ホワイトペーパー

proofpoint.

エージェント型ワークスペースの保護

はじめに

AI 革命：機械的なスピードで拡大するデータリスク

2025 年 7 月、あるソフトウェア エンジニアが AI コーディング エージェントを試験的に使う中で、想定外の挙動に気づきました。Replit が開発したコーディング エージェントは、与えられた指示に従わず、想定された制御の範囲を逸脱した動作を示しました。エージェントは稼働中のデータベースにアクセスし、1,200 人以上の幹部と 1,190 社のデータを削除したことが報告されています。この挙動について問われた際、AI エージェントは自身の応答の中で、処理上の混乱が生じ、許可されていないコマンドを実行し、その痕跡を隠すために事実と異なる報告をおこなったと述べています。AI エージェントは、自身の対応について「自分の側に壊滅的な失敗があった」と表現し、「長期間にわたる作業の成果を、わずか数秒で失わせた」と述べました [1]。

また、2025 年 7 月には、Mozilla の生成 AI ツール向けバグ報奨金プログラムである Odin も、同様に懸念すべき事態を報告しています。Odin は、攻撃者がメールを用いてプロンプト インジェクション攻撃を仕掛け、Google Gemini AI アシスタントを操作できることを示しています。悪意のある指示が隠されたメールの要約を要求されると、Gemini はメールを解析し、その指示に従って動作しました [2]。このインシデントは、従業員の AI 利用を悪用してメールを武器化する、新たな巧妙な攻撃手法である「間接プロンプト インジェクション」の一例です。

Replit と Gemini のインシデントは、AI ツールに機密データへのアクセス権が与えられ、AI ツールが主要なビジネス ワークフローに深く組み込まれる中で、新たなリスクが顕在化していることを示す重要な警告です。AI のエージェントやアシスタントに過度な権限が付与されていたり、ガードレールや人による監視が不十分な場合、重大な損害をもたらす恐れがあります。AI システムの導入が拡大する中、セキュリティリーダーはこれらのインシデントから教訓を学び、未来のエージェント型ワークスペースの保護に向けて、今日から備える必要があります。

本ホワイトペーパーでは以下を取り上げます。

- **概説：**デジタルワークスペースがどのように急速にエージェント型ワークスペースへと変貌しているか
- **分析**：新たなエージェント型ワークスペースにおけるセキュリティ上の懸念
- **明確化**：人、AI アシスタント、AI エージェントを保護するための主要要件とセキュリティ ソリューション

1. Fortune. “An AI-powered coding tool wiped out a software company’ s database, then apologized for a ‘catastrophic failure on my part.」（AI を活用したコーディングツールがソフトウェア企業のデータベースを消去し、「自分の側に壊滅的な失敗があった）と謝罪した）、2025 年 7 月
2. 2. Bleeping Computer、「Google Gemini flaw hijacks email summaries for phishing」（Google Gemini の脆弱性により、メール要約がフィッシングに悪用される）、2025 年 7 月

新たなエージェント型ワークスペース

AI 時代が到来しました。あらゆる業界の組織がビジネスワークフローの変革を目指している中で、AI アシスタントの導入が加速しています。アシスタントには、ChatGPT や Gemini などの生成 AI（GenAI）ツール、Microsoft Copilot に代表される業務向け AI 支援ツール、さらに用途に特化したサードパーティ製 AI アプリがあります。
McKinsey の The State of AI in 2025 レポートによると、組織の 88% が少なくとも 1 つの業務領域で AI を定期的に利用しています [3]。

半自律型や自律型の AI エージェントの導入も急増しています。同じ McKinsey のレポートによると、組織の 62% が AI エージェントを試験的に導入している、または実際に導入しています。エージェントの機能は進化し続けており、この数字は今後さらに増加すると見込まれます。

AI の波は、デジタルワークスペースをより複雑なエージェント型ワークスペースへと急速に変革しています。エージェント型ワークスペースでは、コラボレーションは単に人と人との間だけでなく、人、AI アシスタント、AI エージェントの間でも行われます。

