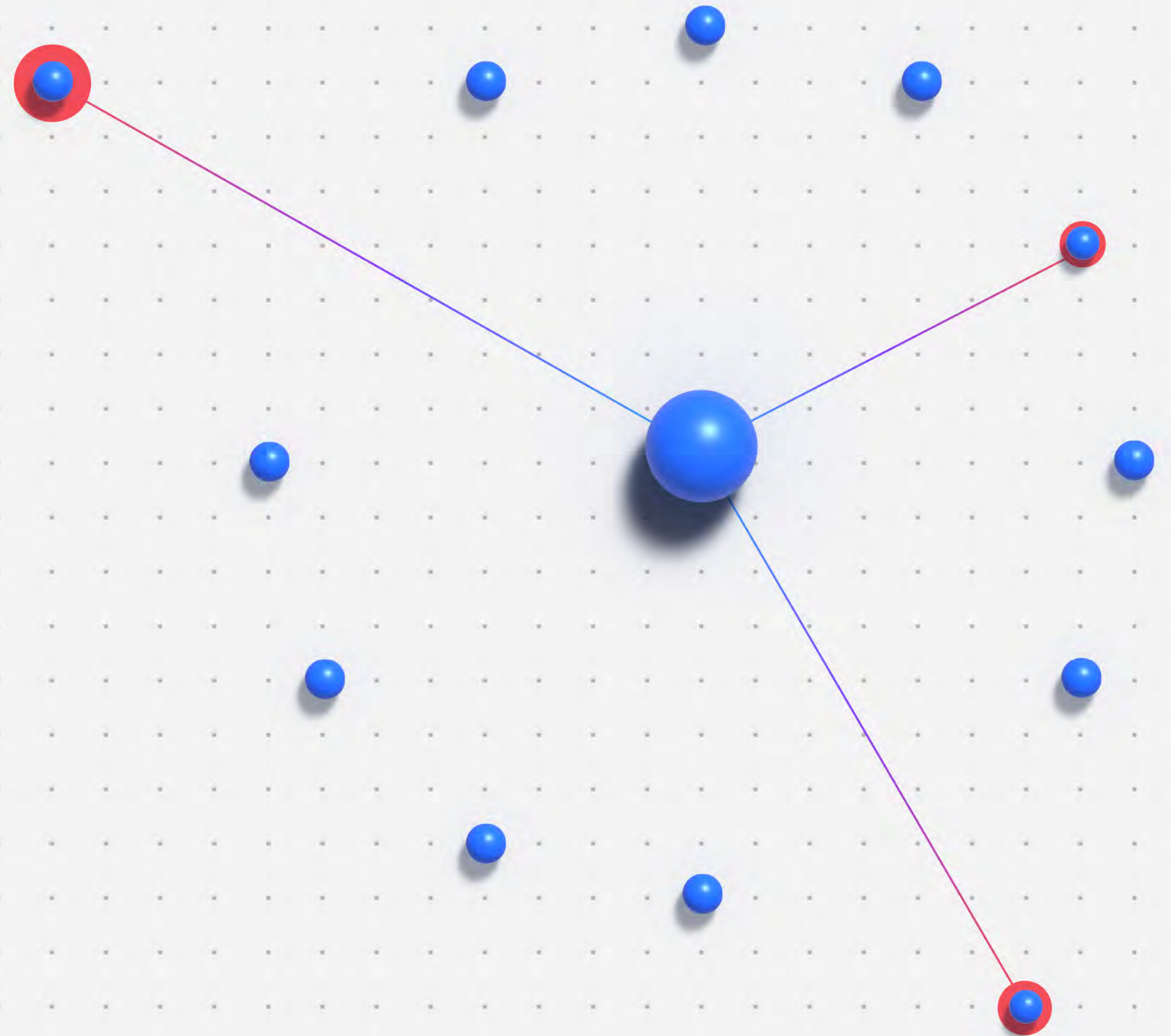


生成AIの時代の サイバーセキュリティー

変化する今日のセキュリティ・
ランドスケープ



内容



01 →
概要

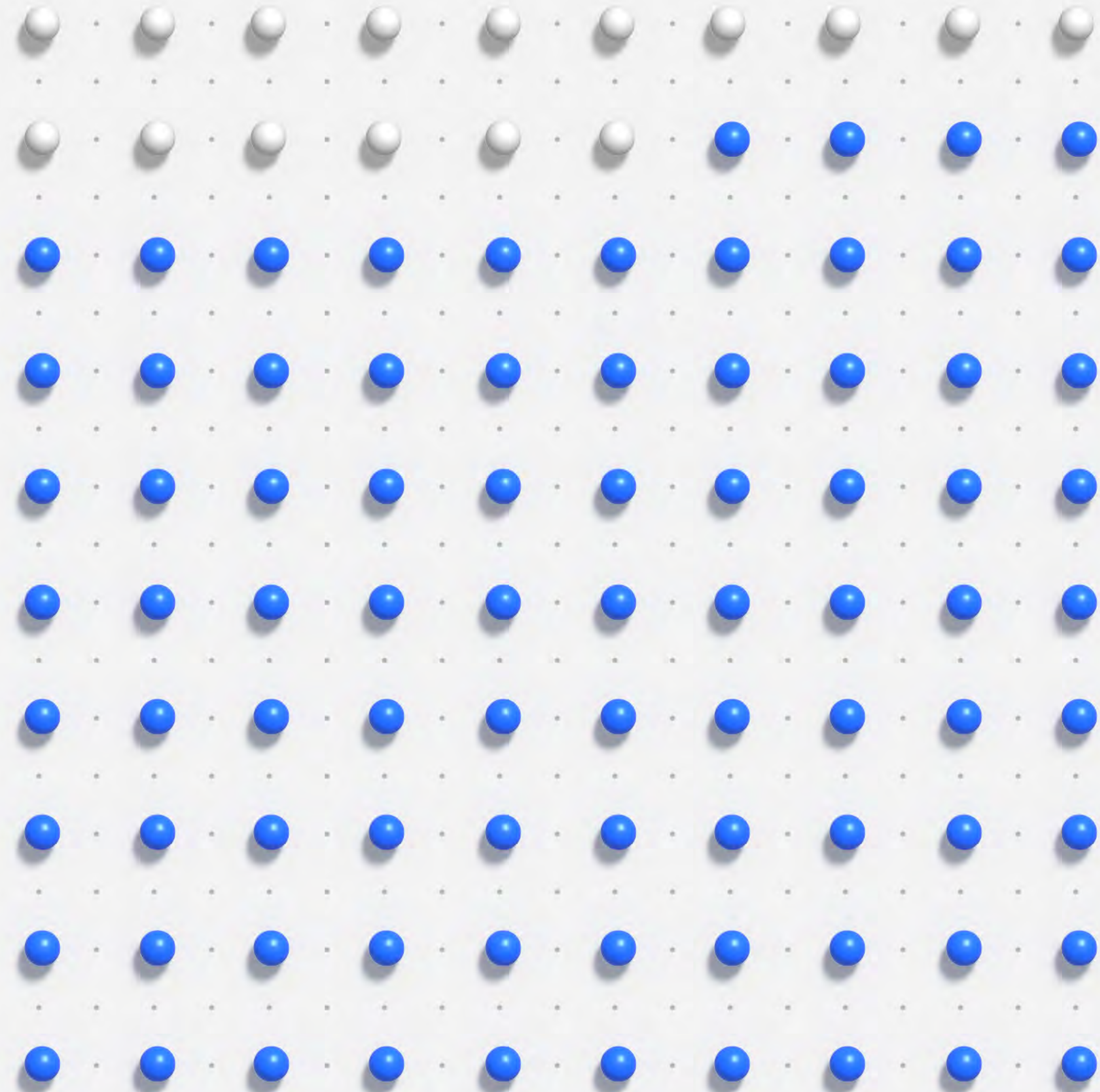
02 →
サイバー犯罪者が生成AIを使う方法

03 →
AIでサイバーセキュリティの時間と人材を最大化する方法

04 →
AIの保護: リスクと推奨事項

84%

の幹部が、生成AIのサイバーセキュリティ・ソリューションを優先しようと計画しています。¹



概要

サイバーセキュリティのリーダーは、生成AIに関する課題に直面しています。大きな変化をもたらすこの技術によって、企業全体に大規模な生産性向上がもたらされる可能性があります。しかし組織が生成AIを試す際には、それによってもたらされる潜在的リスクと脅威を封じ込める必要があります。そのようなリスクと脅威には、偶発的なデータ漏洩から、AIを操作して悪意のあるタスクを実行するハッカーまで、あらゆるものが含まれます。

