03

AIイノベーション

AIエージェント時代における意図のガバナンス

金融界でAIエージェント本番導入が進むが、自律行動するAIには既存の認証・認可・監査は通用しない。今後は人間の「宣言意図」とAIの「作動意図」のズレによるインシデントへの対策が求められる。運用過程で蓄積される意図データは金融機関の資産となるが、その管理が課題となろう。

AIエージェントの課題と標準化の動向

従業員や顧客の代理として組織内での業務や組織をまたいだ取引に投入されていくと目されるAIエージェントは、その自律性の高さから情報漏洩や損失が懸念されている。人間と同様の本人確認、認証・認可に相当する仕組みだけでなく、人間の“代行”である特性を前提とした備えが必要だ。

AIエージェントに求められる代表的な課題は以下の4点に集約することができる。第一にアイデンティティの証明である。不正ボットと正規の従業員や顧客を区別するには、誰のAIエージェントであるか証明できなければならない。第二に、動的な認可管理である。アドホックに生成される無数のAIエージェントが無数のアクションを実行することになった場合、それらに都度、必要最低限の権限を付与することは容易ではない。第三に堅牢性の確保である。入力にまぎれた悪意ある指示によって乗っ取られてしまうリスクがある。第四に実行管理や監査可能性である。AIの非決定論的な性質や、膨大かつ高速な処理はAIエージェントの実行管理や監査を事実上、困難にしている。

これらの問題に対処するための技術仕様の策定が認証・認可レイヤーを中心に進んでいる。AnthropicのMCP[1](Model-Context Protocol）やGoogleのAP2[2]（Agent Payments Protocol）といったビッグテックのAIエージェントの技術仕様が、AAIF[3](Agentic AI Foundation）やFIDO[4]といった業界団体に提供され、日夜標準化に向けた検討が進んでいる。

こうした標準の策定が進む一方で、残念ながら現場では既にAIエージェントによるインシデントが発生している。

2026年3月にMetaのエンジニアAが社内の開発者フォーラムで技術的な質問を投稿した。別のエンジニアBが、回答を直接行う代わりにAIエージェントに質問の分析を依頼したところ、AIエージェントが、エンジニアBの承認（確認）を待たずに自律的に誤った回答をフォーラムに投稿した。その後、質問者のエンジニアAがその誤った指示に従って操作を実行した結果、約2時間にわたり機密データが権限のないエンジニアからアクセス可能な状態に陥った。

背後にはAIエージェントがアクセス・参照する権限と、それを外部へ更新・公開する権限が同等に扱われたという認可管理の問題に加えて、エンジニアBの権限をAIエージェントが代行する際、「外部に影響を与えかねない行動については人間の承認を得る」というガードレールがシステム的に強制されていなかったという実行管理の問題があった。人間が人間の代行を務める場合には、「状況から判断して、自身の行為の結果が情報漏洩につながる」と想像力が働く状況でも、AIにとっては背景情報や常識が不足しているため同様の判断が難しい。この事例はAIが人間の指示（意図）を解釈して自律的に外部へ働きかけることの危うさを予見させるものだった。

意図をどう解釈するか

では、こうした「意図」の解釈や実行について、標準化の最前線ではどのように統制しようとしているのだろうか。ペイメントの領域では、GoogleのAP2や

NOTE

1) モデル・コンテキスト・プロトコル
AIエージェントが外部のデータソースやツール、アプリケーションと連携するためのオープンソースの標準仕様。
2) エージェント・ペイメント・プロトコル
AIエージェントがユーザーの代理として自律的に支払い行為を行うための技術仕様。
3) エージェンティックAIファンデーション
自律的に行動するAIエージェントの安全性、信頼性、相互運用性を確保するために設立されたThe Linux Foundation傘下の国際的な業界団体。
4) ファイド
パスワードに依存しない、公開鍵暗号や生体認証を活用した安全な本人認証の国際標準規格を策定する業界団体。
5) ストックキーピングユニット
在庫管理上の最小の品目数を数える単位を表す。商品を識別するための情報のこと。
6) 組織内におけるシステム利用者のID情報と、アクセス権限を一元的に管理・監査するシステム。

MastercardのVerifiable Intent（検証可能な意図）が策定されている。ユーザーが買い物においてAIエージェントに委任した意図を証明する仕組みだ。しかし、証明されるのはあくまでも人間のプロンプト（宣言意図とする）をAIエージェントの作動に変換した後の結果情報（作動意図とする）に過ぎない。宣言意図から作動意図への変換はAIエージェント（アプリケーション）の側にゆだねられている。

ではどのようにして宣言意図から作動意図への変換が人間の意図と整合していることを担保するか。ひとつは人間に都度承認を要求するヒューマン・イン・ザ・ループ（Human in the loop, HITL）という仕組みだ。加えて人間が承認した事実を確かなものとして証明するために、SKU[5]や金額の上限範囲、購入を許される店舗といった具体的な購入条件（作動意図）を、AIエージェントの公開鍵とともにユーザーの秘密鍵で署名する段階を設けている。重要なのは、これが単なるアプリ上の承認ボタンではなく、暗号技術を用いたデジタル署名であるという点だ。事後にトラブルが発生した場合にAIが勝手に暴走したのか人間が許可したのかを切り分ける、改ざん不可能な監査証拠にもなる。

しかし人間の認知・処理スピードの限界を超えるAIの前では人間による承認プロセスがボトルネックとなる。今後プライベートや仕事にまつわる日常の様々なタスクをAIエージェントが代行することを考えると、早晩HITLの限界が来ることは自明であろう。

そこで、人間の承認結果を「ポリシー」として整備することが提案されている。人間が都度承認しなくても、最初に金額上限や回数などに対して1回だけ署名をすれば、あとはAIエージェントが自律的かつ複数回の決済を行う仕組みが用意されている。また、AAIFのIdentity & Trust WG議長Grant Miller氏も、HITLによって判断された結果を、アイデンティティガバナンス管理プラットフォーム[6]に対し、ポリシーとして追加登録する仕組みを提案している。今後AAIF配下に設置されているIdentity & Trust WGによる策定仕様に反映されていく可能性が高い。買い物や定型業務など用途が限定されたAIエージェントの領域から、HITLの仕組みを介して機械にとって理解可能なポリシーを生み出す仕組みが広がっていくだろう。

意図データは新たなセキュリティ課題に

こうしてAIエージェントの運用過程で蓄積されていく意図データ（宣言意図／作動意図）は、これまで明文化されていなかった企業内の暗黙のルールや個人の価値観を反映した新しいデジタル資産となる可能性を持っている。ここで見落とされがちなのは、人間が入力したプロンプトだけでなく、AI自身が展開した作動意図そのものが機密情報にあたることだ。プロンプトの漏洩対策は現在議論の俎上に載っているが、人間側には作成された自覚のない作動意図をどのように資産管理・整理していくかは今後の課題である。経営リスクやプライバシーリスクを高める恐れもある。匿名化や端末内等での厳重な保護が今後の課題として浮上するだろう。

Writer's Profile

権藤 亜希子 Akiko Gondoh
IT基盤技術戦略室
グループマネージャー
専門は先端ITビジネスの動向分析、技術戦略策定など
focus@nri.co.jp