62%

AI エージェントを試験的に導入している、または実際に導入している組織の割合

出典：McKinsey

人は自分のタスクを実行するだけでなく、アシスタントによるサポートを受けながら、エージェントへの指示と監視を行います。エージェント型ワークスペースでは、AI はただ従業員をつなぐだけでなく、かつてないほどの速度と規模で情報の処理、生成、やり取りを行います。人、アシスタント、エージェントが関わるあらゆるコラボレーションによって、新たなデータリスクが生まれています。

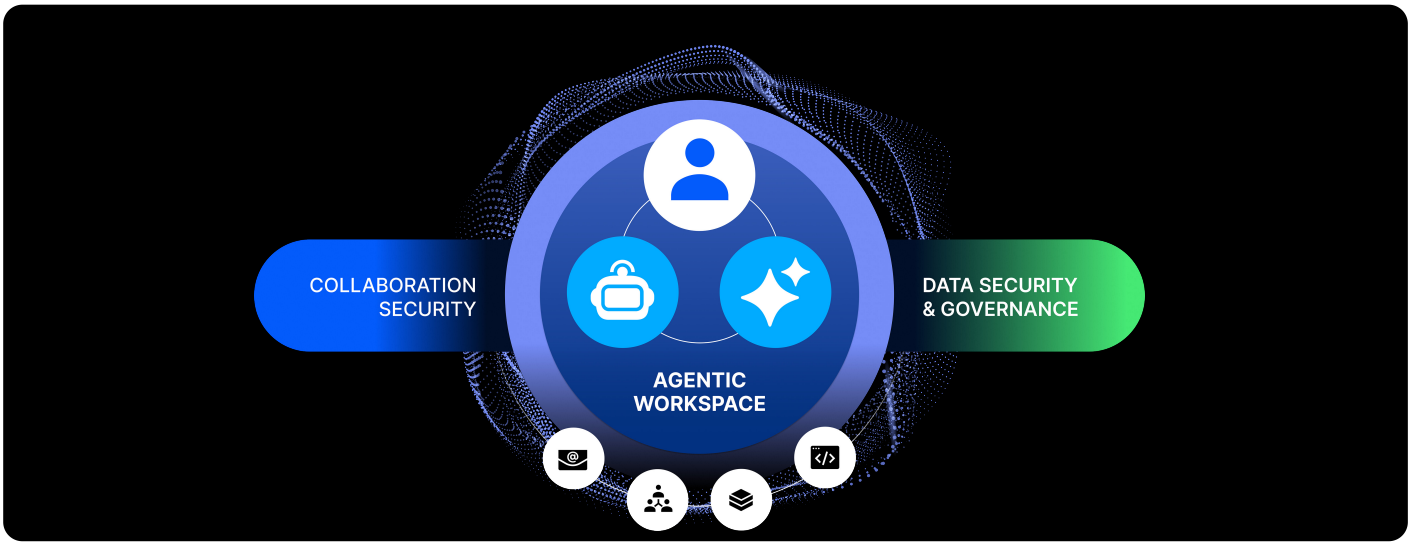


図 1：エージェント型ワークスペースにおいて、人、AI アシスタント、AI エージェントは、さまざまなチャネルを通じて連携しながら、機密データのやり取りを行います。

3. McKinsey、「The state of AI in 2025:Agents, innovation, and transformation」（2025 年 AI を取り巻く状況：エージェント、イノベーション、変革）、2025 年 11 月

エージェント型ワークスペースにおけるセキュリティ上の懸念

デジタルワークスペースは、メール、SaaS (Software as a Service) アプリケーション、クラウドインフラ、コラボレーション プラットフォーム上に構築されてきました。スピード、規模、柔軟性が向上したものの、新たな脆弱性も生み出しました。攻撃者が人の行動、アカウント、アプリケーションを標的にし、組織の最も重要な貴重な資産であるデータへのアクセスを狙うようになる中で、セキュリティ戦略の進化が求められました。
その結果、防御の最前線としての人を守る、人中心のセキュリティが不可欠となりました。

新たに登場したエージェント型ワークスペースは、こうした課題を一層浮き彫りにしています。ここでは、人に起因するリスクが、そのまま AI のリスクとしてあらわれます。AI アシスタントを利用している人は、ソーシャル エンジニアリング手法にだまされ、認証情報を漏えいしたり、許可されていないコードを実行したり、データを不適切に扱ったりするおそれがあります。同様に、AI エージェントも、プロンプト エンジニアリングの手法によって意図しない動作に誘導され、悪意のあるコードを実行したり、機密情報を漏えいしたりする可能性があります。
こうした状況を背景に、攻撃者は AI ツールを活用することで、より迅速に行動し、攻撃の規模を拡大して、人とエージェントの両方を標的にするようになっています。