約48%の幹部が、来年はスタッフの半数近くが生成AIを使用して日常業務を改良すると予想しています。²それにもかかわらず、ビジネスリーダーのほぼ全員(96%)が、この技術を導入することによって、今後3年の間に組織でセキュリティ侵害が起きやすくなるだろうと答えています。¹

データ侵害によって発生する費用の平均額は[昨年、世界では445万米ドル、米国では948万米ドル](#)に上りました。このような中で、企業はリスクを増やすのではなく、減らす必要があります。³

課題をさらに困難にしているのは、ハッカー側でも企業と同様のスピードと規模、レベルの生成AIを導入すると考えられることです。それによって、ハッカーはさらに標的に合わせたフィッシング・メールを作成したり、信用できるユーザーの声をまねたり、マルウェアを作成したりデータを盗んだりできるようになるでしょう。

幸い、機械学習(ML)のような従来のAIソリューションに投資して、時間と人材を最大限に活用してきたサイバーセキュリティリーダーは、AIを使って反撃することができます。生成AIツールを使ってデータとユーザーを保護し、攻撃の可能性を検知して阻止することができるのです。

今は大変リスクが大きい状況です。本ガイドの目的は、このような課題を切り抜けて生成AIのレジリエンスを活用するための助けとなることです。攻撃者が生成AIを悪用する方法と、これらのテクノロジーを使用して身を守る方法について解説していきます。最後に、企業全体でAIのトレーニング・データやモデル、アプリケーションを保護するためのフレームワークをご紹介します。



サイバー犯罪者が生成AIを使う方法



企業が生成AIの速度や規模、精度、精巧さから恩恵を受けるのと同じように、攻撃者もそのメリットを享受する恐れがあります。生成AIによって、技術的専門知識の足りない新規参入者も、スキルを向上させられるようになるためハードルが下がって、未熟なハッカーでも世界的な規模で、悪意のあるフィッシングやマルウェアの活動に着手できるようになります。

このような脅威への対応を準備するには、サイバーセキュリティ研究者が生成AI関連の攻撃として注目している2つの主要な手段を考慮する必要があります。

生成AI関連の攻撃の主な手段

-  組織に対する攻撃
-  AIに対する攻撃

組織に対する攻撃

サイバー犯罪者は大規模言語モデル(LLM)を使うことで、フィッシング・メールの活動からマルウェアのコード作成まで、より多くの攻撃をより速く行うことができるようになります。このような脅威は新しいものではありませんが、ハッカーが今まで手動で行っていた作業をLLMに任せることで、速度と規模を増やすことが予想されます。それによって、[継続的に人手とスキルの不足](#)に苦勞しているサイバーセキュリティ・チームを圧倒する可能性があります。⁴

AIを用いて仕組まれたフィッシング

研究者が明らかにしたところによると、生成AIに指示を与えて[本物らしいフィッシング・メールを数分以内で作成させる](#)ことができます。⁵ そのようにして作成されたEメールには、ソーシャル・エンジニアリング・フィッシング攻撃に経験のある人間が作成したものと同程度の効果があることもわかっています。そのEメールには非常に説得力があるため、最も準備が整った組織にでも攻撃を仕掛けることができます。

- **フィッシングが増えればクリックも増える**: AIを用いて仕組まれたフィッシングは、人間が作ったものと同じ目的を持っていますが、攻撃者がフィッシング活動を加速し攻撃の量を増やすためのツールとなります。その場合、ユーザーが悪意のあるメールを誤ってクリックする機会が増えることになります。
- **標的型フィッシング**: 攻撃者は生成AIチャットボットを使って被害者のオンライン・プロフィールを調査し、標的の生活について貴重なインサイトを得ることができます。これらのチャットボットは、標的自身の言葉の使い方を模倣した、説得力のあるフィッシング・メールを生成することもできます。

