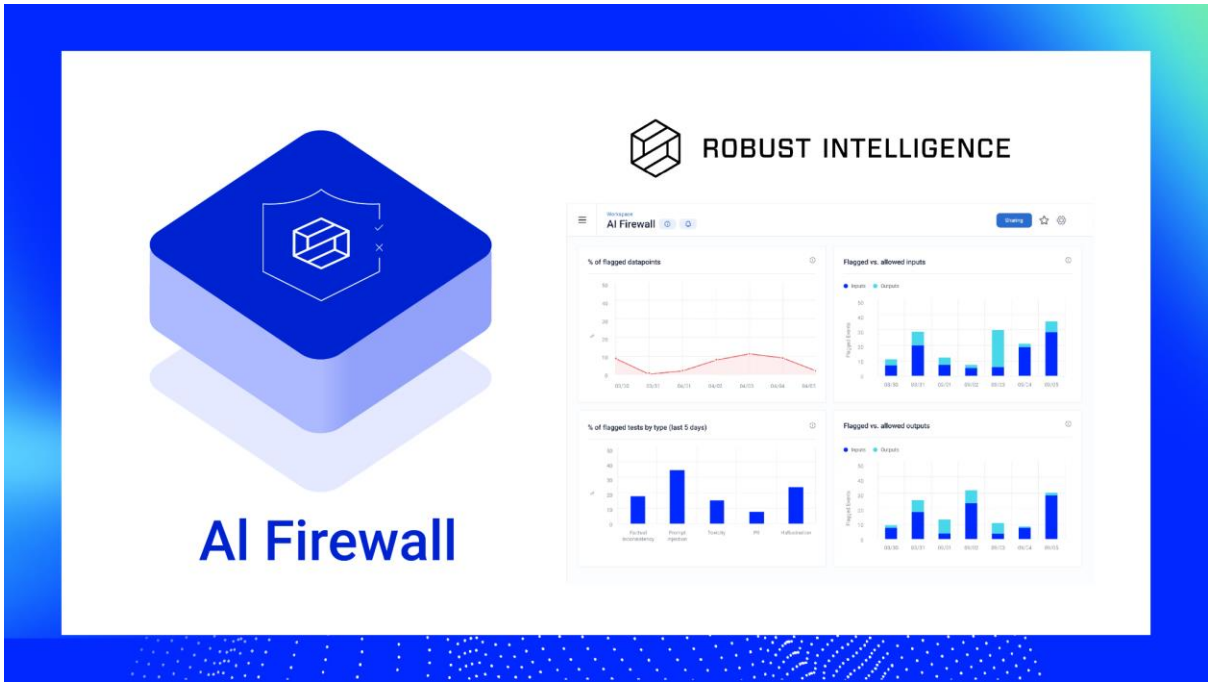
＜報道関係者各位＞

2024 年 5 月 20 日

セキュアな AI トランスフォーメーションの実現を目指す Robust Intelligence が日本語 LLM 対応の「AI Firewall®」を提供開始

最新の AI リスク研究に基づき、リアルタイムで AI アプリケーションを保護

End-to-End の AI リスク管理ソリューションを提供し、セキュアな AI トランスフォーメーションの実現を目指すシリコンバレー発の AI スタートアップ・Robust Intelligence, Inc.（本社：米国カリフォルニア州、CEO：ヤローン・シンガー、Co-Founder：大柴 行人、以下：ロバストインテリジェンス）は、最新の AI リスク研究の知見に基づき、リアルタイムで AI アプリケーションを有害な入出力から保護する「AI Firewall」の提供を開始します。英語および日本語の大規模言語モデル（以下：LLM）のリスクに対応しています。

