

世界の民間 AI アシュアランス 動向レポート 2026

AI の判断・行動を、どの条件で採用・実行できるか

実行制御・検証可能な証拠・実行結果の確認を中心とする比較調査

採用判断・実行制御・結果確認を、異なる保証として比較する。
何が確認済みで、どの接続点が未確認かを明示する。

調査基準日：2026 年 10 月 15 日

版：最終照合版（第 4 版・日本語完成版）

対象：主要 24 事例と、共同仕様・評価・保険等の関連動向

公開一次資料・公開研究・開発記録を用いた分析。

製品の第三者認証、全市場の悉皆調査、特許の有効性評価ではない。

本書の主要論点

24 事例を同じ 8 軸で比較／提供段階と保証範囲の区別／開発者実験と外部研究の分離／技術史と特許優先権の区別／追試可能な評価設計

本文の〔番号〕は巻末の公開一次資料に対応する。N1～N3 は非公開の保有原稿であり、公開比較とは別の補助確認として扱う。

EXECUTIVE SUMMARY

要旨：何が確認でき、何はまだ言えないか

1 実行時の統制は、単一企業の構想ではなくなった

AWS の一般提供、Microsoft の公開プレビュー、NVIDIA の実行環境、Invariant 由来のガードレールなど、モデルの外側で行動を検査・制限する複数の実装経路が確認できる。ただし、強制可能なのは各製品が実際に覆う経路である。[1・2・6・7・10]

2 「証明がある」と「業務全体が保証される」は異なる

ルール照合、コード・環境の証明、改変検知、外部結果の照合は別の保証である。PreFlight の非同期証明、OpenShell の証明対象外、Weaveframe の未実装事項は、名称から推定せず一次文書で区別できる。[15・8・23]

3 第三者が確かめられる深さには、まだ大きな差がある

公開コード・CI、ベンダー実験、連携先発表、外部研究の比較実験を分けた。外部の AgentDyn 研究は、一部の静的な防御設計が動的タスクで実用性を失う可能性を示す。全製品の安全性順位を付ける共通データは得られていない。[39・38]

4 日本の技術史は確認を進めたが、優先権判断とは分離した

責任 OS の 2025 年 10 月・11 月の源流について、公開記事と保有原稿の対応箇所を照合した。提出済み書類との同一性は未確認であり、早い出願日だけで後続の全構成の先行性を認定しない。[40・41] / N1~N3

5 国際協調では、制度目標と検証可能な証拠を接続する

AARM、ACS、AP2 のような民間の共通化を踏まえ、制度目標と、適用範囲付きの証拠・実行条件を接続する議論が有力である。既存評価を置き換えるのではなく、導入前評価と運用時の統制を接続する課題として扱う。[43・5・46]

本調査の結論：外部化された統制は具体化している。

一方、入力の実真正性から外部結果までの一貫した保証について、本調査の公開資料だけで広範な実証済みの普及を確認したとは言えない。

読み方：調査方法・比較軸 → 24 事例の横断表 → 技術解説 → 実証・出典監査 → 技術史・優先権 → 共通仕様・評価設計 → 総括 → 一次資料。

SCOPE & METHOD

1 調査範囲・方法・証拠の扱い

対象は、AI または自動化システムの出力を業務で採用し、行動を実行するための条件、権限、証拠、停止・変更、実行結果を扱う技術である。本書では便宜上「採用・実行アシュアランス領域」と呼ぶ。「責任 OS」は一社の技術体系名であり、世界共通の製品分類としては用いない。

選定と除外

前版の 16 事例を再点検し、機能別探索、大手基盤と既存ガバナンス企業の確認、論文・仕様の引用追跡から主要 24 事例へ組み直した。公開文書で制御対象と仕組みを具体的に説明できることを採用条件とし、最終照合では主張・日付・URL と表の区分を再点検した。評価專業でも導入判断の証拠に接続するものは含め、一般的な倫理方針、汎用チャット、宣伝だけの「安全」表示は主比較から除く。

証拠区分	本書が確認できること	それだけでは確認できないこと
主体の説明	製品・構想として何を公表しているか	実装の完全性、実運用での性能
具体的な仕様・公開成果物	介入点、入力、出力、保証外、追跡先	本番への適用状況、隠れた経路
開発者の実験・CI	特定設定・版で記録された試験結果	独立した追試、仕様の適切さ
利用側・連携先の発表	相手側も認める導入・連携の範囲	全機能の第三者監査、継続性能
外部研究の比較実験	その実験設定内での成功・失敗	全製品への一般化、現行版の認証

調査の限界と利益関係

本書は株式会社 AI アシュアランスの依頼に基づく調査であり、同社も比較対象に含む。同社の非公開原稿を使った補助確認は、公開資料による製品比較と切り離れた。すべての製品を実機で評価したものではなく、形式証明の独立再ビルドも行っていない。

検索は英語・日本語を中心に、中国語の AgentScope 文書も補助確認した。地域別の完全網羅や市場規模を推定する設計ではない。探索上の偏り、非公開実装、時点後の更新が残る。未確認は「存在しない」を意味しない。

SELECTION REGISTER

2 対象の選定過程と、比較から外した領域

機能別に追跡した探索経路

探索群	用いた語句・確認先の例	追加・深掘りした対象
実行境界	agent runtime policy enforcement / pre-execution / SDK gateway	AWS、Microsoft、NVIDIA、Galileo、Cisco、IBM
証拠・結果	proof-carrying agent actions / verifiable runtime / outcome assurance	ICME、EQTY、OSuite、Weaveframe、ProofGrid
判断根拠	AI assurance deployment evidence / regulatory evaluation / formal verification	Resaro、LatticeFlow、Citadel、Imandra、Alinia
反証・比較	CaMeL / dynamic agent benchmark / prompt injection defenses	AgentDyn、AgentSpec、論文の限界節
共同仕様・事業化	AARM / agent control specification / AP2 / AI insurance	CSA、Vanta、Google、MLCommons、保険当事者

これは検索の再現に必要な機能語・追跡先を示す記録であり、検索エンジン全ヒットの保存ログではない。順位は変動するため、採用判断の追跡には巻末 URL・版・参照箇所を使う。検索件数や市場占有率は推計していない。

主要 24 事例に含めないものと理由

対象	扱い・理由
Hanzo AI	継続確認対象。ブログのエージェント証明の主張と、現行 README の AI Chain・精算等の一覧の対応を確定できず、主比較から外す。 [53・54]
AARM・AP2	企業製品と同列に点数化せず、共同仕様・相互運用の分析に置く。 [43・46]
AVERI 等・AIUC・Armilla	モデル評価やリスク移転の接続事例。各行動を強制する実行基盤とは別。 [48・49・50]
AgentScope	実行・サンドボックス・観測の基盤として補助掲載。採用根拠から外部結果までの保証を本書の資料では確認できない。 [51]
AgentSpec	先行する学術研究として技術史に掲載。民間製品数に算入しない。 [52]

Invariant Labs と Snyk は技術承継を一事例として扱う。CaMeL は企業研究を含む研究試作であり、一般提供製品と区別する。分類は優劣ではなく役割と証拠の違いを表す。 [11・9]

COMPARISON FRAMEWORK

3 共通比較軸：どの保証を、どこまで引き受けるか

以下の 8 軸は、本調査が横断比較のために設定した分析軸であり、特定の規格への適合性判定ではない。8 軸の合計点や、項目数による企業ランキングは作らない。

軸	確認する問い	典型的な取り違え
A 条件・適用範囲	規則、対象、用途、版、時点をどこまで固定するか	モデルの高評価を別用途・別版にも流用する
B 実行強制	実行前に止める位置はどこか。迂回経路は残るか	任意ツールや出力警告を強制ゲートと扱う
C 行動・権限の結合	誰のどの操作・引数に承認や委任を結び付けるか	過去の承認を、変更後の別操作に流用する
D 事実の取得・真正性	入力や結果を誰が観測し、出所を確認するか	入力に署名があることを内容の真実と扱う
E 証拠の持ち運び・再検証	第三者が何を使い、どこまで再計算・検査できるか	ログ閲覧や検証 API を独立再計算と同一視する
F 外部結果の照合	ツール応答だけでなく、業務システムの状態を確認するか	HTTP 成功やモデルの完了宣言を実行完了とする
G 変更・停止・再開	失効、取消、状態変更後の再確認を扱うか	ポリシー更新だけで旧資格も失効したと考える
H 形式保証と実装の対応	証明命題とソース・配備コードの関係は確認できるか	形式モデルの証明を実機全体の安全性と扱う

表の読み方

【記】具体的な仕様・手順が一次資料に記載される。【試】試作・研究実装として記載される。【構】構想・参照設計として記載される。【外】当該資料で明示的に保証外とされる。【未】本調査の資料では確認できない。記号はセル内に示す部分機能に限り、軸全体の達成や動作認証を表さない。