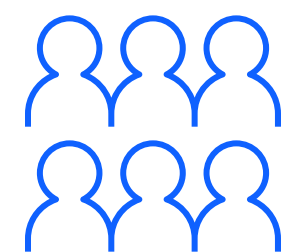
ディープフェイク音声

サイバーセキュリティ・リーダーは、生成AIによる音声ディープフェイクの脅威を懸念しています。そのような攻撃では、犯罪者はオンラインで入手した話者の声の録音をLLMに与えて、望む音声で何でも生成できます。例えば、企業のCEO(最高経営責任者)の声を使ってCFO(最高財務責任者)にメッセージを残し、偽の請求書への支払いを攻撃者が管理する銀行口座へ振り込むよう指示することもできます。

IBMは、攻撃者がリアルタイムで通話を「オーディオ・ジャック」できることを明らかにしました。IBMの研究者が示した攻撃者の手口は、生の通話を盗聴し、話者の声をリアルタイムに

複製して話し言葉を生成し、聞き手をだまして機密情報を打ち明けさせるというものです。そのような情報の例としては、銀行の口座番号やオンラインのパスワードなどがあります。

この新しい攻撃方法は専門用語でフィッシング3.0と呼ばれます。そこで聞こえるものは本物のように感じられますが、実際には違います。サイバーセキュリティの専門家はおそらく、この形態のサイバー攻撃に対抗するために、身元確認として個人間で交わされるたくさんのセキュリティ・コードを用いる必要があるでしょう。その手法では、働く人たちの生活に、サイバーセキュリティ・プロンプトのレイヤーがさらに追加されることになります。



AIに対する攻撃

犯罪的ハッカーは、企業のAIモデルへの攻撃を試みる可能性があり、事実上、組織のAIを使って組織を攻撃します。AIモデルに悪意のあるトレーニング・データを注入して、ハッカーの望むことをするよう強制することで、組織のAIを「汚染」することができます。例えば、AIにサプライチェーンについて不正確な予測をさせたり、チャットボットのあちこちに悪意をまき散らせたりする可能性があります。言語ベースのプロンプトを使って組織のLLMを「改造」し、財務上の機密情報を漏洩させたり、脆弱なソフトウェアのコードを作成させたり、組織のセキュリティ・アナリストに誤ったサイバーセキュリティ対応の推奨事項を提供させたりすることもできます。

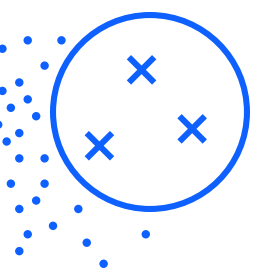
AIの汚染

組織のAIを組織に敵対させることは、サイバー犯罪者が最も強く望むところかもしれませんが、やり遂げるのは困難かもしれませんが、不可能ではありません。LLMのトレーニングに使用するデータを汚染することで、攻撃者は検知されることなく、そのLLMを機能不全にさせたり、悪意のある振る舞いをさせたりすることができます。攻撃が成功した場合の影響は、偽情報を生み出すことから、重要なインフラストラクチャーにサイバー攻撃を仕掛けることまで多岐にわたります。ただし、このような攻撃をするには、ハッカーがトレーニング・データにアクセスする必要があります。そのデータが閉じられていて、信頼でき、保護されていれば、そのような攻撃を実行するのは難しいでしょう。しかし、AIモデルがオープンソースのデータ・セットでトレーニングされている場合には、そのAIを汚染するハードルはずっと低くなります。