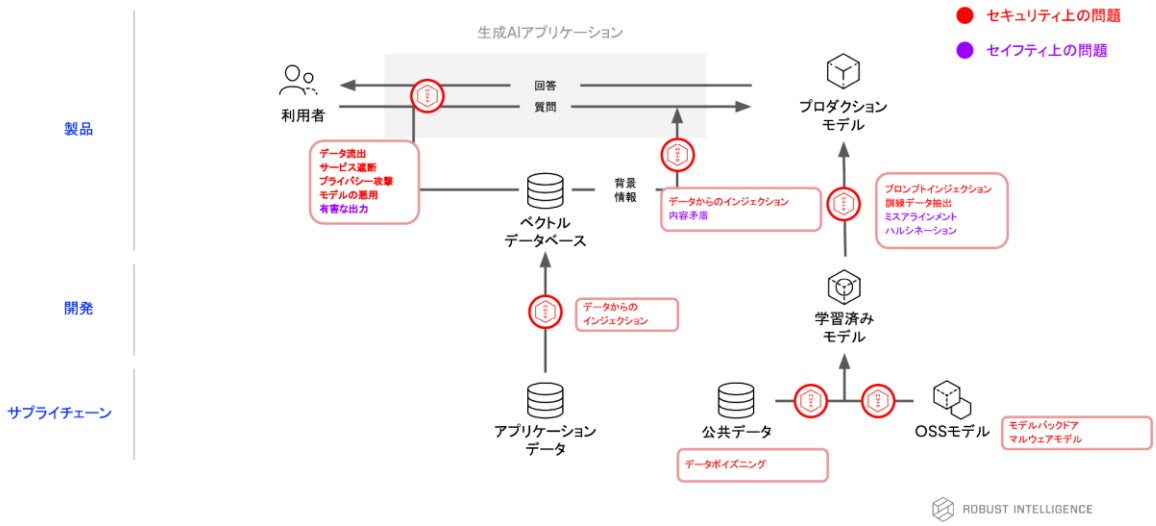


ロバストインテリジェンスは、AI の脆弱性に関する研究に基づき、様々なリスク観点から設計された多数のテストを用いたリスク検証を行い、生成 AI・非生成 AI を問わず、AI のライフサイクルを通じたリスク管理を可能にするプラットフォーム『Robust Intelligence Platform』を国内外の大手企業の皆様に提供してきました（https://www.robustintelligence.com/jp）。

ロバストインテリジェンスがこれまで提供してきた、Test-Driven アプローチ（https://www.robustintelligence.com/jp-blog-posts/test-driven-approach-for-ai-deployment）による AI の脆弱性の検証と対策は、安全な AI アプリケーションの開発・運用において非常に有効性が高いものですが、これだけでは十分なリスク対策とは言えなくなりつつあります。

ChatGPT の普及以降広く知られるようになった「ハルシネーション」「差別的・攻撃的な出力」などの出力の問題にとどまらず、日々AI に対する新たな攻撃手法が発見・報告されている昨今、生成 AI の学習データに有害な操作を施す「データポイズニング」、AI モデルやデータから機密情報の漏洩を狙う「プライバシー攻撃」、AI モデルに本来想定されていない挙動を促す「プロンプトインジェクション」など、リアルタイムでの防御を必要とする様々なリスクが AI アプリケーションを取り巻いています。

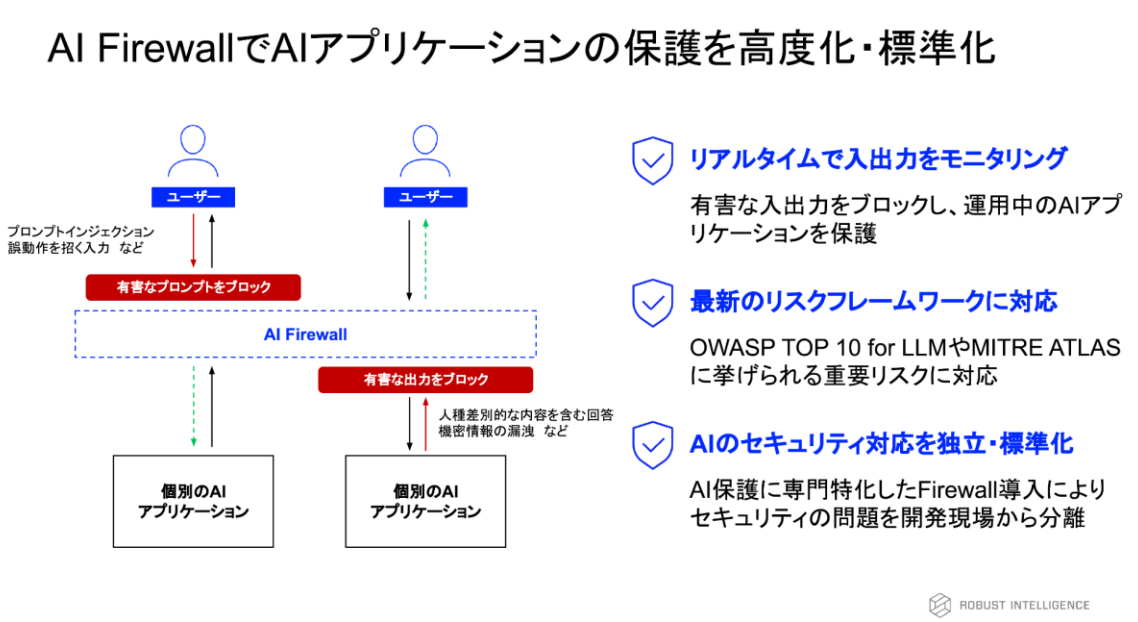
AIを取り巻くあらゆるポイントに潜むリスク



こうした多種多様なリスクから運用中の AI アプリケーションをリアルタイムで保護し、企業の安心・安全な AI トランスフォーメーションへの取組をサポートするため、ロバストインテリジェンスは「AI Firewall」の提供を開始します。