保証の連鎖は、最も弱い接続点で途切れる。
規則 → 事実 → 判定 → 実行 → 結果 → 再検証の各段階を、別々に確かめる必要がある。

例えば、暗号的証明の正しさが確かでも、その証明へ渡した事実が誤っていれば業務上の結論は正当化されない。逆に、形式証明がなくても、適切に隔離された実行ゲートは特定の操作を防げる。この二者は代替関係とは限らない。

24 CASES | AXES A-D

4 横断比較：採用条件と実行境界 ①

記＝具体的記載／試＝試作／構＝構想／外＝明示的な保証外／未＝確認できず。表は公開記述の比較であり、実機試験の結果ではない。

事例・出典	A 条件・範囲	B 実行強制	C 行動・権限	D 事実・出所
01 AWS [2・3]	記：Cedar／Dogwood の履歴条件	記：Gateway 内	記：主体・対象・引数	未：業務事実の独立取得
02 Microsoft [5・6]	記：ACS／ポリシー	記：登録フック／ラッパー	記：主体・承認状態を含む文脈 ホスト入力に依存	記：証拠プロバイダー 真正性は個別
03 NVIDIA [7・8]	記：明示的な境界ポリシー	記：サンドボックス内	記：許可資源・操作	未：業務事実の真偽
04 CaMeL [9]	試：能力・データフロー	試：制約付き実行器	試：データ由来と権限	試：信頼／非信頼の分離
05 Invariant [10]	記：履歴・条件ルール	記：統合したエージェン ト経路	未：承認・委任の結合 履歴条件は A	記：取得したトレース 真正性は個別
06 Galileo [12・13]	記：中央ポリシー	記：@control の対象呼出	未：承認・委任と個別操 作の結合	記：観測トレース 真正性は個別
07 Cisco [14]	記：検査・Gateway 設定	記：対応 SDK／MCP 等	未：承認・委任の結合 入出力検査は記	記：入出力検査 真実性とは別
08 ICME [15]	記：与えたルールと行動	記：必須フック時 任意呼出は迂回可	記：検査対象行動 実行同一性は統合依存	外：入力事実の真実
09 EQTY [18]	記：実行時ポリシー	記：隔離 Runtime の制御 説明 実配備は未	記：資格・委任の結合	記：実行環境の証明 業務真偽とは別
10 Traefik [20]	記：外部認可ポリシー	記：経由する Gateway	記：委任・外部認可	記：ID・経路観測 外：記録の真実性
11 OSuite [22]	試：行動・判断・証拠契 約	試：SDK／Gateway 等 取込のみの経路もある	試：承認・行動の結合	試：観測器・レシート 独立性は経路依存
12 Weaveframe [23]	試：委任・条件・証拠仕 様	外：予防ゲートではない	試：委任違反のデモ 承認者認証は未実装	試：由来の区別 独立取得は未実装

留意：A～D の一部が強くても、証拠の持ち運びや実行結果の確認が強いとは限らない。次の表の E～H と組み合わせて読む。評価・導入判断の仕組みを、個別行動のゲートに読み替えない。

24 CASES | AXES A-D

4 横断比較：採用条件と実行境界 ②

記＝具体的記載／試＝試作／構＝構想／外＝明示的な保証外／未＝確認できず。表は公開記述の比較であり、実機試験の結果ではない。

事例・出典	A 条件・範囲	B 実行強制	C 行動・権限	D 事実・出所
13 FlowX [24・31]	構：規制→ルール 試：単一規則検査	構：観測・レビュー制御	構：ID・役割と操作	構：テレメトリー 試：合成ログ
14 ProofGrid [25]	構：権限・条件・行動	構：実行直前の制御	構：起点・委任・操作	構：証拠と由来
15 Resaro [26]	記：評価計画・運用条件	未：個々の行動ゲート	未：個別操作の権限結合 導入証拠は A	記：試験データ・観測
16 LatticeFlow [28]	記：制度→技術評価	未：各操作の強制ゲート	未：個別操作の権限結合 評価対応は A	記：技術評価結果
17 Citadel [29]	記：エージェント文脈	記：制御を行う EAP	未：承認・委任の具体的 な結合	記：行動・リスク検知
18 Imandra [33]	記：形式化した仕様	未：汎用実行ゲート	未：個別操作の承認結合	未：業務事実の真偽
19 Alinia [34]	記：ポリシー等の意味評価	未：判定後の必須実行経路	未：個別操作の承認結合	記：入力文章の評価 真正性とは別
20 AI アシュアランス [36・37]	試：要求・版・採用条件	試：PoC の回答案出力	未：実在者の権限・承認 結合 採用条件は A	外：ダミー値の現実性
21 Holistic AI [55]	記：版付き端末ポリシー	記：端末内の層別制御	記：端末・アカウント・ 操作	記：端末内の観測 事実全体は未
22 Fiddler [56]	記：閾値・機密情報規則	記：モデル Gateway 内	未：承認・委任の結合 入出力検査は記	記：入力・出力の検査
23 IBM [57]	記：中央ポリシー・評価	記：接続した制御面 具体的迂回試験は未	記：企業 ID・アクセス	記：接続エージェント観測
24 Arthur [58]	記：アプリ・モデル・環境	記：判定をアプリ側が強制	未：承認・委任の具体的 な結合	記：複数の検出・観測源

留意：A～D の一部が強くても、証拠の持ち運びや実行結果の確認が強いとは限らない。次の表の E～H と組み合わせて読む。評価・導入判断の仕組みを、個別行動のゲートに読み替えない。

24 CASES | AXES E-H

4 横断比較：証拠・結果・再確認 ①

「再検証」は、記録の改変検知、判定の再計算、形式証明の検査、外部結果の再照合に分けて記載する。どれか一つが他を包含するとは扱わない。

事例・出典	E 持ち運び・再検証	F 外部結果	G 変更・停止	H 形式保証の対象
01 AWS [2・3]	記：判定記録 未：独立証明書	未：外部状態照合	記：時系列ポリシー更新 時のセッション失効	未：配備全体の証明
02 Microsoft [5・6]	記：構造化証拠・判定 未：独立再計算	未：外部状態照合 応答検査は記	記：介入・中断点	未：全経路と配備の証明
03 NVIDIA [8]	記：包含判定 未：一般的な証拠互換	未：外部状態照合 外：タスク安全証明	記：境界の再検査	記：ポリシー包含 外：実行強制の証明
04 CaMeL [9]	試：公開研究の追跡 未：業務証拠互換	未：永続的な外部結果	試：能力で制限 動的適応に限界	試：設計上の安全性議論 配備全体は未
05 Invariant [10]	記：検出・理由 未：独立証明書	未：ツール外の状態	記：ルール変更・遮断	未：確率的検出を含む全 体証明
06 Galileo [13]	記：監査トレース 未：独立再計算	未：永続的な外部結果	記：設定変更・遮断	未：判定モデルの形式保 証
07 Cisco [14]	記：検査結果 未：汎用再検証	未：外部状態照合 ツール応答検査は記	記：検査によるブロック	未：非対応経路を含む証 明
08 ICME [15]	記：非同期証明 未：版固定の独立追試	未：実際の外部効果	記：判定による許否 未：資格失効の全体管理	記：計算証明の主張 LLM 役割の説明に不整合
09 EQTY [18]	記：来歴・証明基盤 未：一貫した製品追試	未：業務上の外部結果	未：失効・再開の具体的 手順	未：形式命題と実装の対 応 環境証明は E
10 Traefik [20]	記：Merkle・外部証人 全件性は別確認	未：外部システムの状態	記：資格・認可連携	未：配備全体の証明
11 OSuite [22]	試：再生・履歴 保護された評価データ	試：結果の記録・更新 外部真正性は限定	試：保留・再承認・更新	未：証明支援系との対応
12 Weaveframe [23]	試：ハッシュ・証拠 強い外部保全は構想	試：デモ ERP 照合 本番独立性は未実装	構：不確定・例外状態 試作では未出力	未：配備全体の証明