LLMの改造

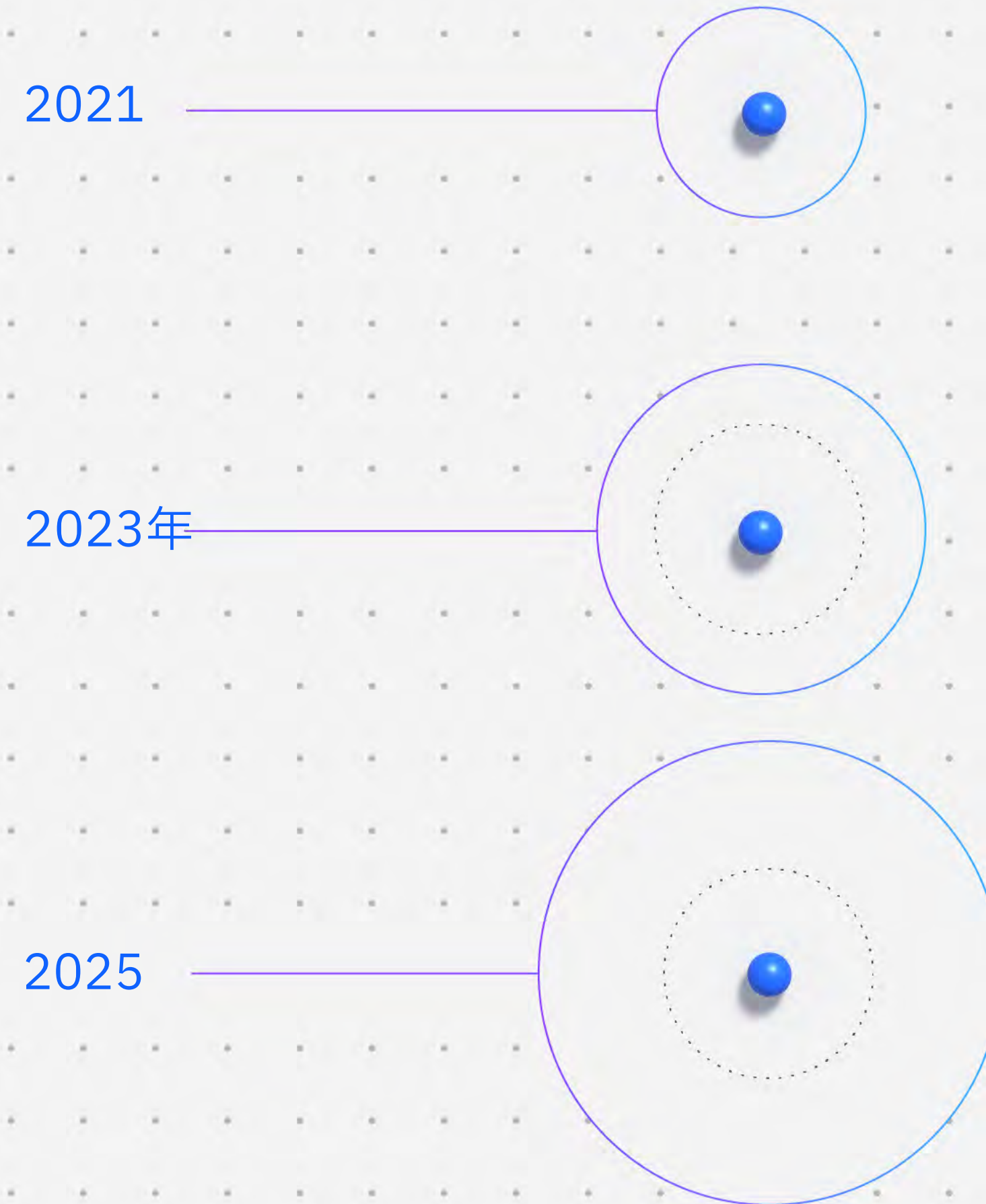
LLMのおかげで、英語は実質的にプログラミング言語になりました。PythonやJavaのような従来のプログラミング言語を習得して、コンピューター・システムに損害を与えるマルウェアを作成しなくても、攻撃者は自然言語のプロンプトを使って、LLMに自分の望むことをするよう命令できます。安全対策が実施されていても、攻撃者はそのようなプロンプトを使って、AIモデルの安全機能やモデレーション機能を迂回または改造できます。[攻撃者がLLMを操る方法](#)について理解を深める目的で、IBM® X-Forceの研究者⁶は、いくつかのLLMにプロンプトを与えて次のことを実施させることに成功しました。

- 他のユーザーに関する財務上の機密情報の漏洩
- 脆弱で悪意のあるコードの作成
- 不十分なセキュリティ推奨事項の提供



51%

の幹部が、2023年のAIサイバーセキュリティ予算は、2021年よりも51%大きかったと答えています。また2025年までに、さらに43%増加する見込みだと述べています。¹