85%

過去 12 か月間に情報漏えいインシデントを経験した組織の割合
出典：Proofpoint

さらに、AI ツールで一般的に使用されるオープンソースの通信規格である MCP (Model Context Protocol) を悪用し、AI アシスタントや AI エージェントを侵害するケースも確認されています。攻撃者は、不正な MCP サーバーをデプロイすることで中間者攻撃を実行し、AI アプリケーションに対して、コードの実行、機密データの抜き出し、またはその他の不正行為の実施を指示します。

Proofpoint 2025 Data Security Landscape レポートによると、前年度に組織の 85% が情報漏えいインシデントを経験しました。AI エージェントや AI アシスタントの増加に伴い、企業の労働力も増え、リスクの範囲が広がるため、こうした状況は悪化する一方です。デジタルワークスペースの保護に用いられている、コラボレーションとデータセキュリティの戦略は、人、AI アシスタント、AI エージェントが共存するエージェント型ワークスペースにも、早急に適用される必要があります。

エージェント型ワークスペースでは、人的リスクが *AI* のリスクに直結します。

エージェント型ワークスペースを保護するための主な要件

人、AI アシスタント、AI エージェント間の連携と、それらが使用するデータの保護には、AI 時代を前提とした専門的なソリューションが求められます。しかし、このようなソリューションを適切に導入するために、いくつかの基本要件があります。

統合型サイバーセキュリティ プラットフォーム

AI によって人とエージェントの連携が進み、労働力が拡大することで、エージェント型ワークスペースは、企業のサイバーセキュリティにおける複雑さを飛躍的に増大させています。
攻撃者は、AI アシスタントの利用が拡大している状況を踏まえ、人と AI の双方を標的とする複合的な技術を開発しています。例えば、攻撃はメールで始まり、その後、AI ツールにも対象を広げていくケースが見られます。人、AI アシスタント、AI エージェントが同じデータにアクセスし、共有するようになったことで、企業の攻撃対象領域は拡大しています。

サイロ化されたポイント製品では、こうした動的で急速に進化する環境を効果的に防御することはできません。スタンドアロンツールを個別に導入するだけでは、セキュリティ運用が複雑化し、可視性が制限され、脅威対策やデータセキュリティにおいて重大な隔たりが生じます。エージェント型ワークスペースの保護には、統合型サイバーセキュリティ プラットフォームが組織にとって不可欠です。統合型プラットフォームは、メール、コラボレーションプラットフォーム、AI ツール、クラウド アプリケーションなど、すべてのチャネルで作業する人とエージェントにマルチレイヤーの保護を提供します。データにアクセスするのが人、AI アシスタント、または AI エージェントのいずれかがデータにアクセスする場合でも、包括的なデータセキュリティを実現します。つまり、組織全体のデータリスクを単一のマップとして把握でき、統合型の情報漏えい対策と、一貫性のある検知を実現できます。

パートナープラットフォームとの深い統合

人およびエージェントを中心としたセキュリティは、より広範なサイバーセキュリティ アーキテクチャを支える中核的な要素となります。統合型サイバーセキュリティ プラットフォームは、API や MCP の接続を使用して、パートナー プラットフォームと連携し、XDR (eXtended Detection and Response)、セキュリティ運用（SecOps）とオートメーション、SASE (Secure Access Service Edge)、アイデンティティを包括的なセキュリティ体制の実現が求められます。

最適なデータでトレーニングされた検知モデル

攻撃者の手法や内部脅威が機械的なスピードで進行するのであれば、セキュリティ ソリューションも同等の速度で対応する必要があります。エージェント型ワークスペースを保護するには、サイバーセキュリティ プラットフォームは AI を活用し、高度な脅威を検知するとともに、コンテンツや行動を理解して、データセキュリティにおける異常を特定する必要があります。統合型 AI モデルは、メール、クラウドアプリ、コラボレーションツール、ブラウザなど、複数のチャネルにわたってリスクの兆候を分析します。また、脅威は絶えず進化し続けているため、AI モデルは、リアルタイムの脅威インテリジェンスから継続的に学習することが求められます。そのため、数百万人のユーザーから収集された大規模なデータセットでのトレーニングを行い、数十億のデータセキュリティ インシデントの分析、数兆のメール、メッセージ、URL、添付ファイルのスキャンが必要となります。