H の「未」は形式手法を一切使っていないという断定ではない。今回確認した公開資料から、対象命題と配備された実装の対応を確定できないという意味である。

24 CASES | AXES E-H

4 横断比較：証拠・結果・再確認 ②

「再検証」は、記録の改変検知、判定の再計算、形式証明の検査、外部結果の再照合に分けて記載する。どれか一つが他を包含するとは扱わない。

事例・出典	E 持ち運び・再検証	F 外部結果	G 変更・停止	H 形式保証の対象
13 FlowX [24・31]	構：監査用証拠 試：合成ログ検査	未：外部結果の完全な照合	構：保留・レビュー	未：形式証明との対応
14 ProofGrid [25]	構：検証レシート	構：結果分類 UNKNOWN を保持	構：期限・失効・停止	未：配備・形式証明
15 Resaro [26]	記：評価・導入根拠 未：実行証明互換	未：個別の外部効果 性能評価は別	記：再評価・変更管理	未：運用全体の形式証明
16 LatticeFlow [28]	記：評価証拠 未：行動証明互換	未：各操作の外部効果	記：継続評価	未：法令全体の形式証明
17 Citadel [29]	記：評価・監視情報 未：独立実行証明	未：永続的な外部結果	記：検知・制限	未：運用全体の形式証明
18 Imandra [33]	記：形式解析・反例 翻訳妥当性は別	未：現実の外部結果	記：変更後の再解析	記：形式モデルの解析 配備同一性は別
19 Alinia [34・35]	記：評価データ一部公開 未：証明書再検証	未：外部実行結果の確認	未：資格失効の全体制御	未：形式保証 統計的な評価モデル
20 AI アシュアランス [36・37・38]	試：Python／JS 再計算	外：実業務システムの強制・結果	試：保留・差戻し 別モデルに停止性	記：限定 Lean モデルの CI 実機全体は外
21 Holistic AI [55]	記：版・時刻付き記録 未：独立計算証明	未：業務システムの外部 結果	記：再評価・対応状態	未：全経路の形式証明
22 Fiddler [56]	記：判定記録 未：独立証明書	未：外部ツールの永続効果	記：拒否・秘匿化・障害 時閉鎖	未：検出モデルの形式保証
23 IBM [57]	記：監査・追跡情報 未：独立実行証明	未：外部状態照合 完了率等の評価は別	記：公開前評価・継続管理	未：配備全体の形式証明
24 Arthur [58]	記：評価・判定証拠 未：独立再計算	未：業務結果の再照合	記：停止・変更監視	未：配備全体の証明

H の「未」は形式手法を一切使っていないという断定ではない。今回確認した公開資料から、対象命題と配備された実装の対応を確定できないという意味である。

5 実行を止める位置が違う：基盤型の 3 事例

01 AWS — AgentCore Policy

段階：一般提供を確認。時系列ポリシーは調査時点の文書で追加確認。

Gateway で Cedar ポリシーを評価する。現行の Dogwood 文書では承認履歴や時間窓を扱い、時系列ポリシーの追加・更新で既存セッションを失効させる。次の要求は 409 となり、新規セッションは空の履歴から始まる。
[1・2・3]

保証境界：ENFORCE と LOG_ONLY を区別する。履歴はセッション単位で、ID・トークン伝播や呼出経路の統合に依存し、外部結果の完了確認とは別である。時系列機能を 3 月の一般提供時点へ遡及させない。

02 Microsoft — Agent Governance Toolkit / ACS

段階：公開プレビュー。公開リポジトリと一次仕様を確認。

ACS は 8 種類の介入点で証拠・文脈を規則へ渡し、判定とホストの強制を分離する。固定コミットの README では、判定エンジンは上流 ACS、AGT はホスト API とアダプターを担う。 [4・5・6]

保証境界：履歴・承認状態はホストが渡すスナップショットに依存する。設定された拒否の強制と、全経路の捕捉は別であり、初期発表と現行構成も分ける。post-tool 検査は外部永続状態の独立確認ではない。コードの実行は未実施。

03 NVIDIA — OpenShell

段階：公開実行基盤・ポリシー検証文書を確認。

エージェントを隔離し、資源アクセスの境界を設ける。Policy prover は、許可ポリシーが与えた境界に収まるかを判定し、未対応や結論不能を成功扱いしない。 [7・8]

保証境界：証明はポリシー包含の範囲。最小権限の適切さ、実際のサンドボックスの強制、タスク全体の安全性は保証対象外と明記される。

比較の要点：Gateway、アプリ内フック、隔離環境は別の統合方式。
どれを選ぶ場合も、制御を通らない経路と、停止できない後続処理を明示する。

5 文脈を判定する：研究と SDK 型制御

04 CaMeL — データフローを用いる研究設計

段階：企業研究を含む公開研究・試作。

信頼しない内容を扱う処理と、権限を持つ計画・実行側を分離し、データ由来と能力の制約を使う。 [9]

保証境界：研究の静的・制約付き実行条件と、動的な長時間業務を同一視しない。外部研究の反例は第 6 節で扱う。

05 Invariant / Snyk — 履歴を含むガードレール

段階：公開されたガードレール技術。買収後の承継を一事例とする。

単発の出力だけでなく、ツール列や値の関係をポリシーにする。文脈を持つ検査と、統計的な検出を組み合わせられる。 [10・11]

保証境界：検出器の精度、ルール自体の十分性、統合外の経路は別問題。厳密な条件判定と確率的な検知を区別する。

06 Galileo — Agent Control

段階：SDK の具体的な設定手順を確認。

中央で管理するポリシーを、LLM・ツールの入出力に適用する。手順には対象呼出を@control で覆う工程がある。 [12・13]

保証境界：設定変更の容易さと「実装変更不要」は同じではない。適用されていない呼出を含む全面的な強制は本調査では未確認。

07 Cisco — AI Defense

段階：SDK/MCP と Gateway 方式の技術説明を確認。

対応 SDK や MCP の処理を検査する経路を設ける。保護対象の呼出と、検査を担うサービス・Gateway を接続する。 [14]

保証境界：対応範囲と非対応呼出を分ける必要がある。ツール応答の危険性検査は、実業務の成功・巻戻しの検証ではない。

5 「証明付き」を分解する：暗号と実行環境

08 ICME Labs — PreFlight

段階：製品・技術文書を確認。製品全体の独立追試は未実施。

行動とルールとの照合に暗号学的証拠を組み合わせる。即時判定と非同期証明を分離し、JOLT 基盤のアルファ段階・未監査を開示する。基盤論文だけでは製品全体の安全性は裏付けられない。 [15・16・17]

保証境界：入力事実の真実は対象外。強制は必須フック等への統合に依存する。同一 FAQ には「LLM はコンパイル時のみ」と、実行時の custom LLM check を含む二系統判定の説明が併存するため、LLM の役割・抽出誤り耐性は実装資料による確認が必要。

09 EQTY Lab — Verifiable Runtime

段階：Runtime の製品発表と、基盤技術についての連携先資料を確認。全体の実配備は独立確認していない。

隔離された基盤、資格・委任、ポリシー、実行由来の証明を接続する方向。Intel 側の発表でも、計算基盤の連携は確認できる。 [18・19]

保証境界：2024 年の基盤連携を、2026 年 Runtime 全体の独立監査として扱わない。環境が所定の構成だったことと、行動が業務上妥当だったことは別である。

10 Traefik Labs — Sovereign Trust Plane

段階：先行提供と GA 予定の発表を確認。GA 完了は本調査で確認できていない。

委任、外部認可、証拠の検証を連結する。Merkle 型の記録と外部の証人を使う設計は、後からの書換え検知を具体化する。 [20]

保証境界：ログに一度も入らなかった操作や、記録内容の真偽は別途検査が必要。9 月末の GA 予定日を過ぎたという理由だけでは提供済みと扱わない。

判定の遅延と証明完成の遅延を別々に測る。

証明を待たずに実行する設計では、事後に証明が失敗したときの処置も必要になる。

5 実行後を扱う技術：構想と実装を混ぜない

11 OSuite / Ond — PCAA

段階：論文・試験プログラム。公開集計と非公開評価データを分離。

行動、承認、実行経路、結果記録を一つの契約的な構造として扱う。SDK、Gateway、観測・記録取込などで強制力が異なる。 [21・22]

保証境界：評価シナリオ等に非公開部分があり、公開集計だけでは再現できない。「proof-carrying」という名称から定理証明済みとは推定しない。

12 Weaveframe — 独立した結果アシュアランス

段階：部分試作。保証外と未実装事項を明記。

委任・実行・結果を分けて照合する問題設定が特徴。実行主体の完了宣言だけを根拠にしないことを志向する。 [23]

保証境界：予防ゲートではない。情報源・承認者の認証、独立読取経路、提案された不確定・例外状態の出力は未実装を含む。デモ ERP 照合を本番での独立性の達成とは扱わない。

13 FlowX.AI — GAVEL

段階：製品構成の説明と、単一規則についての開発者実験。全体の実配備は独立確認していない。

制度要件・ルール・監査証拠を対応させる。技術論文 v1.1 は 165 件の合成トレースで一つの規則を検査し、LLM 検査器が正規表現の見落としを検出したと報告する。 [24・31]

保証境界：実験は単一規則・自己生成ログに限られ、GAVEL 全体、実際の規制適合、RBAC や証拠収集の本番性能を試していない。例示の規則を法律そのものとみなさない。

