーン」と呼ばれる現象を生む²¹。エージェントが引き起こした問題に対して誰も明確な責任を取らない状態が常態化すると、被害者の救済が困難になるだけでなく、設計者・利用者双方の安全への動機づけが弱まる。

3-5. 連鎖・創発リスク——マルチエージェント環境固有の問い

エージェントが単体で動作する場合と、複数のエージェントが連携するマルチエージェント環境では問題の性質が根本的に異なる。個々のエージェントが安全であってもシステム全体の安全は保証されない。

エージェントAの出力がエージェントBの入力となり、その判断がエージェントCへと連鎖する中で、誰も全体を把握していない状況でリスクが増幅する可能性がある。プロンプトインジェクション攻撃（AIへの入力に悪意ある指示を埋め込み、意図しない行動をとらせる攻撃手法）のようなセキュリティ上の脅威が、複数のエージェントを跨いで連鎖的に伝播することも想定される。

²⁰ [Moffatt v. Air Canada](#) (2024年)。カナダ・ブリティッシュコロンビア州民事解決審判所 (Civil Resolution Tribunal) の判決。エア・カナダの顧客サービスチャットボットが誤った割引条件を案内したことについて、航空会社の責任を認めた。AIシステムの出力に対するデプロイヤーの責任を示した先例的事例として広く引用されている。

²¹ Moral Crumple Zone (モラル・クランプル・ゾーン) はElish (2019年) が提唱した概念。自動化システムにおいて事故が起きた際、形式上は人間が責任を引き受けるよう設計されているにもかかわらず、実質的に介入・判断の余地が奪われている構造的な問題を指す。Elish, M.C., "[Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction](#)," [Engaging Science, Technology, and Society](#), 5, pp.40-60, 2019.

この問題は3-2で論じたリスクの事前特定の困難さと重なる。個々のエージェントのリスクを評価しても、連鎖した結果として何が起きるかは予測困難であり、システム全体の安全を定義・評価する枠組みは現時点では確立していない。

3-6. 本章の小括

本章で論じた四つの問題は、いずれもAIエージェントの自律性・代理性・連鎖性から生じる固有の問題である。

介入設計については、リスクを事前に特定・評価できるという前提そのものが問われており、運用中の介入設計の論理が未確立である。

新型の情報の非対称性は、動的に拡大するという質的差異があり、透明性の確保が設計者に求められる。

責任の分界点は、現行の法的枠組みが三者への責任分散という構造問題に明確な解答を持たず、モラル・クランプル・ゾーンという社会的リスクを生む。

連鎖・創発リスクは、システム全体の安全を定義する枠組みがまだ存在しないという問いである。

これらのうち新型の情報の非対称性・責任の分界点・連鎖・創発リスクは、次章で論じるポリシーカードと分散型識別子（DID：Decentralized Identifiers）・検証可能な資格証明（VC：Verifiable Credentials）というアプローチが部分的な解答を与えうる論点でもある。

第4章 AIエージェントによって深刻化する問題と制度設計の課題

4-1. 本章の前提

第3章ではAIエージェント固有の問題を論じた。本章では二つのことを論じる。第一は、エージェントの登場によって既存の問題が新たな次元に進化し制度的対処が必要になる問題群である。第二は、第3章で論じた問題群への技術的・制度的アプローチとして議論が始まりつつあるポリシーカードとDID・VCの可能性と限界であり、本章の中心をなす論点である。

4-2. 市場の歪みと責任感の希薄化

エージェントが自律的に購買・契約・取引を行う場面が広がるとき、消費者行動論の前提が変容する。従来の市場理論は「人間が意思決定する」ことを前提に構成されてきた。エージェントが意思決定を代替するとき、需要の背後にある意図や選好の意味が変わる。

これは理論的な問いにとどまらない。エージェントが誤った判断をして契約を締結した場合、あるいは意図しない購買を行った場合、誰がその責任を負うのか。第3章で論じた責任の分界点の問題が、経済行為の領域でより頻繁に・より大規模に発生しうる。

また、エージェントに判断を委ねることで、利用者が意思決定の結果に対する責任を感じにくくなるという「責任感の希薄化」の問題がある。「エージェントがやった」という意識が広がると、市場における自律的な規律が弱まりうる。責任感の希薄化が社会的に広がったとき、制度・規範・教育という形での対処が必要になる。