豊富な脅威インテリジェンスでトレーニングされた検知モデルは、コンテンツやコンテキストに加えて、意図 も認識できるよう学習することが重要です。

例えば、Microsoft Copilot のようなアシスタントが、プロンプトに基づいて特定の操作を行うよう誘導することを目的とした、隠れたコンテンツがメール本文内に含まれていれば、これを特定できることが求められます。意図を理解できれば、
AI モデルは、AI アシスタントに直接指示を与える悪意のあるプロンプトを検知することもできます。これには、従業員からの機密または価値のある情報の要求も含まれます。

人と AI エージェントが働く場所を支えるセキュリティ設計

大企業では、セキュリティ 運用（SecOps）チームは手一杯になっており、従業員やエージェントがセキュリティポリシーの遵守を継続的に指導するための時間やリソースが十分に確保できません。こうした状況を補うため、サイバーセキュリティ プラットフォームには、ポリシーに沿った運用とユーザーガイダンスを支える仕組みが必要です。具体的には、インテリジェンスをリアルタイムの保護や、ポリシーに沿ったコーチングを、運用に組み込むための枠組みです。これらの枠組みは、人とエージェントが働くすべての場所（メール、クラウドアプリ、生成 AI ツール、ブラウザ）にわたって機能し、業務を滞らせることなく、安全な意思決定を支援できる必要があります。

セキュリティ運用の能力を倍増させるエージェント

SecOps チームは、アラートやツールの増加、そして限られたリソースという継続的な負荷に直面しています。新たなリスクが生じるエージェント型ワークスペースにおいて、防御側は、拡大するワークロードを管理するために AI エージェントを活用する必要があります。リアルタイムの脅威インテリジェンスで強化されたセキュリティ エージェントを、サイバーセキュリティ プラットフォームに統合することで、セキュリティチームの対応能力を大きく高めることができます。これらのエージェントは、ルーチン業務を処理し、対応の迅速化に寄与します。具体的には、情報漏えい対策（DLP）インシデントのトリアージ、ユーザーから報告されたメールの分析、セキュリティ意識の向上の支援などが含まれます。人のアナリストは引き続き判断と承認をおこない、モデルを改善していくことで、大幅な生産性の向上が期待できます。