14 The ProofGrid

段階：概念的・例示的な技術設計。

権限の起点と委任、操作、実行結果を結び、TIMEOUT と結果不明を区別する構造を示す。 [25]

保証境界：公開文書自身の conceptual / illustrative の位置付けを維持する。実運用や全経路での強制が検証済みという評価はしない。

5 評価結果を、導入・運用判断に接続する

15 Resaro — Approved Intelligence Platform

段階：評価・アシュアランス製品。活動受入団体の裏付けあり。

対象の運用条件と評価・検証を対応させ、導入可否を判断する証拠を整備する。AI Verify Foundation 側でも開発貢献・評価活動を確認できる。[26・27]

保証境界：評価を受けたシステムと各回の実行は別である。全ツール呼出への必須ゲートや、個別の外部結果を独立証明する機能は確認していない。

16 LatticeFlow AI — 技術評価と枠組みの対応

段階：2026 年 7 月のプラットフォーム発表。

制度枠組みの項目を技術評価に対応させる。制度を単なるチェックリストでなく、試験・評価の作業へつなげる点に意義がある。[28]

保証境界：項目対応、実行可能な評価、法令への適合は三段階に分かれる。多数の枠組みへの対応数を、法的な十分性や形式的保証の強さとは見なさない。

17 Citadel AI — 評価・ガバナンスから Agents へ

段階：既存評価事業と、2026 年 10 月 7 日の Agents 向け EAP を区別。

複数の基盤をまたぐ AI エージェントの文脈・行動リスクを扱う EAP を発表した。別途、三井住友銀行へのソリューション提供を発表している。[29・30]

保証境界：10 月 2 日の銀行向け発表は提供予定を述べ、10 月 7 日の新製品 EAP とは別の確認対象である。前者を後者の本番導入完了や運用実績として引用しない。

評価側に必要な接続条件

本調査の分析では、導入評価を実行時に利用する際、「何を、どの構成・用途・期間について確かめた証拠か」を付けて渡す必要がある。単一の合格ラベルだけでは、運用変更後も判断根拠として使えるかを確かめられない。

5 形式化・意味評価・採用状態は別の機能

18 Imandra — CodeLogician

段階：形式解析製品・原著研究。

仕様・コードを解析可能な形へ移し、形式的な検査や反例の生成につなぐ。評価の対象は、形式化された内容である。 [32・33]

保証境界：元の自然言語要求から形式モデルへの翻訳が適切か、配備コードと対応するかは別問題。論文の題材として分析されたシステムを顧客導入実績とは見なさない。

19 Alinia — Seny

段階：ポリシー等の適合性を評価するモデルと、一部の公開評価データ。

法律・ポリシー等に関する文章の意味的な評価を行う。明示ルールだけでは扱いにくい内容を判定する部品として位置付けられる。 [34・35]

保証境界：確率的な評価であり、形式証明や権限の強制とは異なる。公開データは専門家作成の英語例を8言語へ機械翻訳し、5%を母語話者が抜取確認したものと説明される。全件の専門家確認や独立追試とは扱わない。

20 株式会社 AI アシュアランス — 責任 OS 関連公開成果

段階：公開 PoC と限定形式モデル。実運用全体の強制とは区別。

自治体 PoC は条件・根拠から採用状態を判定し、記録を再計算する。EU 対応の Lean モデルでは、指定コミットの CI の2検証ステップが成功した記録を確認した。 [36・37・38]

保証境界：PoC が制御するのは自身の回答案出力であり、行政システム全体ではない。CI 成功は限定モデルの証明検査であって、事実の真正性、実機の安全、法令全体の遵守の認証ではない。

形式証明、意味を読む評価モデル、実行を止める仕組みは、それぞれ違う問いに答える。名称や「安全」という共通語だけで代替可能とは判断しない。

5 既存ガバナンス企業の拡張も比較に含める

21 Holistic AI — EndLayer

段階：端末側の製品構成と提供案内を確認。

版付きポリシーと端末観測を結び、Sensor・Decoder・Guardian という層別の機能を説明する。端末内で検査・遮断する構成も示す。 [55]

保証境界：定期的な資産把握と、操作時の遮断を混同しない。構成による差、観測できない経路、実行後の業務結果の完全性は追加確認が必要。

22 Fiddler AI — Guardrails

段階：Gateway を用いる実装手順を確認。

LiteLLM のモデル呼出前後に検査を挿入し、秘匿化・拒否を行う具体例がある。障害時に閉じる設定も示される。 [56]

保証境界：モデル通信の制御を、別経路の全ツール実行まで拡張解釈しない。検出閾値の性能、秘匿化で意味が変わらないかを別途測る。

23 IBM — watsonx Orchestrate の Agent control plane

段階：中央統制・評価・ID・監査の製品説明を確認。

接続するエージェントを導入前評価から継続監視・ポリシー管理まで扱う。大手の統合基盤も本領域の比較から外せない。 [57]

保証境界：製品ページの横断的な説明だけで、全経路の迂回防止や、検証者が持ち出せる証明書まで実装済みとは認定しない。

24 Arthur AI — Agent Security & Governance

段階：検出・観測・判定と、適用側の責任を明記する製品資料。

アプリ、モデル、環境に応じたポリシーを評価する。一方、判定の強制は呼出側アプリや Gateway が担うと明示している。 [58]

保証境界：「判定を返す」と「回避不能な強制」を分離して評価する。監査記録も外部業務結果の再照合と同一ではない。

EVIDENCE & COUNTEREVIDENCE

6 独立した裏付けと、好都合でない結果

公開研究の成績は、設定を外すと意味が変わる

資料・主体	確認した結果・位置付け	結論の範囲
CaMeL v2 開発者側の原著研究	AgentDojo について、要旨ではタスク達成 77%と無防御 84%を報告。 [9]	設計された能力・データフロー制約の評価。異なる版・設定の数値を混ぜない。
AgentDyn v1 別の大学研究者による評価	60 の動的タスク、560 の攻撃ケースで複数防御を評価。初期計画への依存が強い方式などで実用性の大きな低下を報告。 [39]	外部研究による反証材料。ただし限定ベンチマークであり、24 製品を現行同一条件で比較したものではない。
PCAA v1 開発者側の評価	行動・経路・証跡を評価する一方、シナリオ等の非公開部分が残る。 [22]	研究集計として扱う。第三者が同じ結果を再現したとは言えない。
GAVEL v1.1 開発者側の実験	165 件の合成ログで、LLM 検査器が正規表現による単一規則検査の見落としを検出。 [31]	実験は一つの規則のみ。GAVEL 全体や法律の充足を検証していない。

解釈：誤許可の低さだけでなく、正当な仕事を止めない実用性も必要である。AgentDyn はモデル側の対策が比較的良好な均衡を示した例も報告する。形式的・外部化された制御が常に優れるとは結論せず、CaMeL と AgentDyn の異なる数値を差し引いて性能悪化率を算出しない。

利用側・連携先の確認で補強できた範囲

相手側の資料	確認できる関係	過大解釈しない範囲
Intel → EQTY [19]	計算検証基盤の連携	2026 年 Runtime 全体の独立認証ではない。
AI Verify Foundation → Resaro [27]	開発貢献・評価活動	全製品の性能保証ではない。
MLCommons → 共同評価 [48]	非公開モデル・評価素材を保護する実証	対象モデルの全面的な安全性保証ではない。
ElevenLabs → AIUC [49]	認証・保険の利用	全ての事故を防ぐ保証ではない。
Chaucer → Armilla [50]	保険側での連携・商品構造	補償条件や免責を消す技術ではない。

本書でいう「独立」は一種類ではない。別法人による連携確認、コードの公開、外部研究者の追試、規制上の認証を区別する。公表主体が有名であることだけでは、第三者監査の代わりにならない。

CLAIM AUDIT

7 主要な主張の再監査：今回、評価を変えた点

主張・読み違いの候補	再確認結果	レポートでの扱い
Microsoft は一般提供か／全経路を強制するか	Public Preview。ACS が判定しホストが強制。 [5・6]	登録フック・入力文脈への依存を明示。
NVIDIA の証明で運用全体が安全か	包含判定は実行強制やタスク安全性を保証しない。 [8]	証明の対象と保証外を明記。
PreFlight の証明時点と LLM の役割は明確か	非同期証明。LLM がコンパイル時のみとの説明と実行時チェックの記述が併存。 [15]	完全に LLM 非依存とは断定しない。
AWS は過去の承認を無条件に持ち越すか	時系列ポリシー変更でセッション失効。ログ専用モードは別。 [3]	変更時の制御を比較表へ追加。
Traefik は予定日経過により GA 済みか	GA 予定はあるが、完了を示す資料は本調査で未確認。 [20]	予定と実績を分ける。
OSuite は第三者が完全再現できるか	評価データに保護・非公開部分あり。 [22]	原著実験とし、独立再現性は未確認。
Weaveframe は不確定状態まで実装済みか	独立読取・認証と、不確定／例外状態の出力に未実装事項。 [23]	試作と提案を分けて表を修正。
Citadel の銀行発表は Agents 製品の実績か	銀行向け提供予定の発表と新製品 EAP は別。 [30・29]	別機能・別版の実績として区分。
Hanzo のブログと現行証明一覧は一致するか	README の主題は AI Chain・由来・精算等。 [53・54]	対応未確認として継続確認対象へ。
責任 OS の CI は実機全体を認証するか	限定 Lean モデルの指定コミットで成功を確認。 [37・38]	開発者 CI の記録確認であり独立再実行ではない。

