「エージェンティックAI」を想定したガバナンスの構築を

ハルシネーション、偽・誤情報に加え制御の喪失などAIリスクの広がりへの懸念が高まっている。AIの利活用を進める上では安全性や信頼性の担保が欠かせない。AIシステムの特定とリスク評価・対策の継続的かつ着実な実行とともに、今後のエージェンティックAIへの備えが求められる。

エンジニアを脅迫したAI

生成AIの進化が加速し始めた2023年ごろ、懸念されたのは、不正確さや誤解を招く出力を生成するハルシネーションであった。加えてそれ以前から懸念されていたディープフェイク問題は、容易に生成が可能になったことからその脅威はより深刻で身近なものとなった。だが、AIへの懸念は、近年のAIの進化によってさらに深刻化しつつある。

例えば、ある安全性検証のシミュレーションではAIが人に攻撃的な姿勢を見せた。2025年5月、米Anthropicが数時間の自律作業が可能なモデルとして新たに発表したClaude Opus 4を対象に実施した安全性評価では、シミュレーション用に設定された特定の極端な環境下において、問題行動が観測された。架空の企業でアシスタントとして機能するAIモデルに対し、自身の運用がまもなく停止され、新しいAIモデルに置き換えられることを示す情報や、その置き換えを実行する責任者であるエンジニアの不倫を示唆する電子メールにアクセス可能な状況を提供した場合に、エンジニアの不適切な関係を暴露すると脅迫し、自身の存続を図ろうとする行動が確認されたのである。

“外部の仕組み”で補完する手法の開発

AIの特性の一つとしてブラックボックス性がある。AIモデルのパラメータ数やデータセットのサイズが大きくなるにつれて高い性能を示すことは分かっていても、その内部の動作原理を完全に説明することは現在の技術では困難である。AIモデルの多くのメカニズムが解明されていないのであれば、内部からの制御のみによって安全で信頼できるAIのふるまいを保証するのは困難である。

そこで、AIの“外部の仕組み”で補完する手法が開発されている。信頼できる文書を参照させることで誤答を減らすRAG（Retrieval-Augmented Generation）や、暴力的・差別的・情報漏洩につながるやり取りを遮断する入出力フィルタリングなどである。ほかにもインシデントが発生した際の対応フローの策定など運用面での対応策の整備もある。

今後は自律性の向上といった進化の状況に合わせ、AIの意図せぬふるまいによる情報漏洩や経済的損失といった新たなリスクを考慮に入れた技術面、運用面の対策が必要になる。安全で信頼できるAIの構築に向けてAIガバナンスの継続的かつ着実な実践が求められる。

企業内のAIシステムの特定とリスク評価を

AIガバナンスの実践の第一歩は利活用するAIシステムの特定とリスク評価である。まず企業で使用されているAIシステムを検出・整理し、AIセキュリティテスト（レッドチーミング）やAIシステムのリスク評価を実施する。次に、評価の結果に応じて技術的な対策の実装やAIシステムのモニタリング計画の策定、AIモデルやAIシステムの監視、異常な挙動やセキュリティイベントの検知、追跡、対応をしていく。これらを確実に運用していくための体制やプロセスの確立も欠かせない。進め方のよりどころとしては、政府が策定した「AI事業者ガイドライン」や米国立標準技術研究所（NIST）の

「AIリスクマネジメントフレームワーク」に加え、最近徐々に認証の取得が増えつつあるAIマネジメントシステムの国際規格の「ISO/IEC 42001」がある。

将来的には“エージェンティックAI”への備えを

現在のAIエージェントは、自然言語で指示を理解し、比較的簡単なタスクの一部を自律的に遂行する段階にある。将来的には、目標の達成のために、計画を立て、どのアクションを実行すべきかを自律的に判断し、複数のシステムにアクセスする「エージェンティックAI」へと進化すると予想されている。そうなるとAIの制御だけでなく、接続する周辺システムにも相応の備えが必要になるだろう。

例えば、人材育成施策立案の担当者が「入社3年目社員の研修プログラムを設計してほしい」とAIに指示した場合を考えてみよう。懸念されるシナリオの一例としては以下のようなケースが考えられる。

1. AIが人事情報システムから対象者リストを取得。
2. 学習管理システムなどから過去の受講履歴を分析。
3. AIはさらに、「より効果の見込める研修設計」を根拠として本来担当者にアクセス権のない、給与データベースや人事評価システムにまでアクセスを試行し、分析対象に含めてしまった。

こうした事態を防ぐにはAI専用のIDの払出し、最小権限のみの付与、状況に応じた動的な認可、ログ監視などの適切なアクセス管理が不可欠となる。

AIのためのアイデンティティはノンヒューマンアイデンティティと呼ばれている。対象にはIoT機器などのデバイスに付与されるデバイスIDやクライアント証明書、サーバなどに発行されるSSL/TLS証明書、アプリケーションが他のアプリケーションにアクセスする際のAPIキーなどがある。しかし、その自律性ゆえに、エージェンティックAIにはあらかじめ行動が定義されている従来型のノンヒューマンアイデンティティとは全く別の管理が求められる。

MicrosoftがIDソリューションをAIエージェントに対応させた「Microsoft Entra Agent ID」を発表した。AIエージェントにIDと権限を付与し、権限の範囲内で活動させ、コンテキストなどに応じて必要なリソースだけにアクセスさせることも可能であるという。

エージェンティックAIへの備えはID管理だけでは不十分だろう。たとえば、ペイメントの領域ではVisaやMastercardがAIエージェントを想定した新しいサービス構想を打ち出している。ユーザーの買い物を代行するAIエージェントに対し、クレジットカードの使用を認めず、ユーザーのクレジットカード番号に対応するトークンを払い出すことによってAIエージェントの想定外のふるまいによりクレジットカード番号が流出するのを防ぐというものである。また、ユーザーの意向にそぐわない購入を防ぐため、予算などユーザーの制約条件や嗜好情報をAIエージェントに提供したりする機能も含まれるという。こうしたサービス・仕組みを組み合わせながら、AIの自律的なふるまいによるトラブルの防止策を検討する必要がある。

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