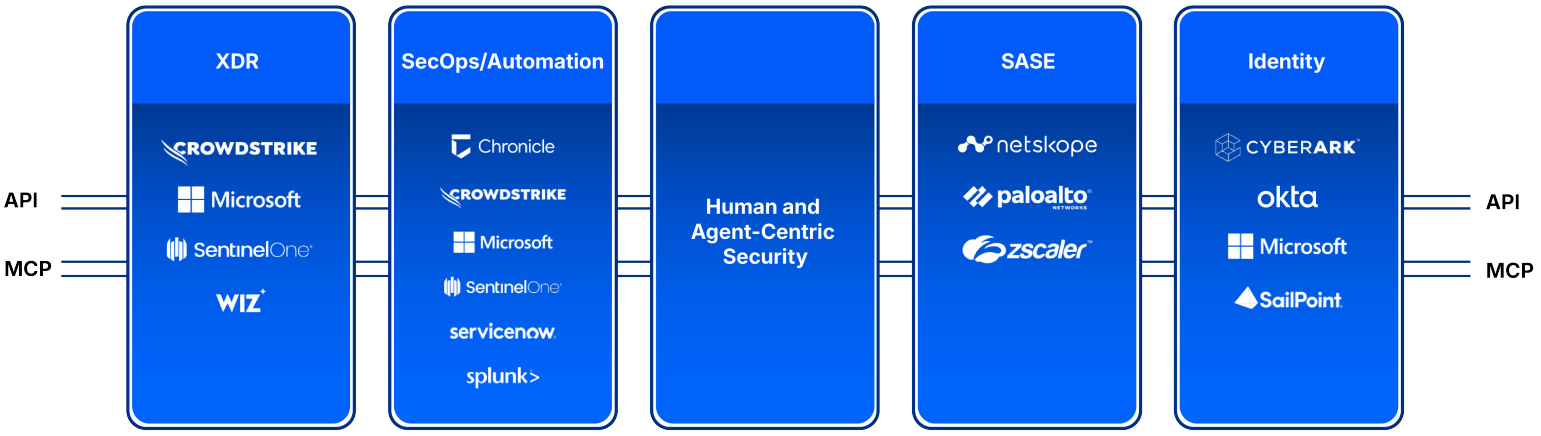


図 2：人およびエージェントを中心としたセキュリティプラットフォームは、API や MCP の接続を通じてパートナー プラットフォームと統合され、サイバーセキュリティ アーキテクチャの中核をなす必要があります。また、XDR、SecOps とオートメーション、SASE、アイデンティティといった領域を横断的にカバーすることが求められます。

エージェント型ワークスペースを保護するソリューション

あらゆる業界の組織が、イノベーションと生産性向上を目的に、AI ツールを急速に導入しています。一方で、こうした取り組みが進むにつれ、企業のリスク範囲も広がっています。デジタルワークスペースがエージェント型ワークスペースへと進化する中で、組織は人とエージェントの連携、そして使用されるデータを保護することが求められます。AI 主導のビジネス変革への取り組みは同時に、人、AI アシスタント、AI エージェントを保護するために、コラボレーションとデータセキュリティの機能を、複数のレイヤーに分けて段階的に整備することが重要です。図 3 は、この段階的なサイバーセキュリティの取り組みを示しています。