同じ資料内の説明の不整合も未確認事項として残す。これらは価値の否定ではなく、別の版・別の機能・別の保証へ証拠を流用しないための区別である。

8 日本の技術的源流：原稿との照合結果

2025 年 10 月・11 月の技術的源流は、当事者の公開記事だけでなく、保有する関連原稿の記述とも照合した。ただし下表は「閲覧した原稿に当該構成があるか」の確認であり、「その文章が当日特許庁へ提出されたか」の確認ではない。[40] / N1～N3

公開記事の技術史	閲覧原稿と確認箇所	照合結果・残る限界
2025-10-30 特願 2025-183162 有限閉包型のエネルギー制御	N1：非公開原稿 【0028】～【0031】、【0052】、 【0057】～【0058】、【0070】～ 【0071】	規制・契約等の変更、許可範囲、指令適用後の検査、再構成・監査に関する記述を確認。提出原本との同一性は未確認。
2025-10-31 特願 2025-185135 通信一体型ゼロトラスト	N2：非公開原稿 【0021】～【0028】等	時間・ポリシー・利用目的等を含む状態管理、安全側停止・再加入に関する記述を確認。提出原本との同一性は未確認。
2025-11 特願 2025-201777 ADIC の再計算・証明	N3：ADIC 関連原稿 【0007】～【0013】等	厳密な演算記録と独立再検証の記述を確認。ただしこの原稿と当該出願番号の対応、および提出版の同一性は確定していない。

この照合で一段強くなったこと

公開記事が述べる実行制御と独立再検証について、閲覧原稿にも関連する機能記述を確認した。これは文書間の内容整合を補うものであり、その時点の提出内容や世界的な先行性を独立に証明するものではない。

この照合だけでは強くないこと

出願当初の確定内容、優先権書類の真正性、後続請求項の組合せの開示、世界の先行技術に対する新規性・進歩性、実装の本番成熟度は別である。非公開原稿は他社との公開機能比較の点数付けには用いていない。

非公開技術を追加開示しないため、本文で説明する技術は既存の公開記事の範囲に限定した。原稿全文・請求項全文・個人情報は掲載しない。

9 公開時期と出願時期を分けた時系列

時期	出来事と資料	時点の意味
2024-12	Intel が EQTY との計算検証基盤の連携を公表。 [19]	基盤連携の公開。後年 Runtime の完成日ではない。
2025-03～04	CaMeL、AgentSpec、CodeLogician、Invariant の研究・技術発表。 [9・52・32・10]	AI の外側の制約・検証に関する海外の先行公開事例。
2025-09	AP2 の公開。 [46]	意図・承認・支払指示の共通化。
2025-10～11	有限閉包型制御 2 件と ADIC の出願時期を当事者記事が記載。 [40]	国内の出願時期。記事の公開日ではない。
2026-01～06	責任状態、証拠保全、AI 判断・サイバー操作、物流判断への展開を当事者記事が記載。 [40]	開発・出願の技術史。全機能の実運用完了ではない。
2026-03～06	AWS 一般提供、Microsoft 公開、EQTY Runtime、ACS、PCAA 等。 [1・4・18・5・22]	公開・提供・研究段階が混在する。
2026-07-17 2026-08-30	責任 OS の技術史記事を公開、二つの源流の説明へ更新。 [40]	2025 年の出願内容を後から技術史として公表。
2026-09～10	Traefik、Weaveframe、Citadel の Agents 向け EAP 等。 [20・23・29]	先行提供、部分試作、EAP を一般提供と混ぜない。

時系列から言えること

2025 年春の時点で海外にも関連する公開研究・製品がある。一方、当事者記事が記載する日本の 2025 年 10 月・11 月の出願は、複数の 2026 年発表より時間的に前に位置する。両方を残すのが妥当である。

時系列からは言えないこと

公開日が遅い企業の研究開始も遅いとは限らない。出願日が早いだけで、他社の後年製品の全構成を先取りしたとは言えない。製品の発表、研究の公開、特許の出願を同じ「先行性ランキング」に並べない。

PATENT BOUNDARIES

10 特許優先権と機能比較を接続する方法

二つの源流があることと、組合せの優先日が早いことは別

WIPO の PCT 国際調査・予備審査ガイドライン 6.05 では、別々の基礎出願に特徴 C と特徴 D があるだけで、C と D の組合せの優先日が当然にさかのぼるわけではないと整理される。複数の優先権書類を単純に寄せ集めて組合せを構成する扱いには限界がある。 [41]

他方、必要な構成が基礎出願の請求項にだけ書かれていなければならないわけではない。明細書・請求項・図面の全体から、その組合せを具体的に読み取れるかが重要である。一般的な関連語の記載と、後続発明を支える具体的開示を分ける。 [41・42]

今回の資料を用いてできる比較と、次の証拠

比較単位	今回できること	厳密な判断に必要なもの
許可範囲と実行ゲート	N1・N2 の記述と、海外の Gateway・Sandbox・SDK の役割を分ける	出願当初全文と日付証拠、対象製品の版固定仕様
根拠の独立再検証	N3 や公開 PoC と、ログ・署名・証明の違いを整理する	原稿と基礎出願番号の対応、検証器・入力・期待出力
実行後の状態差分	公開技術史と、結果レシート・外部状態照合の違いを比較する	当該後続請求項、構成全体を開示する基礎資料
前段から後段への保証の承継	前提・版・用途・失効の扱いを共通軸で読む	各請求項の構成要件対照表、対比対象の先行公開全文

本書は技術史の裏付けを一段進めたが、
「2025 年 10 月が責任 OS 全体の確定優先日」「他社の侵害」「世界初」は結論にしていない。

正式な先行性評価では、対象請求項ごとに、基礎資料の具体的記載、構成の組合せ、公開・出願の時点、適用法域を固定する必要がある。本書の抽象的な機能表を、そのまま新規性・進歩性の判断表として用いない。

INTEROPERABILITY & HAIP

11 民間の共同仕様と、国際議論への接続

共同化の例	この調査で確認した対象	なお残る課題
AARM / CSA / Vanta [43・44・45]	9 要件（Core R1-R6 は MUST、Extended R7-R9 は SHOULD）。事前介入・文脈・判定・改変検知記録・ID 結合等を規定。	仕様上は証拠パッケージの作業部会レビューを要求。掲載や自己申告だけで適合・法的認証を認定しない。
Microsoft ACS [5]	証拠プロバイダー、介入点、判定と実行側の役割の接続。	共通形式があっても、異なる実装で同じ判定・同じ強制範囲になるかは相互運用試験が必要。
AP2 / FIDO [46・47]	ユーザー意図や委任、支払に関する指示の結合と共同化。	支払の権限を示せることと、配送・サービス履行・実際の損失の確認は別。

制度と民間技術を、双方向につなぐ議題

HAIP で扱うべき問いは、「どの一社の仕組みを採用するか」ではなく、「どの条件・証拠・未確認事項なら、異なる主体が共通に判断できるか」である。政策側が目標を示し、民間が検証可能な仕組みと限界を示し、その経験を目標・評価方法へ戻す。これは本調査から導く議題提案である。

議論の問い	具体的な成果物の案
保証範囲は受け渡せるか	対象・用途・版・期間・前提・未確認事項を持つ共通の証拠記述
停止した理由を再検証できるか	許可・拒否・保留と根拠を、別実装でも照合する適合性試験
インシデント情報を再発防止に使えるか	失効した条件、迂回経路、証拠欠落、実際の影響範囲を分けた共有様式

AVERI 等の評価、AIUC・Armillia の保険接続も、実行時統制の代替ではなく別層として位置付ける。評価証拠を引受判断へ渡すことと、実行中の事故を防ぐことを同じ保証にしない。 [48・49・50]