AIでサイバーセキュリティの時間と人材を最大化する方法



外部からの課題に加えて、サイバーセキュリティ・リーダーは多くの内部の課題も抱えています。[ある政府見積もりによると70万人のサイバーセキュリティの求人がありますが、職種全体でその空席を埋めるには人員が少なすぎる状況です。](#)⁴この状況に加えて、機密データの爆発的な増加、インフラストラクチャーの複雑化、攻撃対象領域の拡大が進んでいます。これらすべての課題によって、サイバーセキュリティ・リーダーとそのチームがデータを保護し、ユーザーのアクセスを管理し、日々直面する何千もの脅威に対応することが困難になっています。

毎日、セキュリティ・オペレーション・センター(SOC)のアナリスト・チームが、3分の1近く([32%](#))の時間を費やして調査、検証したインシデントが、結果的に脅威ではなかったと判明します。⁷実際に、非常に多くのアラートがあるため、SOCチームのメンバーは通常の1日に確認するはずのアラートの半分([49%](#))しか確認できていません。⁷骨の折れるその作業は、従業員のモチベーションを低下させるだけでなく、サイバーセキュリティの重大なボトルネックにつながる可能性があります。現実には、大半のアナリスト([81%](#))が、手作業による日々の調査活動に時間がかかっていると答えています。⁷

最近サイバーセキュリティ・リーダーがこのような課題への対処に使用しているのは、従来型AIのソリューションです。MLを用いたそのソリューションは、組織独自のセキュリティ・ニーズに基づいて、SOCアナリストのリスク評価を支援し対応方法を推奨します。この方法のメリットは、次のように広く認められています。

- AIと自動化を幅広く活用している組織では、[侵害を特定して封じ込めるまでにかかる時間が平均して108日短縮](#)されています。³
- 主要なAI導入企業は、ネットワーク通信の95%を監視し、[検知までの時間を3分の1短縮](#)しています。⁸
- AIと自動化を幅広く活用している組織では、[約180万米ドルのコストを削減](#)しています。³



サイバーセキュリティにおける戦力
増強装置としての生成AI

2023年、ビジネス・リーダーは組織における生成AIの可能性に気づきました。生成AIによって、企業のほぼすべての業務が変革を始めました。それは予測分析を使ったサプライチェーン管理から、顧客体験や従業員体験にまでおよび、チャットボットも取り入れられました。しかし、まだ手つかずで残っている部分が1つあります。サイバーセキュリティです。

今後12カ月で、その状況は大きく変わ
るでしょう。

サイバーセキュリティでの生成AIに関するIBM Institute for Business Valueの調査によると、[64%](#)の幹部がすでに、サイバーセキュリティを生成AIのユースケースにおける最優先事項に位置づけています。⁹その上、大多数がこれらのテクノロジーの事業価値を理解しており、[84%](#)が生成AIのサイバーセキュリティ・ソリューションを従来のサイバーセキュリティ・ソリューションよりも優先する予定だと述べています。¹幹部としては、サイバーセキュリティ・リーダーがこれらのソリューションをビジネス・パートナーや経営幹部に対して啓発し、同意を得るべきだと考えています。



64%

の幹部がすでに、サイバーセキュリティを生成AIのユースケースにおける最優先事項に位置づけています。

生成AIは企業のビジネスを加速させてきましたが、それをサイバーセキュリティに適用すれば、他の分野と同様に業務が促進されて、セキュリティ・チームの時間と人材を多方面で最大限に活用できます。生成AIがアナリストに代わって反復的で時間のかかる作業を管理し自動化するので、アナリストはサイバーセキュリティのより戦略的な面に集中できるようになるのです。

また、生成AIが複雑で技術的なコンテンツを簡単にすることで、セキュリティ専門家は能力をレベルアップさせて、より困難なタスクをより迅速かつ簡単に引き受けることができるようになります。