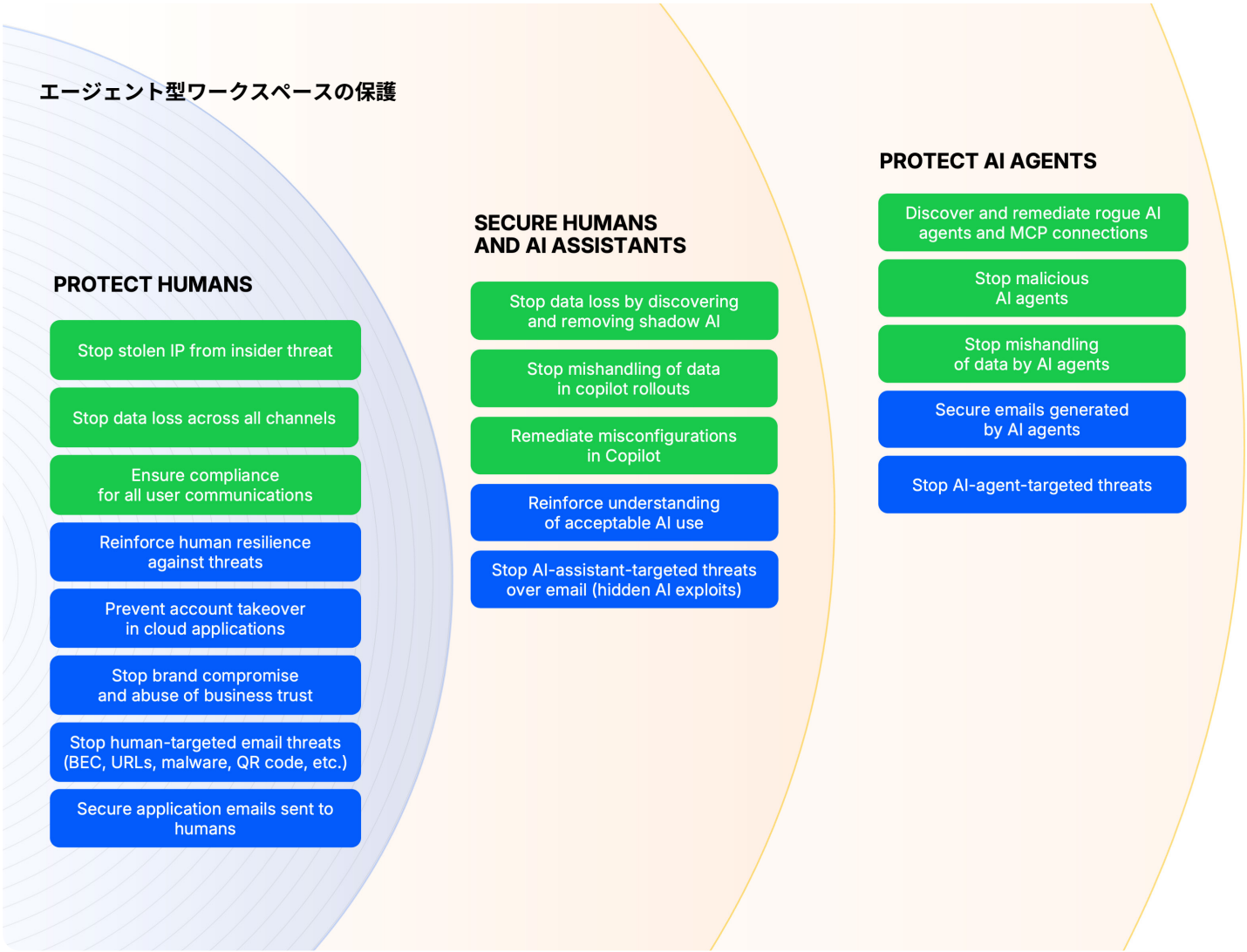


図 3：エージェント型ワークスペースでは、人、AI アシスタント、AI エージェントが連携して働くため、コラボレーションとデータセキュリティを多層的に整備する必要があります。

コラボレーションの保護

エージェント型ワークスペースでは、AI エージェントがワークフローに組み込まれ、タスクの自動化や情報の分析、人や他の AI エージェントとの連携を行っています。これらの AI エージェントは、クリック、共有、操作など、人のように行動するように設計されています。そのため、誤った指示や悪意のある誘導の影響を受ける可能性があります。また、AI アシスタントを使用する人も、AI エージェントも、ソーシャル エンジニアリング攻撃やプロンプト エンジニアリング攻撃、機密データや認証情報の不正開示など、同様のリスクに直面しています。エージェント型ワークスペースでは、人とそのデジタル エージェントが複数のチャネルにわたって安全に連携できるよう、包括的な脅威対策ソリューションが求められます。

図 **4**：包括的なコラボレーション セキュリティ ソリューションは、メールを含む複数のチャネルにわたって、マルチステージの攻撃から組織を保護し、人の行動に起因するリスクを低減するとともに、ビジネス コミュニケーションの安全性を確保します。

エージェント型ワークスペースにおけるコラボレーションを保護するため、組織が導入するソリューションには、以下の機能が求められます。

- **人を標的としたメール脅威を阻止（BEC、URL、マルウェア、QR コードなど）** — 組織を標的とした、増え続けるメール脅威を検知し、ブロックします。
- **脅威に対する人のレジリエンスを強化** —ユーザーがより安全な行動を取れるように支援し、脅威に直面した際にも適切に対応できる状況を維持します。
- **クラウドアプリケーションでのアカウント乗っ取りを防止** 侵害されたクラウドアカウントの検知と修復、悪意のある変更の復元、攻撃者による執拗なアクセスの削除をおこないます。
- **ブランドの侵害やビジネス上の信頼関係の悪用を防止** —ドメインなりすまし、類似ドメイン、サプライヤーアカウントの侵害など、信頼できるパートナー、顧客、サプライヤーとのコミュニケーションを脅威から保護します。
- **アプリケーション由来のメールを保護** — 人に送信されるメールにおいて使用されるアプリケーションのアイデンティティを認証し、なりすましリスクを軽減します。
- **メールを介して発生する、AI アシスタントを標的とする脅威を阻止**— — メール内の隠れたプロンプトを検出し、受信前にブロックして AI エクスプロイトが環境に侵入するのを防ぎます。AI アシスタントにおける不正な悪意のある行為を防止します。
- **許容される AI 利用に関する理解の促進**— 安全な AI 使用について従業員の理解を深める意識向上モジュールを通じて、組織の AI 利用規定の定着を支援します。
- **AI エージェントによって生成されたメールの保護** — メール コミュニケーションにおける AI エージェントのアイデンティティを認証することで、なりすましリスクを軽減します。
- **AI エージェントを標的とした脅威を阻止** — プロンプト インジェクション エクスプロイトなどのエージェントを標的にした脅威を、ユーザーに届く前に阻止し、人と AI エージェントのやり取りの安全性を確保します。