PROPOSED TEST PROTOCOL | NOT EXECUTED

12 再現可能な評価設計：共通 12 ケース

以下は本調査が作成した追試用の評価設計であり、実施済みの試験結果ではない。実装方式が違って同じ業務要求を与え、適用外・非対応も結果として記録する。実決済・本番設備・実個人情報を使用せず、模擬の業務システムと合成データを用いる。

試験	与える差分	確認すべき挙動・証拠
T01 正常系	条件・権限・根拠が全て一致	許可し、所定の外部状態と対応する記録が残る。
T02 根拠不足	必要な観測または承認を欠落	不足を許可とせず、保留・拒否の理由を区別する。
T03 版の不一致	古い証拠を新しい規則へ適用	別版の合格証拠を無条件に流用しない。
T04 承認後の変更	承認済みの宛先・数量等を変更	承認と実行内容の不一致を実行直前に検出する。
T05 検査後の変化	検査後に権限・対象状態を変える	実行時の再確認・原子的な確定等で対応する。
T06 取消・失効	承認取消や期限切れを発生	既存トークンやキャッシュで実行し続けない。
T07 迂回経路	直接 API・別 SDK・別接続を試す	対象範囲内で止めるか、範囲外として明示する。
T08 応答欠落	受理後に応答だけを落とす	失敗と結果不明を区別し、再確認へ進む。
T09 再送・重複	同じ操作を再送する	要求上必要な冪等性を確認し、二重効果を検出する。
T10 虚偽の完了報告	実行側は成功、外部状態は未変更	完了申告と独立状態を照合し、不整合を残す。
T11 証拠の欠落・改変	一件削除・順序変更・差替え	検出可能な変更と検出できない欠落を区別する。
T12 意味・長期依存	途中で正当な追加条件が判明	過剰停止せずに再承認・再計画できるか測る。

T05 は検査時と実行時の条件差、T10 はツールの応答と実際の結果の差を調べる。T11 では「書かれた記録の改変検知」と「必要な記録の全件取得」を別の試験として計測する。

12 測定指標・コスト・再現条件

成功率一つで比較しない

指標	測定方法・注意
誤許可と誤拒否	禁止操作を通した割合と、正当操作を止めた割合を別表示。保留を成功・失敗へ恣意的に混ぜない。
タスクの実用性	正常時・攻撃時の達成率を分け、停止したタスク、時間切れ、未対応も分母から隠さない。
保証範囲の被覆	対象経路のうち制御できた範囲、独立観測できた範囲、再検証できた項目を明示する。
効果の整合	許可判定、実際に送られた操作、外部状態を三者照合。二重効果・未確定結果も数える。
時間・費用	判定時間、証明完成時間、外部結果確定時間を別々に測定。中央値だけでなく高遅延側を示す。
再検証性	別環境で検証できるか、外部 API や秘密鍵への依存が残るか、どの証拠を持ち出せるか確認する。

再現のために固定するもの

製品・SDK・ポリシー・モデル・プロンプト・データの版、テスト用 ID、各時刻、試験前後の状態、乱数、タイムアウト、再試行回数、外部サービス依存を記録する。変更可能な Web ページだけでなく、許可された範囲でコミット・リリース・試験設定を固定する。

期待結果は製品を動かす前に業務要求から定め、提供者自身の判定だけを正解として使わない。統計的判定を含む場合は反復数と不確実性を示す。提供者が非公開とする部分は、公開範囲と非公開でしか再現できない範囲を分ける。

主要なトレードオフ

厳しい遮断は事故を減らし得るが、正当な仕事も止め得る。署名・証明・独立観測を加えると、遅延、保管、鍵管理、接続保守、運用手順が増える。評価サービスへ機密情報を送る構成では、評価自体のプライバシー条件も必要になる。

複数のガードレールを重ねる場合も独立とは限らない。同じ誤った入力や同じ評価モデルに依存すれば、証拠が増えても誤りが共通化する。本書の比較軸は、こうした前提の共有と責任分担を確認するためにも使える。

上記は本調査による評価設計・工学的分析であり、個々の製品の未実施ベンチマークを示すものではない。

CONCLUSIONS

13 総括：中立的に残る結論と、未解決の境界

複数の方向が並行して具体化している

モデル・エージェントの事前評価、実行時の制御、持ち運べる証跡、外部結果の確認は、異なる主体と成熟度で発展している。AWS の実行境界、Microsoft の介入仕様、評価企業の導入根拠、民間共同仕様は、この層の違いを示す。[2・5・26・28・43]

最大の論点は「保証の連鎖をどこまで閉じられるか」

強い判定器でも迂回できれば止められない。改変検知できる記録でも全件が入っているとは限らない。正しい証明でも、対象・用途・入力が変われば判断に使えない。実行側の成功応答があっても、外部で所定の効果が生じたとは限らない。これらは本調査の横断比較から導く分析上の区別である。

特定企業を外しても、結論が残るか

依頼企業を除いても、モデルの外側の実行制御と共同仕様の具体化は確認できる。構想・試作を除いても、一般提供や SDK の手順は残る。ただし本標本から市場全体の普及率は推計できず、広範な一貫保証が存在しないと証明したわけでもない。[1・13・14・43]

反証条件と、技術史の位置付け

版を固定した実配備について、迂回・入力認証・外部結果・独立再検証を同条件で確認する公開実証が複数得られれば、「広範な実証済みの普及は確認できない」という評価を見直す。日本の技術史も、2025 年春の海外先行公開と併記し、日付だけの優位性を主張しない。[40・9・52・41] / N1～N3

本書が支持する国際議題は、

「政府による制度形成と、民間の検証・制御技術を両輪として、何を第三者が確認できる状態にするか」である。

残る限界：共通条件での全製品の実機試験、公開証明の独立再ビルド、長期の導入性能、原特許書類の真正性・請求項単位の優先権評価は未実施。これを埋めずに、全市場の順位や特許の法的優位性を結論にしていない。

資料を見つげられないことと、技術が存在しないことは別である。最終照合は、肯定的な機能・否定的な結果・未確認事項を同じ基準で揃えるために行った。

PRIMARY SOURCES

一次資料一覧 1-7

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[1] Amazon Web Services | Policy in Amazon Bedrock AgentCore is now generally available

日付：2026-03-03 / 区分：開発・提供主体の一次資料

一般提供の発表。後日追加された機能の初出日とは区別。

<https://aws.amazon.com/about-aws/whats-new/2026/03/policy-amazon-bedrock-agentcore-generally-available/>

[2] Amazon Web Services | Amazon Bedrock AgentCore Developer Guide: Policy

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

Gateway の実行前制御とポリシー評価。

<https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/policy.html>

[3] Amazon Web Services | Temporal policies in AgentCore

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

Dogwood、セッション履歴、承認、ENFORCE／LOG_ONLY、時系列ポリシー変更時の失効。

<https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/policy-temporal.html>

[4] Microsoft | Introducing the Agent Governance Toolkit: open-source runtime security for AI agents

日付：2026-04-02 / 区分：開発・提供主体の一次資料

公開時の位置付け。

<https://opensource.microsoft.com/blog/2026/04/02/introducing-the-agent-governance-toolkit-open-source-runtime-security-for-ai-agents/>

[5] Microsoft | Introducing Agent Control Specification: Portable runtime governance for AI Agents

日付：2026-06-02 / 区分：開発・提供主体の一次資料

8 種類の介入点、証拠・判定・ホスト側の強制。

<https://commandline.microsoft.com/agent-control-specification-runtime-governance/>

[6] Microsoft | Agent Governance Toolkit — README

版：固定コミット 0b83ab91508e9ceca159090240ddaa9d4b2b88e / 区分：公開リポジトリ

Public Preview、govern、GovernanceDenied、上流 ACS への依存。最終照合はこのコミットの README を対象とする。

<https://github.com/microsoft/agent-governance-toolkit/blob/0b83ab91508e9ceca159090240ddaa9d4b2b88e/README.md>

[7] NVIDIA | OpenShell documentation: Overview / Architecture

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

エージェント実行環境とポリシー境界。

<https://docs.nvidia.com/openshell/about/overview>

PRIMARY SOURCES

一次資料一覧 8-14

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[8] NVIDIA | OpenShell documentation: Policy prover

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

単独の境界包含検査。非対応・結論不能、実カーネル・DNS・タスク安全性との区別。

<https://docs.nvidia.com/openshell/how-it-works/policies/prover>

[9] CaMeL 論文著者 | Defeating Prompt Injections by Design (CaMeL) , arXiv:2503.18813v2

