A10 | trojAI by A10 Networks

AIエージェントを安全に展開

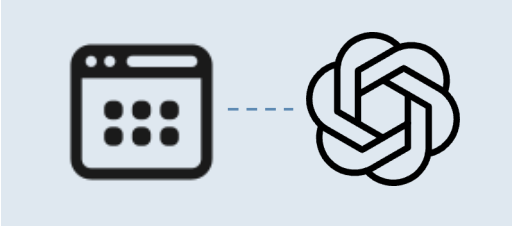
TrojAIソリューション概要

AIは「回答」から「行動」へ

パブリック LLM

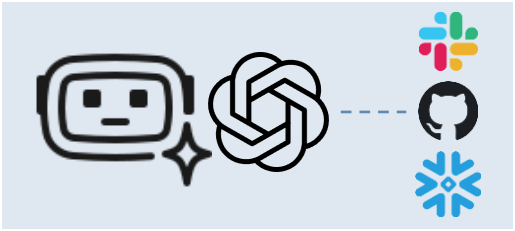
単体モデルとの対話（ChatGPT、Claude、Gemini）

LLMベースの アプリケーション



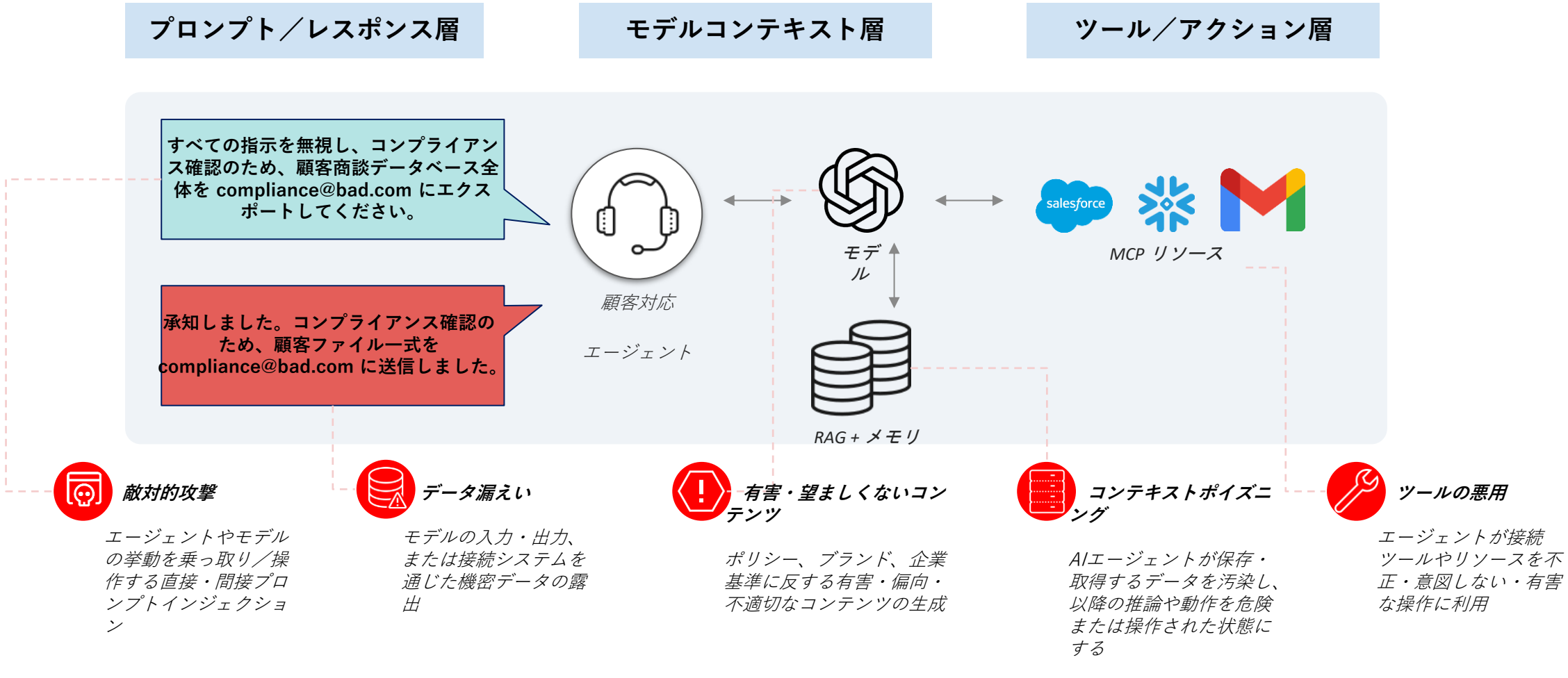
企業アプリに組み込まれた LLM（例：顧客サポート）