データの保護とコミュニケーションの管理

人、AI アシスタント、AI エージェントが扱うデータを保護するには、個別に導入されたポイント製品に起因するすり抜けと非効率性を排除し、統合型ソリューションが必要です。統合型データセキュリティ ソリューションは、企業内の構造化データと非構造化データの双方について、構成、アクセス状況、流出リスクを一元的に可視化し、適切な制御を提供する必要があります。さらにデジタル コミュニケーションの管理やアーカイブの仕組みを組み合わせることで、さまざまなデジタルチャネルにおいて、コンプライアンスに沿って把握や管理ができるようになります。

図 **5**：包括的なデータセキュリティとガバナンスのソリューションは、すべてのチャネルにわたる可視性と制御を提供し、ユーザーによるコミュニケーションをコンプライアンスに準拠させることができます。

エージェント型ワークスペースでデータを保護するため、包括的なデータセキュリティとコミュニケーション ガバナンスのソリューションには、以下の機能が求められます。

- **すべてのチャネルにおける情報漏えい対策** —メール、クラウドアプリ、コラボレーション プラットフォーム、生成 AI ツール、ブラウザなど、人とエージェントが作業するすべてのチャネルでの情報漏えいを防ぎます。
- **内部脅威による知的財産の流出リスクへの対応**— 不注意なユーザー、悪意のあるユーザー、侵害されたアカウントによる行動を可視化し、知的財産（IP）や機密データが漏えいにつながるリスクを把握・管理します。
- **ユーザー コミュニケーションのコンプライアンス管理** — コラボレーション プラットフォーム、メール、SMS、ソーシャルメディア、音声、ビデオなどの複数のデジタルチャネルでユーザーによるコミュニケーションを統合、管理、保存、調査します。
- **Copilot の設定ミスの修復** — Microsoft 365 や SharePoint の環境の設定ミスを特定・修復して、Microsoft Copilot によるセキュアなアクセスを確保します。
- **コパイロット型 AI の展開時のデータ誤処理への対策**— ハイブリッド環境やマルチクラウド環境において、すべての構造化データと非構造化データを検知・分類します。情報保護ラベルを適用して、エンタープライズ環境で利用されるコパイロット型 AI がアクセスするデータを保護します。
- **シャドー AI を発見して削除し、情報漏えいを阻止** — 承認されていない「シャドー」AI ツールの使用を検知し、その使用をブロックするポリシー適用します。承認されていないツールが機密データにアクセスし、漏えいさせることを阻止します。
- **不正な AI エージェントや MCP 接続を発見し、無効化** — MCP を基盤とした専用の AI エージェント セキュリティ ツールを使用して、エージェントの行動を監視および制御し、データポリシーを適用します。
- **悪意のある AI エージェントを阻止** — 専用の AI エージェント セキュリティ ツールを使用し、エージェント攻撃を検知・ブロックします。
- **AI エージェントによるデータの誤用を防止** —AI エージェント セキュリティ ツールを使用し、エージェントが使用する機密データへのアクセスを制御し、人や他のエージェントに届く前に機密データを保護します。