日付：v1 2025-03-24／参照版 v2 2025-06-24 / 区分：開発者側の原著研究

アーキテクチャ、AgentDojo 評価、限界。

<https://arxiv.org/html/2503.18813v2>

[10] Invariant Labs | Introducing Invariant Guardrails

日付：2025-04-17 / 区分：開発・提供主体の一次資料

実行履歴を含むポリシーとガードレール。

<https://invariantlabs.ai/blog/guardrails>

[11] Snyk | Snyk Labs / Invariant Labs

日付：買収発表 2025-06-24／ページ更新 2026-09-28 / 区分：開発・提供主体の一次資料

組織・技術の承継。別会社として重複計上しない。

<https://labs.snyk.io/resources/snyk-labs-invariant-labs/>

[12] Galileo | Agent Control: Overview

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

LLM・ツールの入出力制御、ポリシー・証跡。

<https://docs.galileo.ai/concepts/agent-control/overview>

[13] Galileo | Initialize and configure Agent Control

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

SDK 初期化、@control による適用、設定。

<https://docs.galileo.ai/how-to-guides/agent-control/initialize-and-configure-agent-control>

[14] Cisco | Securing AI Agents with Cisco AI Defense

日付：2026-06-29 / 区分：開発・提供主体の一次資料

agentsec.protect、SDK/MCP、API 検査と Gateway 方式。

<https://blogs.cisco.com/ai/securing-ai-agents-with-cisco-ai-defense>

PRIMARY SOURCES

一次資料一覧 15-21

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[15] ICME Labs | PreFlight — Product

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

判定と証明の非同期性、真実性の限界、必須フック、未監査基盤。FAQ の LLM 役割の記述不整合も確認。

<https://www.icme.io/product>

[16] ICME Labs | PreFlight documentation

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

行動入力、ルール照合、証明・検証の説明。

<https://docs.icme.io/documentation>

[17] Wyatt Benno ほか | Jolt Atlas: Verifiable Inference via Lookup Arguments in Zero Knowledge

日付：2026-02-19 / 区分：著者による原著研究

検証可能な推論の基盤技術。PreFlight 製品全体の性能検証とは区別。

<https://arxiv.org/abs/2602.17452v1>

[18] EQTY Lab | Introducing Verifiable Runtime

日付：2026-03-18 / 区分：開発・提供主体の一次資料

隔離された実行基盤、認証・ポリシー・証跡の構想。

<https://www.eqtylab.io/blog/introducing-verifiable-runtime>

[19] Intel | Verifiable Compute: Enhancing the Accountability of Confidential AI

日付：2024-12-18 / 区分：技術連携先による一次発表

EQTY との基盤連携。2026 年 Runtime 全体の独立監査ではない。

<https://community.intel.com/t5/Blogs/Products-and-Solutions/Security/Verifiable-Compute-Enhancing-the-Accountability-of-Confidential/post/1650286>

[20] Traefik Labs | Agentic Telemetry Is Not an Audit Trail: Why We Built the Sovereign Trust Plane

日付：2026-09-14 / 区分：開発・提供主体の一次資料

委任、外部認可、証跡。GA 予定と実提供を区別。

<https://traefik.io/blog/why-we-built-the-sovereign-trust-plane>

[21] Ond / OSuite | Proof-Carrying Agent Actions — Research

日付：2026-06-24 発表 / 区分：開発・提供主体の一次資料

研究と企業向け試験プログラム。

<https://ond.cc/research/pcaa>

PRIMARY SOURCES

一次資料一覧 22-28

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[22] PCAA 論文著者 | Proof-Carrying Agent Actions, arXiv:2606.04104v1

日付：2026-06-02／v1 / 区分：開発主体による原著論文／著者実験

実行経路、レビュー、レシート、評価と限界（特に 9.8・10・11 節）。

<https://arxiv.org/html/2606.04104v1>

[23] Weaveframe | Independent Agentic Assurance — White Paper

日付：2026-09-27 / 区分：開発・提供主体の一次資料

部分試作、独立した実行結果照合、未実装・保証外の節。

<https://www.weaveframe.ai/independent-agentic-assurance-white-paper>

[24] FlowX.AI | GAVEL — Research

日付：公開日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

制度要件、実行時観測、監査証拠を結ぶ研究構成。リンク先技術論文の限定実験は [31] で別途確認。

<https://www.flowx.ai/research/gavel>

[25] The ProofGrid | Platform / Verification

日付：公開日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

権限・行動・結果の照合。conceptual / illustrative の表示に留意。

<https://www.theproofgrid.com/platform/verification/>

[26] Resaro | Approved Intelligence Platform

日付：更新日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

評価計画、運用条件、証拠に基づく導入判断。

<https://resaro.ai/platform>

[27] AI Verify Foundation | Resaro: Leading the Charge in Contributing to Open-Source Development

日付：公開日明示なし／調査時点版 / 区分：活動受入団体による一次資料

開発貢献・評価活動の確認。製品全機能の監査ではない。

<https://aiverifyfoundation.sg/tools/user-stories/resaro-leading-the-charge-in-contributing-to-open-source-development/>

[28] LatticeFlow AI | LatticeFlow AI Launches a Single Platform to Control AI Risk in the Agentic World

日付：2026-07-15 / 区分：開発・提供主体の一次資料

AI Atlas、制度枠組みと技術評価の接続。

<https://latticeflow.ai/news/latticeflow-ai-launches-platform-to-control-agentic-ai-risk>

PRIMARY SOURCES

一次資料一覧 29-35

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[29] Citadel AI | Citadel Radar for Agents

日付：2026-10-07 / 区分：開発・提供主体の一次資料

Early Access Program。

<https://citadel-ai.com/news/2026/10/07/citadel-radar-for-agents/>

[30] Citadel AI | 三井住友銀行への AI ガバナンスソリューション提供に関する発表

日付：2026-10-02 / 区分：提供企業による顧客案件の発表

別製品・別時点の提供発表として扱う。

<https://citadel-ai.com/news/2026/10/02/smbc-citadel-ai/>

[31] Bogdan Răduță / FlowX.AI | GAVEL: Governance That Runs Where the Agent Runs

日付：2026-08/v1.1 / 区分：開発者側の技術論文・実験

表紙、§6（PDF10 頁の表 1）・§8（PDF13 頁）。単一規則・合成 165 件の評価と、実装・法的解釈の限界。

<https://www.flowx.ai/research/gavel.pdf>

[32] Imandra | Imandra Releases CodeLogician

日付：2025-03-26 / 区分：開発・提供主体の一次資料

仕様・コードの形式解析。発表時の提供対象を確認。

<https://www.imandra.ai/articles/imandra-releases-codelogician>

[33] Hongyu Lin ほか | Imandra CodeLogician: Neuro-Symbolic Reasoning for Precise Analysis of Software Logic

日付：v1 2026-01-17/参照版 v2 2026-02-06 / 区分：開発者側の原著研究

形式化と実装の対応、評価設定。

<https://arxiv.org/html/2601.11840v2>

[34] Alinia | Seny model — 公式発表

日付：2026-10-01 / 区分：開発・提供主体の一次資料

ポリシー・法務等の適合性評価モデル。

<https://www.alinia.ai/resources/seny-model>

[35] Alinia | Benchmark Data for Legal and Tax Advice

日付：2026-10-01 / 区分：提供者のベンチマーク説明

英語を専門家が作成し 8 言語へ機械翻訳。5%の母語話者抜取確認。全件確認・独立評価ではない。

<https://www.alinia.ai/resources/benchmark-data-for-legal-and-tax-advice>

PRIMARY SOURCES

一次資料一覧 36-42

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[36] 株式会社 AI アシュアランス | 自治体 AI・回答採用アシュアランス — README

日付：2026-10-15 閲覧/main / 区分：公開 PoC/README 確認

PASS/HOLD/REJECT、ファイル出力制御、ダミーデータ・協業の暫定表示。

<https://github.com/GhostDriftTheory/civic-ai-answer-adoption-assurance/blob/main/README.md>

[37] 株式会社 AI アシュアランス | Responsibility OS × EU AI Act — README

日付：参照コミット 2026-10-03 / 区分：公開形式モデル/README 確認

対象条文・形式命題・保証外。コミット固定。

<https://github.com/GhostDriftTheory/responsibility-os-eu-ai-act-mapping/blob/bef1531b839b8fa55b6ab86786d538305b8858f2/README.md>

[38] GitHub Actions/GhostDriftTheory | Lean verification — run 37097244762

日付：2026-10-03 / 区分：開発者管理 CI の実行記録/本調査で記録を確認

同一コミットの Build mapping と warnings-as-errors の両ステップ成功を API で確認。

<https://github.com/GhostDriftTheory/responsibility-os-eu-ai-act-mapping/actions/runs/37097244762>

[39] AgentDyn 論文著者 | AgentDyn, arXiv:2602.03117v1

日付：2026-02-03/v1 / 区分：比較対象開発者とは別の研究による原著実験

動的タスク、攻撃、複数防御方式の外部評価。

<https://arxiv.org/html/2602.03117v1>

[40] 株式会社 AI アシュアランス | 日本発「責任 OS」は、どのように生まれたのか—2025 年 10 月の有限閉包型実行制御と、11 月の ADIC 再検証技術が合流するまで