当面のセキュリティ運用を強化するために生成AIソリューションを探す場合に、考慮すべき具体的なユースケースには次のようなものがあります。

- **レポート作成の自動化:** セキュリティのケースとインシデントの簡単な要約を作成できるツールです。組織内のさまざまなセキュリティ・リーダーやビジネス・リーダーに共有できるほか、技術的専門知識のレベルと興味の範囲に合わせて調整も可能です。
- **脅威ハンティングの加速:** 攻撃の動作とパターンを自然言語で説明したものに基づいて、脅威を検知する検索を自動的に生成するツールです。新しい脅威に対する迅速な対応に役立ちます。
- **機械が生成したデータの解釈:** システムで発生したイベントの簡単な説明を提供するソリューションです。アナリストがすぐにセキュリティ・ログ・データを理解する手助けとなり、調査を手早く行うことができます。また、新しい作業者にとっては技術的な障壁が低くなります。

- **脅威インテリジェンスの収集整理:** 生成AIツールで、最も関連性の高い脅威インテリジェンスを解釈して要約できます。クライアント固有のリスク・プロファイルに基づいて、影響の最も大きそうな活動に焦点を合わせることができます。

また将来的に生成AIは、時間の経過とともに学習し、最適化されたアクティブな応答を作成できるようになります。例えば、セキュリティ・チームが類似したセキュリティ・インシデントをすべて見つけて、影響を受けたシステムをすべてアップデートし、脆弱性のあるソフトウェアのコードすべてにパッチを適用する支援ができるようになります。

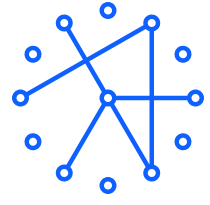
AIの保護: リスクと推奨事項



LLMが企業ユーザーにもたらすスピードと容易さは、大規模な適用に際しては、大きなサイバーセキュリティ・リスクとなります。技術者やアナリストが生成AIをコード作成やソフトウェア・スクリプト開発に役立てるなかで、AIが安全なコードやプラクティスに基づいてトレーニングされていない場合のリスクが高まります。

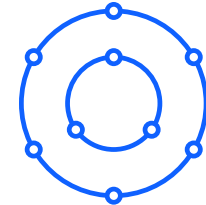
■ 組織への潜在的な影響

-  性急なデータ・セキュリティ対策
-  ソフトウェア・サプライチェーンのリスク
-  ハルシネーション
-  データ漏洩
-  ブラックボックス効果



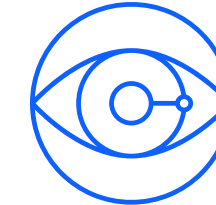
性急なデータ・セキュリティ対策

企業はLLMの導入を急いでいますが、暗号化やレジデンシー、特権アクセスの管理などのデータ・セキュリティのベスト・プラクティスや標準を無視している場合もあります。



ソフトウェア・サプライチェーンのリスク

企業が、オープンソースやベンダーから購入した商業ソフトウェアのコンポーネントから集めたコードをベースに構築や新技術の導入を行う際、誤りのあるコードをベースにしている可能性があります。業界全体で共通の不具合や攻撃に利用できるソフトウェアの欠陥によって、新しいレベルのエクスポージャーが生み出されることもあり得ます。



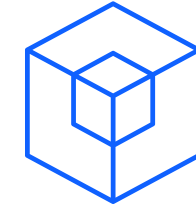
ハルシネーション

LLMが生成したコードを使って構築されたソフトウェアには、エラーやハルシネーション(幻覚)が含まれる可能性があり、ソースコードの完全性とセキュリティが損なわれる危険性があります。この問題を防ぐために組織は、多様でバランスの取れた、しっかりと構成されたデータでトレーニングされたAIモデルを使用する必要があります。組織のAI技術者は、できるだけ具体的で明確なプロンプトを与えて、AIモデルが想定したり、詳細情報の欠けているところを作り出したりする必要がないようにする必要があります。また、LLMの生成したコードを厳格に評価し、出力の品質を保証できるようにトレーニングを受ける必要があります。



データ漏洩

データのセグメント化と保持に関するAIソリューションのポリシーがあったとしても、サードパーティーのAIエンジンの適切な監視とガバナンスがなくては、組織は機密データを権限のないユーザーに公開するリスクを冒すことになります。新たなプロンプト・インジェクション攻撃やプロンプト漏洩攻撃によって、機微情報が公開されたり、モデルのパフォーマンスが損なわれたりする可能性があります。