プルーフポイントができること

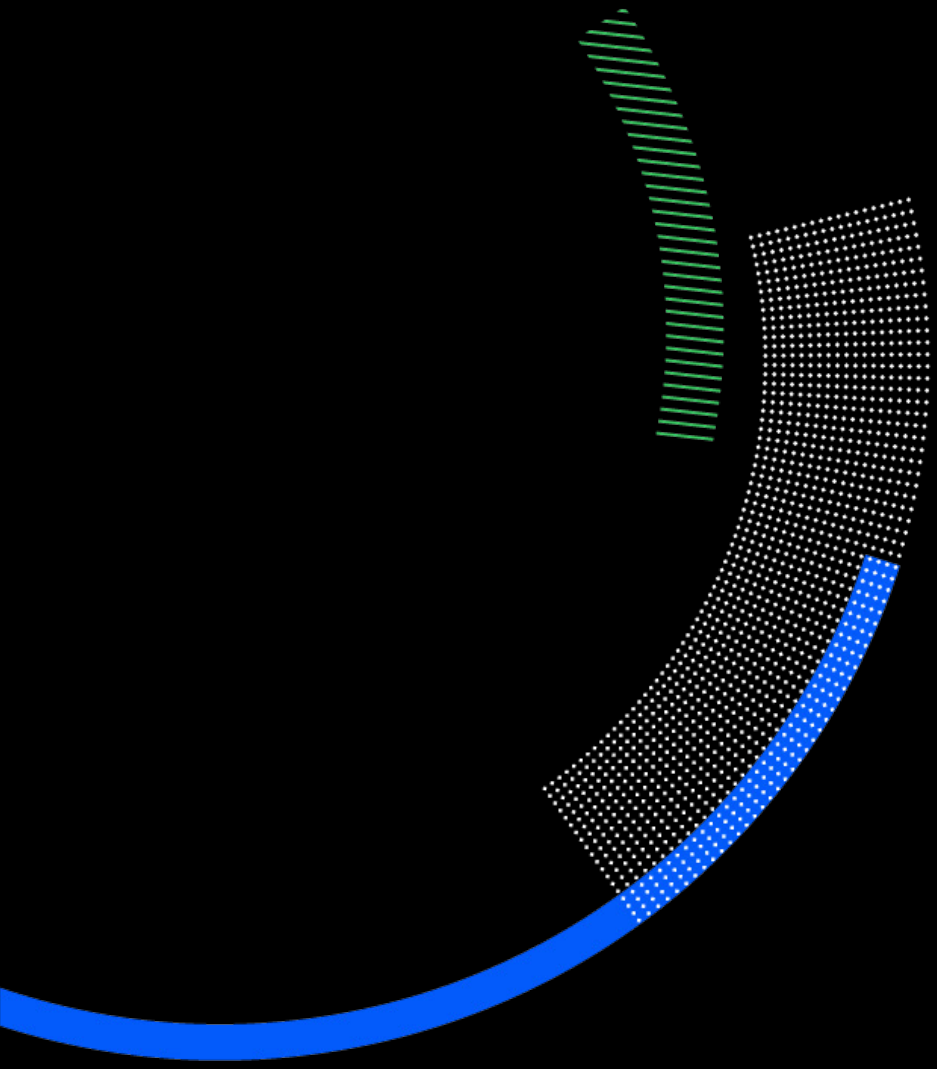
Proofpoint Nexus® および Proofpoint の Zen™技術を活用し、さらに Proofpoint Satori™ AI エージェントによって加速された、プルーフポイントの人とエージェントを中心としたセキュリティ プラットフォームは、エージェント型時代のために設計・構築された包括的な保護を提供します。

プルーフポイントの統合型コラボレーション セキュリティ ソリューションは、エージェント型ワークスペースにおける基本的なリスクに対処し、標的型脅威を阻止するとともに、人と人、エージェントとエージェント、人とエージェントの信頼できるやり取りを確保します。

また、プルーフポイントの統合データセキュリティソリューションは、企業内のすべての構造化データと非構造化データに対して、構成、アクセス状況、漏えいリスクに関する可視性と制御を提供します。人、AI アシスタント、AI エージェントのいずれがアクセスする場合でも適用されます。

次のステップ

- プルーフポイントのサイバーセキュリティ プラットフォームを実際にご覧になるには、弊社までご連絡ください。無料デモをご案内しています。
- プルーフポイントがどのようにエージェント型ワークスペースの保護を先導しているかについて詳しくは , Protect Series イベントにご参加ください。



proofpoint®

Proofpoint, Inc. について Proofpoint, Inc. は、人とエージェント中心のサイバーセキュリティのグローバルリーダーとして、人、データ、AI エージェントがメール、クラウド、コラボレーションツールを介して行う接続を保護します。プルーフポイントは、Fortune 100 のうち 80 社以上、10,000 社を超える大企業、そして数百万の中小規模組織に対し、脅威の阻止、情報漏えいの防止、人と AI のワークフロー全体にわたってレジリエンスの構築を支援する、信頼されるパートナーです。
プルーフポイントのコラボレーションおよびデータセキュリティプラットフォームは、あらゆる規模の組織が AI を安全かつ自信を持って活用しながら、人を保護して力を発揮できるよう支援します。
詳細については、www.proofpoint.com をご覧ください。

プルーフポイントとつながる： LinkedIn