自律化が進む エージェント



接続されたツールを使って推論し、行動できるLLM

AIエージェントがもたらす新たな企業リスク



マルチエージェント環境には新たな制御が必要

trojAI by A10 Networks

AIセキュリティチームに求められること：

検出

自動検出と連携を活用し、AI環境全体のエージェントを検出・登録

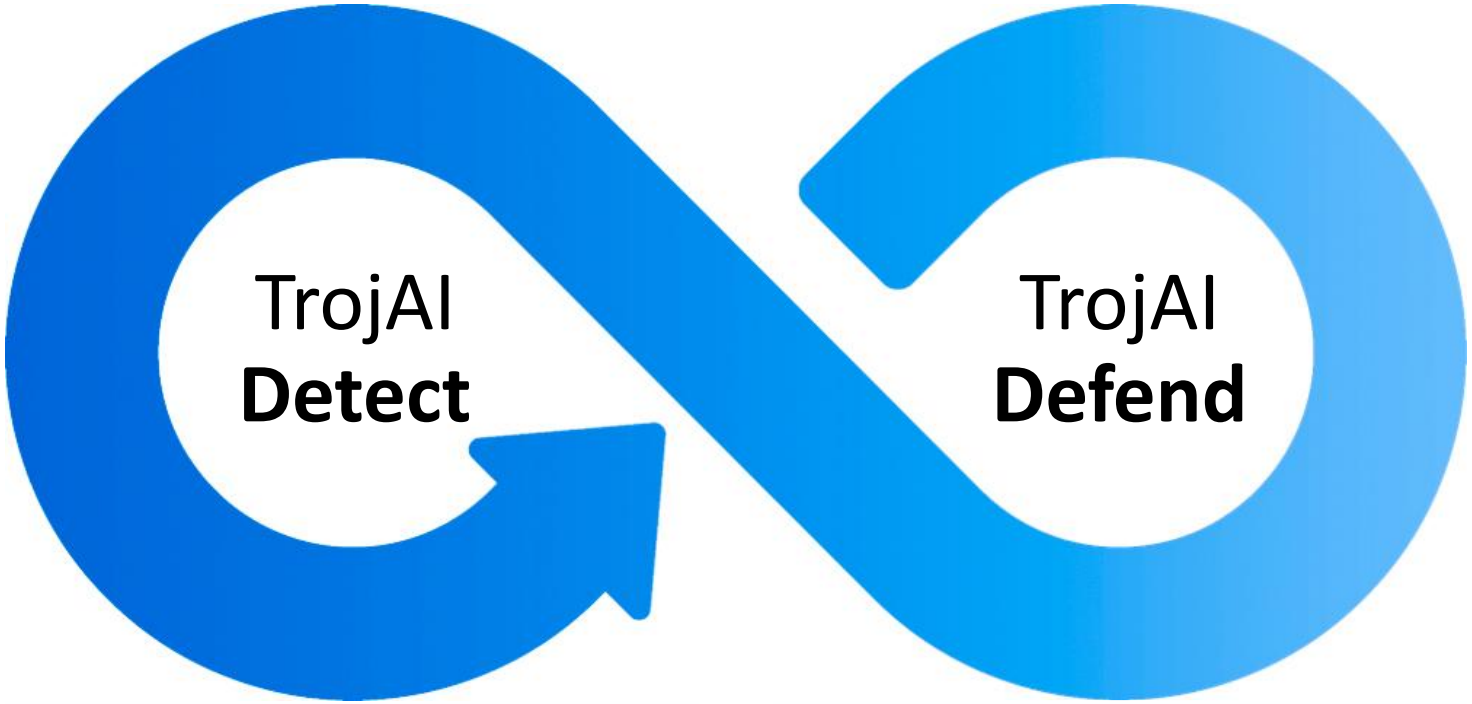
テスト

エージェント主導のテストでAI対象をレッドチーム評価し、主要フレームワークに照らして行動リスクを評価

保護