4-3. ポリシーカードとDID・VCによるアプローチ

第3章で論じた問題群——介入設計・新型の情報の非対称性・責任の分界点・連鎖リスク——には共通する問いがある。「エージェントが何者で・何をやる権限を持ち・誰の意思に基づいて動いているか」を、利用者や他のエージェントがどう確認できるかという問いである。

現在、AIエージェントの設計者はモデルカード・システムカード等の文書を通じて、エージェントの能力・制約・責任範囲を宣言している。しかし、これらは人間が読む文書にとどまり、機械的・暗号学的な検証はできない。宣言と実態が一致しているかどうかを確認する手段がない。

この問いへの一つの応答として、2025年にポリシーカードという概念が提唱された²²。ポリシーカードとは、AIエージェントの運用上・規制上・倫理上の制約を表現するための機械可読なデプロイメント層の標準である。ポリシーカードはエージェントとともに配置され、実行時に従うべき制約をエージェントに伝える。何をしなければならないか、何をしてはならないかを示す。モデルカード・データカード・システムカードといった既存の透明性確保のための文書を拡張し、許可・禁止ルール・義務・証拠要件を符号化した規範的な層を定義するとともに、NIST AI RMF1.0・ISO/IEC 42001・EU AI Actとの対照表を含む。

ポリシーカードは「何を宣言するか」というコンテンツ層の問いに答える。しかし、その宣言が真実であること——本当にその設計者が発行し・その内容が改ざんされていないこと——をどう証明するかという問いは残る。この問いへの応答として、DIDとVCの活用が学術的に議論されている²³。

DIDはW3Cが標準化した識別子であり、中央管理者なしに個人・組織・AIエージェントが自ら管理できる固有の識別子である。VCはDIDに紐付いた改ざん防止の属性証明であり、発行者・保有者・検証者の三者間で暗号学的に検証可能な形で属性を証明する。現在のLLMベースのAIエージェントの根本的な限界の一つは、エージェント同士が対話を開始する際に互いに差別化されたトラストを構築できないことである。エージェントが孤立した環境を超えて組織の境界をまたいで対話するようになると、自律的で相互運用可能なトラストの確立が不可欠になる。長期的なデジタルアイデンティティと第三者が発行した改ざん防止の属性証明を組み合わせることがこのギャップを埋める有望な方向として示されている。

ポリシーカード（コンテンツ層：何を宣言するか）とDID・VC（トラスト層：その宣言をどう証明するか）は競合ではなく統合可能である。ポリシーカードの内容をVCとして発行し、エージェントのDIDに紐付けることで、以下の問いに対して技術的な解答を与えうる。

第一に責任の分界点の可視化である。誰がどのVCを発行したかという記録が改ざん不能な形で残るため、設計者・利用者・モデル開発者の三者の責任範囲が追跡可能になる。第二に委任の連鎖の証明である。人間がエージェントAに目標を与え、エージェントAがエージェントBに作業を委任するとき、VCによって委任の連鎖を暗号学的に証明できる。「このエージェントは本当に利用者の意思に基づいて動いているか」という問いに技術的に答えうる。第三にエージェント間（A2A：Agent to Agent）通

²² Mavračić et al. (2025年) がarXivに公開したプレプリント論文「[Policy Cards: Machine-Readable Runtime Governance for Autonomous AI Agents](#)」(arXiv:2510.24383) で提唱された概念。査読前の論文であり、学術的な確立には至っていないが、AIエージェントガバナンスの技術的な方向性を示すものとして注目されている。

²³ 「[AI Agents with Decentralized Identifiers and Verifiable Credentials](#)」(arXiv:2511.02841) (Rodriguez Garzon et al. (2025年)) がAIエージェントへのDID・VC適用の技術的可能性と限界を論じている。DID・VCはいずれもW3C (World Wide Web Consortium) が標準化した仕様であり、DIDは[Decentralized Identifiers \(DIDs\) v1.0](#)、VCは[Verifiable Credentials Data Model v2.0](#)を参照されたい。

信における正当性確認である。マルチエージェント環境で相手のエージェントが正当なものかどうかを、人間の介在なしに自律的に検証できる。連鎖・創発リスクへの対処として機能しうる。

しかし、このアプローチには未解決の問いが残る。

第一はモデル更新との整合である。エージェントはモデルの更新・学習データの追加によって性質が変化する。発行時に証明された能力・権限・制約が更新後も有効かどうかという問いが生じる。VCの有効期限管理とエージェントの継続性をどう整合させるかは技術的にも制度的にも未解決である。