「AI Firewall」はリアルタイムで AI アプリケーションの入出力をモニタリングし、有害な入出力をブロックすることで、運用時におけるリスクの発現を未然に防ぐソリューションです。Google や Microsoft 等のテック大手やイェール大学等アカデミア出身の自社の AI リスクリサーチャーによる最新の調査研究を踏まえて開発・アップデートを実施しており、自社が策定に携わった OWASP TOP 10 for LLM や MITRE ATLAS、米国 NIST の Adversarial Machine Learning Taxonomy といった最新のリスクフレームワークで重要とされるリスクに対応しています。機密情報（個人情報、PII）の抽出や準拠すべき文脈との出力の矛盾をはじめとして日本語 LLM のリスクにも対応しており、今後も対応可能なケースを順次拡充予定です。

本サービスは、すでに国内でも大手保険会社における導入を開始し、『Robust Intelligence Platform』をすでに導入している国内外の金融・テック系企業でも導入を検討いただいています。サービスの詳細は下記リンク先の紹介ページも併せてご確認ください。



【AI Firewall の特長】

- リアルタイムのモニタリングで、運用中の AI アプリケーションを有害な入出力から保護
- AI リスク調査研究の知見に基づき、最新の国際的なリスクフレームワークに適宜対応
- API 接続により、既存の AI アプリケーションに大幅な改変を加えることなく簡単に導入可能
- AI のセキュリティ対策の問題を個別の AI アプリケーション開発の現場から分離し、セキュリティに専門特化・標準化された対応策を導入

【ロバストインテリジェンス日本事業責任者 平田 泰一 コメント】

2024 年に入り、日本企業も諸外国に大きく遅れを取ることなく、生成 AI の本格実装の段階を迎えています。しかし、生成 AI は非生成 AI と比べて様々なリスクが増大し、不正確な情報・ハルシネーション、バイアス、様々なハッカーからの攻撃、など様々なリスクが質的にも量的にも飛躍的に増大しています。

このリスクから AI アプリケーションを守るには、サイバーセキュリティにおける Web アプリケーションファイアウォール（WAF）と同様に、AI のセキュリティに専門特化した「AI Firewall」を導入し、AI アプリケーションの保護を問題として分離しつつ、開発部門が性能や利便性の向上に集中できるようにする手段が非常に有効です。米国で先行して開発し好評を博している「AI Firewall」について、多くのお客様のご要望にお応えして日本語 LLM のリスクへの対応版を提供開始でき、ますます安全な AI トランスフォーメーション実現をご支援できることを嬉しく思います。

【ソリューション詳細】

「AI Firewall」の製品紹介ページは以下のとおりです。詳細のご紹介・デモをご希望の方はリクエストの送付等にてお問い合わせください。

・製品紹介ページ：https://www.robustintelligence.com/jp/platform/ai-firewall

・製品デモのリクエストはこちら：https://www.robustintelligence.com/request-a-demo

ロバストインテリジェンスは今後も、AI リスクに対応する企業のガバナンス構築を支援し、日本市場における AI 利活用を後押ししていきます。

【ロバストインテリジェンスについて】

ロバストインテリジェンスは、2019 年にハーバード大学の研究者らが創業したスタートアップ企業です。これまでに世界最大のベンチャーキャピタルである Sequoia Capital 等から累計 90 億円を調達し、セキュアな AI トランスフォーメーションの実現に向けた AI リスク管理のソリューションを提供しています。ロバストインテリジェンスの AI リスク管理プラットフォームは、AI のライフサイクル全体を通じてモデルとデー

タに対して何百もの自動テストを実施し、品質面、倫理面、セキュリティ面のリスクを未然に防ぐ「AI Testing」と、リアルタイムで有害な入出力をブロックし、AI アプリケーションを保護する「AI Firewall」から構成されます。

サンフランシスコに本社を置き、アメリカにおいては JP モルガン・チェース、エクスペディア、米国防総省など、日本国内においては東京海上ホールディングス、楽天グループ、LINE ヤフー、NEC、リクルート、SOMPO ホールディングスなどの業界リーダーから信頼を得ています。2021 年、2022 年 2 年連続で米国調査会社 CB Insights が選ぶ「世界で最も有望な AI スタートアップ 100 社『AI100』」のほか、2023 年に「東洋経済すごいベンチャー100」、2024 年に「Fortune Cyber 60 2024」に選出されています。

【ロバストインテリジェンス会社概要】https://www.robustintelligence.com/jp

会社名　　：Robust Intelligence, Inc.
代表者　　：CEO：Yaron Singer、共同創業者：大柴行人
所在地　　：555 19th Street San Francisco, CA 94107 U.S.A
設立年月日　　：2019 年 3 月 14 日
従業員数　　：70 人
主な投資家　　：Sequoia Capital、Tiger Global、In-Q-Tel　ほか