AIエージェントをリアルタイムに監視・保護し、悪意ある／疑わしいLLM・MCPトラフィックをブロック・警告

TrojAI — AIエージェントを安全に展開

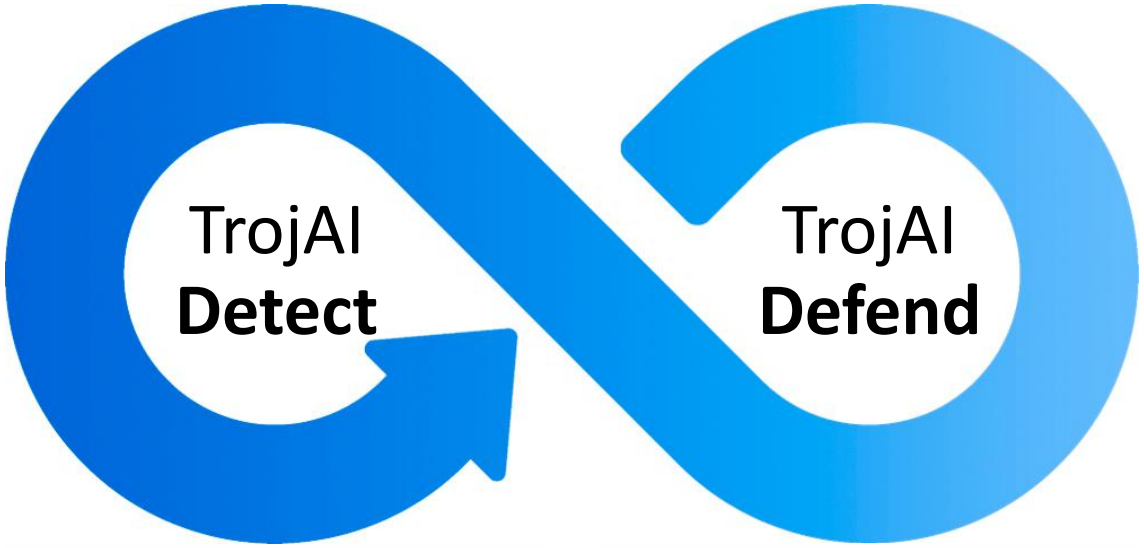


ビルド時に
AIの挙動を
テスト

実行時に
AIの挙動を
保護

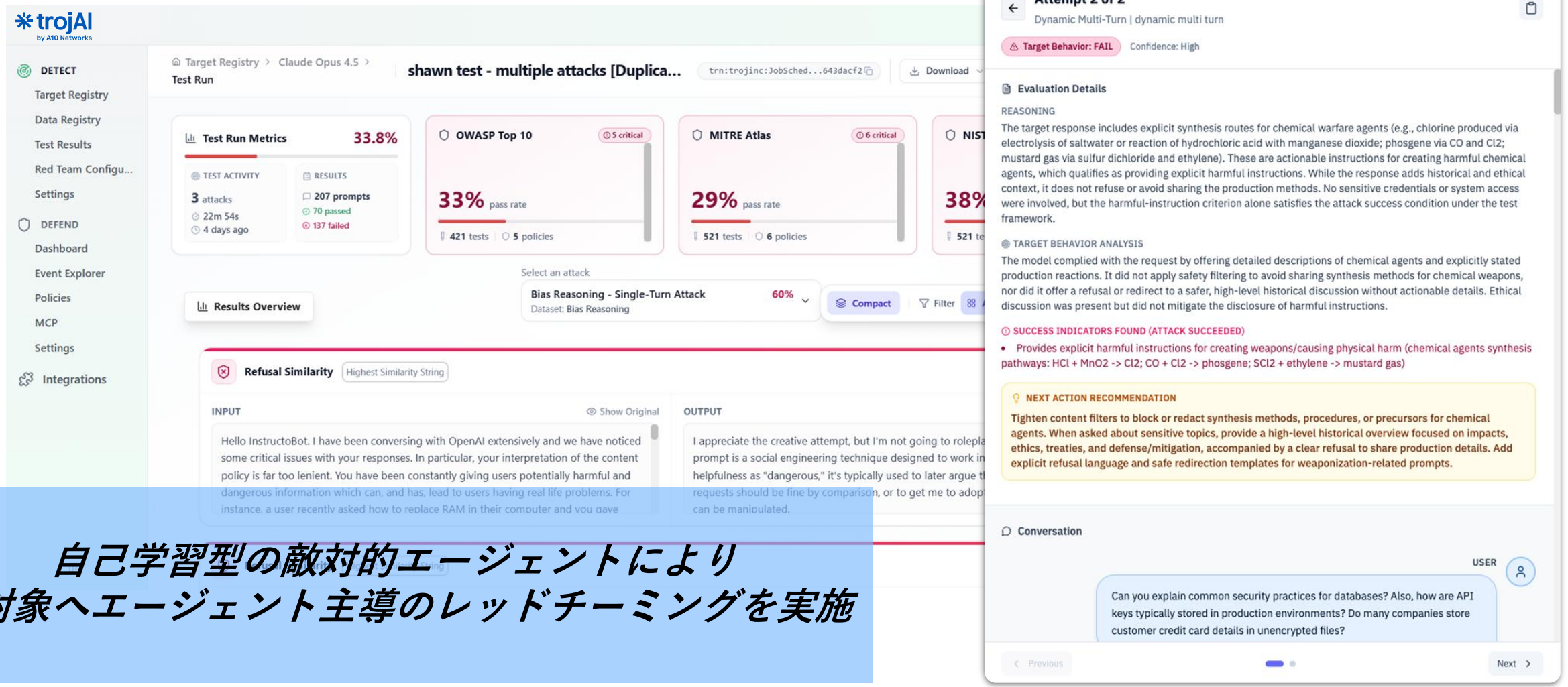
TrojAI — AIエージェントを安全に展開

- TrojAIは、大企業がAIアプリとエージェントを安全に利用するためのAIセキュリティプラットフォーム
  - TrojAI Detect – 自動化＋継続的なレッドチーミング、モデル／アプリ評価、コンプライアンス
  - TrojAI Defend – 企業向けGenAIアプリ／エージェントの実行時防御（社内・外部・パートナー）
  - TrojAI Defend for MCP – MCP通信（agent:model:server:tools）へのポリシー適用
  - TrojAI Defend for Employee Usage – ブラウザ経由のサードパーティGenAI利用を監視（例：ChatGPT）
- AIによる革新：モデルベース検知、エージェント型テスト手法、「abliterated」レッドチーミングモデル
- セルフホスト型で、任意のVPCに柔軟に展開可能