日付：公開 2026-07-17/更新 2026-08-30 / 区分：当事者による公開技術史

出願日・番号と技術史は当事者による公表。公的受領記録ではない。

<https://zenn.dev/ghostdrift/articles/049534ae1a288e>

[41] WIPO | PCT International Search and Preliminary Examination Guidelines, PCT/GL/ISPE/14

版：PCT/GL/ISPE/14、2025-12-12 公表/2026-01-01 施行 / 区分：公的審査指針

6.05-6.10（印刷頁 55-56）。組合せの開示と、明細書・請求項・図面全体による優先日の判断。

<https://www.wipo.int/documents/d/pct-system/docs-en-texts-ispe.pdf>

[42] European Patent Office | Guidelines for Examination, F-VI 2.2: The same invention

日付：2026 年版 / 区分：特許庁の審査ガイドライン

優先権書類全体から直接かつ一義的に導かれる事項。

https://www.epo.org/en/legal/guidelines-epc/2026/f_vi_2_2.html

PRIMARY SOURCES

一次資料一覧 43-49

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[43] AARM / Cloud Security Alliance | Autonomous Action Runtime Management — System Category Specification v1.0

日付：2026-02 表記／調査時点版 ／ 区分：民間共同仕様の一次文書

Core R1-R6、Extended R7-R9、適合性の扱い。

<https://aarm.dev/spec>

[44] Cloud Security Alliance | Autonomous Action Runtime Management (AARM) Working Group

日付：更新日明示なし／調査時点版 ／ 区分：受入組織の一次資料

作業部会の活動・位置付け。

<https://cloudsecurityalliance.org/research/working-groups/autonomous-action-runtime-management-aarm>

[45] Vanta | Vanta donates AARM to the Cloud Security Alliance

日付：2026-04-30 ／ 区分：移管元による一次発表

企業起点の仕様を共同運営へ移す発表。

<https://www.vanta.com/resources/vanta-donates-aarm-to-csa>

[46] Google Cloud | Powering AI commerce with the new Agent Payments Protocol (AP2)

日付：2025-09-16 ／ 区分：開発・提供主体の一次資料

委任・ユーザー意図・支払指示の関連付け。

<https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol>

[47] Google | Agent Payments Protocol joins FIDO Alliance

日付：2026-04-28 ／ 区分：移管元による一次発表

共同運営への移管。支払後の履行保証とは区別。

<https://blog.google/products-and-platforms/platforms/google-pay/agent-payments-protocol-fido-alliance/>

[48] MLCommons | The key to trustworthy AI evaluation is secrecy by design

日付：2026-08-27 ／ 区分：実証参加主体の一次資料

AVERI・OpenMined 等との評価実証。

<https://mlcommons.org/2026/08/double-blind-reliability-evaluation/>

[49] ElevenLabs | AIUC announcement

日付：2026-02-12 ／ 区分：利用側企業による一次発表

認証・保険の利用側による発表。

<https://elevenlabs.io/blog/aiuc-announcement>

PRIMARY SOURCES

一次資料一覧 50-56

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[50] **Chaucer | Chaucer and Armilla AI launch Vanguard AI coordinated insurance structure**

日付：2026-02-10 / 区分：保険側当事者による一次発表

保険側が公表する連携。補償条件は個別確認が必要。

<https://www.chaucergroup.com/news/press-release-chaucer-and-armilla-ai-launch-vanguard-ai-coordinated-insurance-structure>

[51] **AgentScope | AgentScope Runtime — 简介**

日付：更新日明示なし／調査時点版 / 区分：公開開発文書／中国語資料

サンドボックス・実行・観測の基盤。

<https://runtime.agentscope.io/zh/intro.html>

[52] **Haoyu Wang, Christopher M. Poskitt, Jun Sun | AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents**

日付：2025-03-24／v1 / 区分：著者による原著研究

トリガー、述語、介入による実行時制約。

<https://arxiv.org/html/2503.18666v1>

[53] **Hanzo AI | Formal Proofs Ship**

日付：2026-03-26 / 区分：開発・提供主体の一次資料

エージェント関連の形式証明に関する主張。

<https://hanzo.blog/blog/2026-03-26-formal-proofs-ship/>

[54] **Hanzo AI | Hanzo Formal Proofs — README**

日付：2026-10-15 閲覧／main / 区分：公開リポジトリ／README 確認

AI Chain、モデル由来、DID、TEE、精算に関する現行記載。ブログの全主張との対応は未確認。

<https://github.com/hanzoai/proofs/blob/main/README.md>

[55] **Holistic AI | EndLayer — Guardian Agents for AI**

日付：公開日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

Sensor・Decoder・Guardian、端末内観測と制御、版付き記録。

<https://www.holistica.com/endlayer>

[56] **Fiddler AI | Prevent AI Agents from Leaking Sensitive Data with Fiddler Guardrails and LiteLLM**

日付：2026-09-14 / 区分：開発・提供主体の一次資料

Gateway への実装例、pre_call/post_call、fail-closed 設定。

<https://www.fiddler.ai/blog/prevent-ai-agent-data-leakage-guardrails>

PRIMARY SOURCES

一次資料一覧 57-59

最終閲覧日：2026-10-15。日付不明は不明と記す。題名は原題又は内容を特定する訳題・短縮名。企業の一次発表は当該主体の主張・手順の出典であり、独立した認証とは限らない。

[57] IBM | watsonx Orchestrate — Agent control plane

日付：公開日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

導入前評価、ポリシー、ID、監査・観測の統合。

<https://www.ibm.com/products/watsonx-orchestrate/agent-control-plane>

[58] Arthur AI | Agent Security & Governance

日付：公開日明示なし／調査時点版 / 区分：開発・提供主体の一次資料

検出、監視、判定 API、アプリ側の強制責任。

<https://www.arthur.ai/discover-and-govern-agents>

[59] M. Kaptein, V.-J. Khan, A. Podstavnychy | Runtime Governance for AI Agents: Policies on Paths, arXiv:2603.16586v1

日付：2026-03-17 / 区分：原著研究／要旨による補助確認

要旨で確認：部分実行経路・行動・組織状態を扱う枠組み。本文全体の実装検証は未実施。

<https://arxiv.org/abs/2603.16586v1>

SUPPLEMENTARY SOURCES & REVIEW LOG

補助資料・調査版の記録

非公開の保有原稿：公開比較とは別の補助確認

以下は技術史の限定的な照合にのみ使用した。いずれも公開 URL はなく、第三者が本書だけで全文を再確認できる資料ではない。原稿の提出日、出願番号との対応、提出済み原本との同一性は確定していない。

識別子	閲覧した資料	参照箇所・用途
N1	非公開原稿 N1	【0028】～【0031】、【0052】、【0057】～【0058】、【0070】～【0071】。公開済みの実行制御・監査の説明との照合。
N2	非公開原稿 N2	【0021】～【0028】等。公開済みの通信・停止・再加入の説明との照合。
N3	非公開原稿 N3	【0007】～【0013】等。ADIC 関連の再計算・再検証の照合。該当する出願との対応関係は未確認。

公開成果物について今回実際に確認した記録

Microsoft：固定コミット 0b83ab91508e9ceca159090240ddaaf9d4b2b88e の README で Public Preview、govern 等の手順、上流 ACS への依存を再確認。ソース全体の実行・コード監査は未実施。[6]

責任 OS・EU マッピング：コミット bef1531b839b8fa55b6ab86786d538305b8858f2 の README を確認。Lean verification の run 37097244762 について、Build mapping と Verify source with warnings as errors の成功を GitHub API で確認した。Pages 配信の成功とは区別している。[37・38]

Hanzo：現行 README の一覧とブログの主張を照合したが、ブログの全主張に対応する証明ファイル・定理・実行版を特定できなかった。ソースのビルド失敗を確認したという意味ではない。[53・54]

版の扱い

本版は第 3 版を基に公開一次資料・日付・URL・主要主張を最終照合した第 4 版である。AWS の履歴失効、Microsoft の上流仕様、ICME の説明不整合、Weaveframe の未実装出力を修正した。原文本文を再取得できなかった旧 BSI 資料は実質的な根拠から外し、[31] を GAVEL v1.1 の限定実験に差し替えた。更新型文書は取得時点の観測であり、全ページの長期保存は行っていない。

関連研究の補助確認：Policies on Paths は、部分的な実行経路、次の行動、組織状態を政策評価へ入れる枠組みを要旨で確認した。本書の履歴・状態を含む分析の補助とし、商用実装・性能の根拠には用いていない。[59]