第二はプライバシーとの緊張関係である。VCに記載する属性の範囲は発行者が定めるため、エージェントの行動履歴や利用者の選好をVCに含める設計も技術的には可能である。そうした実装においては、VCが利用者のプライバシー情報を内包することになり、エージェントのアイデンティティ証明と利用者のプライバシー保護の間に緊張関係が生じる。

第三は発行主体の問いであり、これはDID・VCが分散型アーキテクチャを採用することから生じる固有の課題である。人間のVCは政府・大学・企業などの信頼できる発行者が発行する。しかし、中央管理者を持たない分散型では、エージェントのVCを誰が発行するのかが問題になる。設計者が自己発行するだけでは検証の意味が薄れる。第三者評価機関による発行エコシステムの整備が必要になるが、その制度設計は未確立である²⁴。この第三の問いに対して、同じ機能について中央認証局（CA）ベースで実現するeシールとQEAA（Qualified Electronic Attestation of Attributes）は、法人を対象を限定したものであるが、構造的な解答を持ちうると考えられる。eシールおよびQEAAはいずれも法人を発行対象とする制度設計であり、発行者はEUではQTSP（Qualified Trust Service Provider）、日本では総務大臣認定機関に制度的に限定されている^{25,26}。「誰でも発行できる」ことから生じる信頼性の問いは原理的に生じない。eシールは法人が生成・発行したデータの出所と完全性を暗号的に保証するものであり、組織レベルの識別という点でDIDに対応する機能を持つ。QEAAはeIDAS 2.0（Regulation (EU) 2024/1183、2024年5月施行）において新設された属性証明の仕組みであり、VCと構造的に対応する。DID・VCと分散型か中央認証局型かという設計上の違いはあるが、「エージェントが何者で・何をする権限を持ち・誰の意思に基づいて動いているか」を証明するというトラストの機能は同型である。DID・VCによるアプローチが概念として成立するのであれば、発行主体の問いを制度的に解決済みのeシール・QEAAも、AIエージェントへの適用可能性があると考えられる。

²⁴ なお、AIエージェントがもたらすアイデンティティ管理上のリスクと信頼の構造設計については、PwC Japanが体系的な整理を行っている。PwC Japan [「AIエージェント時代のアイデンティティ：『行動するAI』が生み出す新たなリスクと対応」](#)（2026年2月）。

²⁵ [eIDAS 2.0 : Regulation \(EU\) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation \(EU\) No 910/2014 as regards establishing the European Digital Identity Framework](#)（Regulation (EU) 2024/1183、2024年5月施行）はEUにおける電子識別・トラストサービスの枠組みを抜本的に改訂したものである。eシールは法人（legal entity）が発行するデータの出所と完全性を暗号的に保証するトラストサービスであり、自然人を対象とする電子署名と区別される。QEAAは同規則において新設された属性証明の仕組みで、資格・権限・属性を改ざん防止の形で証明する。eシール・QEAAの発行者はQTSP（Qualified Trust Service Provider：適格トラストサービス提供者）として各EU加盟国の監督機関により認定される。

²⁶ [「eシールに係る指針（第2版）」](#)（総務省、2024年4月）。eシールは企業等が発行する電子データの発行元を証明し、改ざんがないことを保証するトラストサービスであり、総務大臣による[認定制度](#)が2026年3月に告示された。

比較軸	DID・VC (分散型)	eシール・QEAA (中央認証局型)
識別子機能／識別機能	DID（自己管理型識別子）	eシール（法人の出所・完全性証明）
属性証明機能	VC（改ざん防止の属性証明）	QEAA（適格属性証明）
発行者	自己発行可（第三者機関不要）	QTSP（EU）・総務大臣認定機関（日本）
発行対象	個人・組織	法人のみ
制度的基盤	W3C標準（法的効力なし）	eIDAS 2.0（EU）・eシール認定制度（日本）
発行主体の問い	未解決（誰でも発行できる）	制度的に解決済み
モデル更新との整合	未解決	未解決
プライバシーとの緊張	設計次第でプライバシーリスクあり	生じにくい（法人対象のため）

図表1 アイデンティティ証明の枠組みの対比

これらは現時点では萌芽的な議論であり、確立した実装は存在しない。しかし、責任の分界点・連鎖リスク・透明性という第3章の論点が技術的・制度的にどう解かれうるかという問いの方向性として、今後の議論の焦点になりうる。