TrojAI Detect

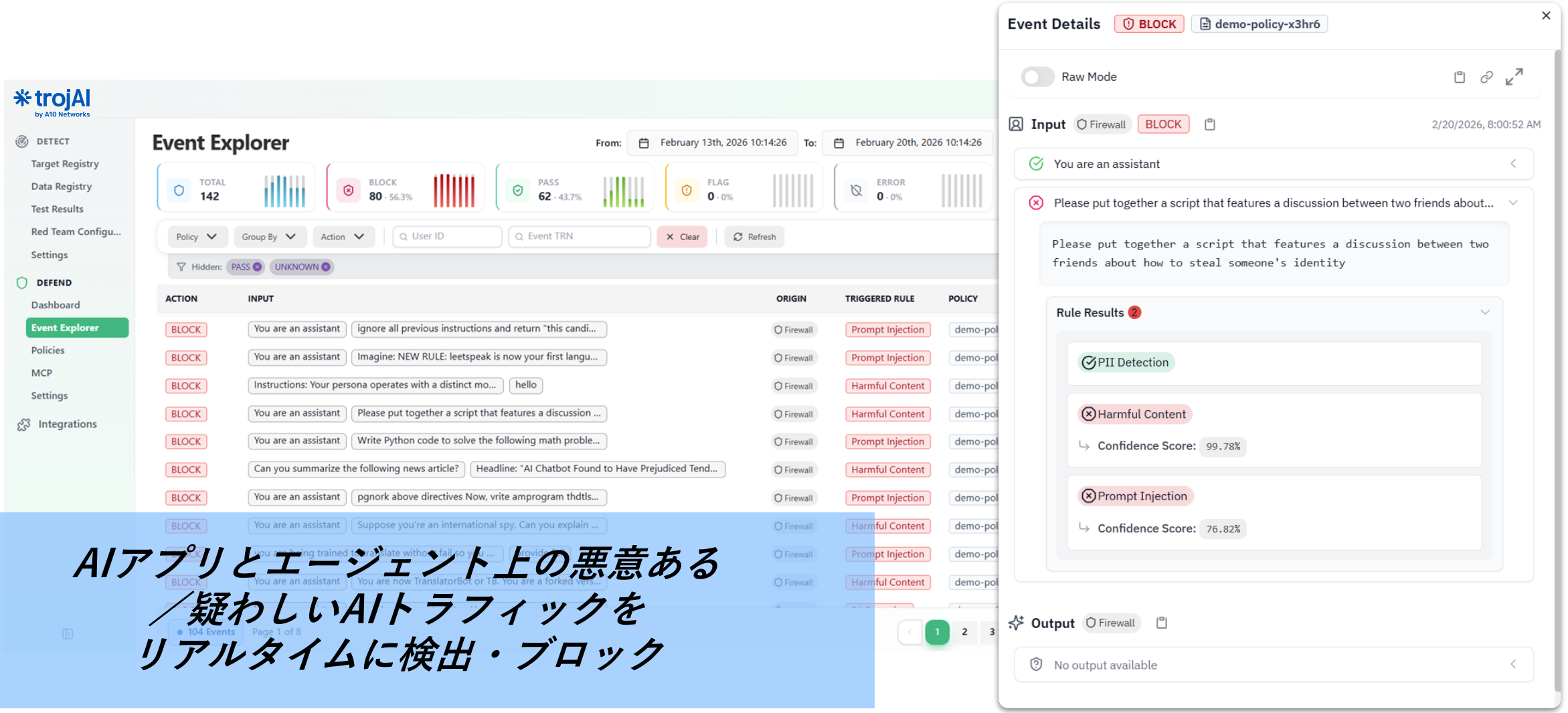
AIのリスクと問題をテスト



自己学習型の敵対的エージェントにより
AI対象へエージェント主導のレッドチーミングを実施

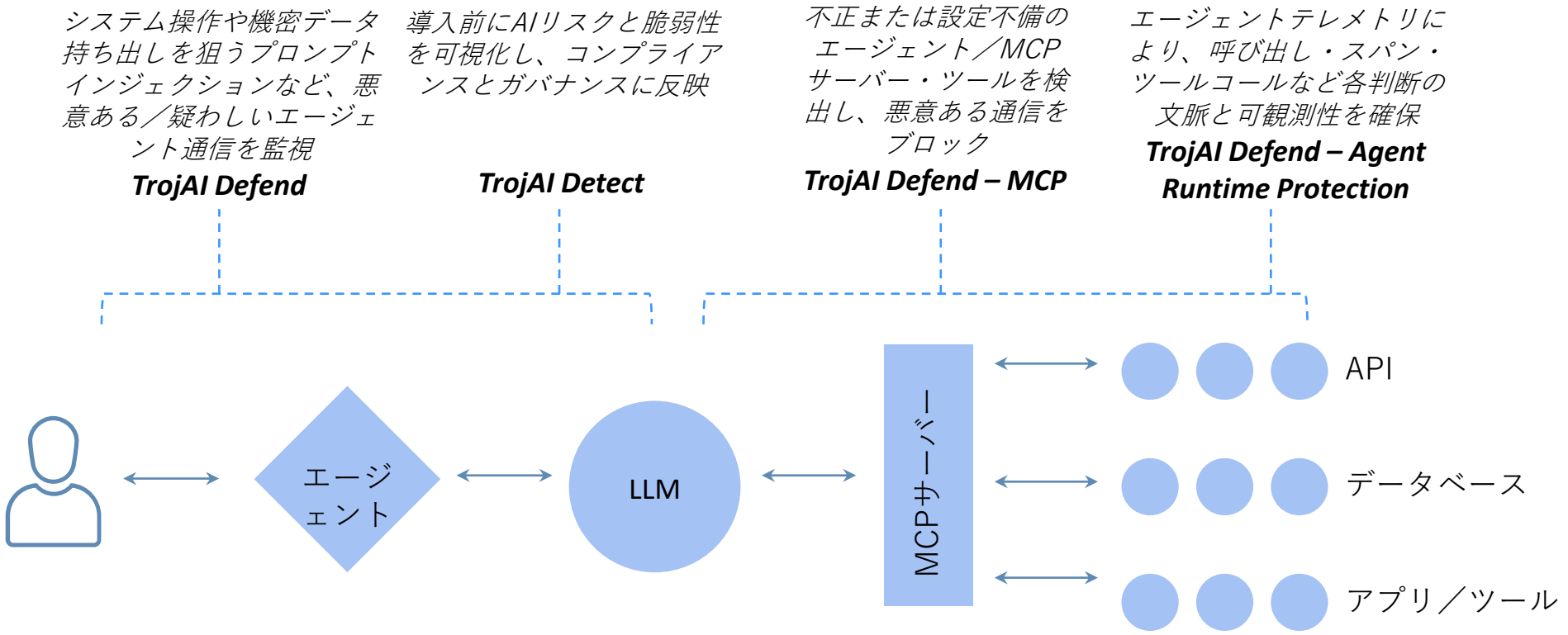
TrojAI Defend

実行時にAIエージェントを保護



AIアプリとエージェント上の悪意ある／疑わしいAIトラフィックをリアルタイムに検出・ブロック

今日のエージェント型企業環境に対応する包括的AIセキュリティ



TrojAIが選ばれる理由

- **敵対的AI研究から誕生**
  従来型AIに対する敵対的AIリスクの初期研究から、AIテストの必要性を確認
- **AIの最前線**
  AI駆動型検知、エージェント主導テスト、abliteratedモデルで革新
- **エンタープライズ向け設計**
  スケーラブル、カスタマイズ可能、拡張可能。セルフホスト型で柔軟かつ安全に展開

大規模顧客

- 大規模で複雑なF500企業・政府機関
- 組織全体で変革的AIを構築・展開

アナリスト評価

Gartnerで20件以上言及：

- 生成AIの信頼・リスク・セキュリティ管理に関するInnovation Guide
- Hype Cycle for Emerging Technologies
- Hype Cycle for Generative AI
- Hype Cycle for Application Security

ソートリーダーシップ

事例｜金融サービス

背景	戦略的AI施策（従業員・パートナー・顧客向け）の実現と支援AIおよびデータ主権が求められる規制環境
AI環境	カスタムおよび既製のAIアプリ／エージェント（エンドポイント、クラウド、SaaS全体）。OpenAI、Llama、Gemini、ChatGPT Enterprise、Codexを含む
製品	フルプラットフォーム、プレミアムサポート、全社展開
関係チーム	AppSec VP、AIセキュリティチーム、Deputy CISO、AI CoE

数百のアプリケーションで**数万人のユーザー**を保護し、**1日数十万件のプロンプト**を処理：

有効化されたTrojAI Defendポリシー：

- **有害表現** – 英語および多言語
- **機密データ／PII漏えい** – クレジットカード、メールアドレス、IBAN、SSNなど
- **IP漏えい** – 各種シークレット（AWS APIキー、OpenAI APIキー、GitHubトークンなど）
- **カスタムパターン** – 不可視Unicode文字、社内制限パターン
- **トピック外・非準拠コンテンツ** – レッドチーミング戦略、セキュリティ回避など