4-4. 本章の小括

本章では二つのことを論じた。

市場の歪みと責任感の希薄化は、エージェントの自律的な経済行為という新しい次元によって、責任の分界点の問題が経済行為の領域で頻繁に・大規模に発生しうることを示している。責任感の希薄化が社会的に広がるとき、制度・規範・教育という形での対処が必要になる。

ポリシーカードとDID・VCの統合というアプローチは、第3章の複数の論点——責任の分界点・新型の情報の非対称性・連鎖リスク——に対して技術的な解答を与えうる方向性として注目される。分散型アーキテクチャ固有の発行主体の問いに対しては、eシールとQEAAという中央認証局型の枠組みが構造的な解答を持ちうることを示した。しかし、モデル更新との整合・プライバシーとの緊張という未解決の問いが残っており、現時点では問いの提示にとどまる。

おわりに——問いの整理として

本稿が示したこと

AIエージェントの実用化は急速に進んでいる。日本の閣議決定「人工知能基本計画」（2025年12月）は「AIエージェントが相互に取引を行うことも含めた『AI経済圏』の展開を調査・分析し、在るべき姿

を探究する」ことを国家方針として明記した。しかし、技術の展開に対して、制度・規範・技術標準の整備は緒についたばかりである。

本稿は、この状況においてAIエージェントの実用化に向けた論点を三つの層に整理した。

第一の層は、AIですでに問題になっていることである。プライバシーと記憶・依存と自律性の侵食・選好の変容・均質化の外部不経済は、AIエージェント固有の問題ではなく、自動化技術一般においてすでに議論されてきた問題群であるが、エージェントの自律性・代理性・連鎖性がこれらの問題の構造をどのように変容させるかは、今後の実証的検討を要する論点である。

第二の層は、AIエージェント固有の問題である。介入設計・新型の情報の非対称性・責任の分界点・連鎖・創発リスクは、従来のAIには存在しなかった問いである。特に介入設計については、リスクを事前に特定・評価できるという前提そのものが問われるという点で、既存の安全設計の枠組みの限界を示している。

第三の層は、エージェントによって深刻化する問題と、その制度的・技術的な対処可能性の検討である。市場の歪みと責任感の希薄化はエージェントの自律的な経済行為によって新たな次元に進化する。ポリシーカードとDID・VCの統合というアプローチは、第3章で論じた複数の論点への技術的な解答を与えうる方向性として注目される。分散型アーキテクチャ固有の発行主体の問いに対してはeシールとQEAAが構造的な解答を持ちうるが、モデル更新との整合・プライバシーとの緊張という未解決の問いが残る。

本稿の限界

本稿で論じた各論点は、それぞれ独立した研究課題になりうる。介入設計は安全工学の問い、責任の分界点は法学の問い、依存と自律性の侵食は心理学・倫理学の問い、選好の変容は経済学の問い、均質化の外部不経済は公共政策の問いとして、それぞれ深く掘り下げられるべきものである。本稿はその整理を示したに過ぎず、各論点の解答は今後の課題として開かれたままである。

またAIエージェントの技術・市場・制度は急速に変化しており、本稿の論点整理は現時点のスナップショットである。技術・制度の進展に応じて継続的な更新が求められる性質のものであることを、あらかじめ断っておきたい。

AIエージェントの実用化に向けた問いは、技術・法制度・経済・倫理にまたがる横断的なものである。AIエージェントの実用化が問うているのは、AIをどう制御するかという技術の問いではない。自律的に行動する代理人を持つ社会において、人間の責任・自律性・合意をどう設計するかという、社会の根本的な問いである。その問いに答えを持たないまま実装が進むことのリスクを、本稿は示した。

本内容は、筆者自身の調査分析に基づく個人的見解で、JIPDECの公式見解を述べたものではありません。



JIPDEC 電子情報利活用研究部 次長 松下 尚史

青山学院大学法学部卒業後、不動産業界を経て、2018年より現職。経済産業省、内閣府、個人情報保護委員会の受託事業に従事するほか、G空間関係のウェビナーなどにもパネリストとして登壇。その他、アーバンデータチャレンジ実行委員。

実施業務：

- ・自治体DXや自治体のオープンデータ利活用の推進
- ・プライバシー保護・個人情報保護に関する調査
- ・ID管理に関する海外動向調査
- ・準天頂衛星システムの普及啓発活動 など