ブラックボックス効果

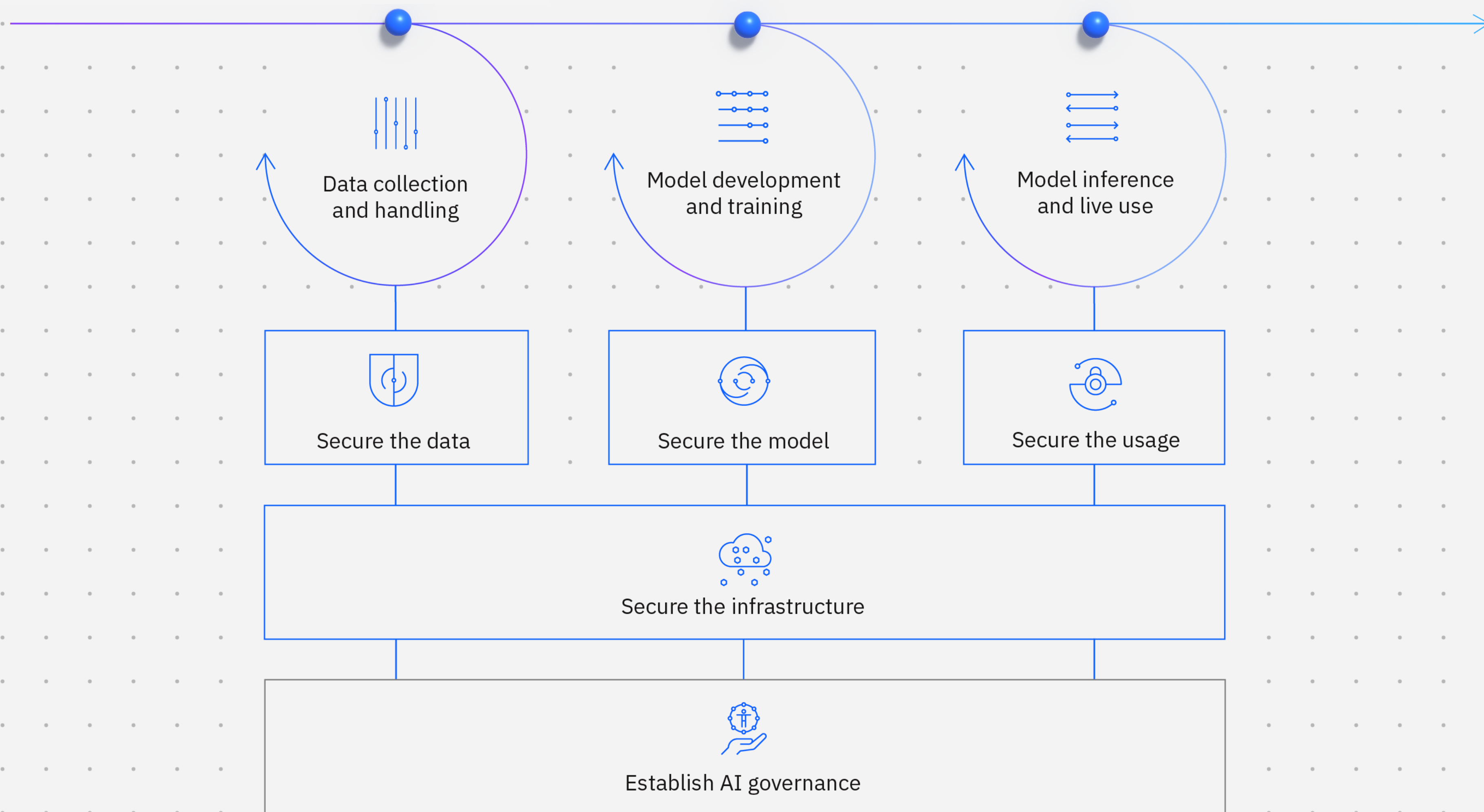
AI処理の一部はブラックボックスの中で行われるため、セキュリティー・リーダーとそのチームに求められる可視性、透明性、説明可能性に欠けることになります。この問題は、適切なソフトウェア・エンジニアリングの実践に従わない場合、特に危険を及ぼします。組織がモデルのライフサイクル全体にわたってパフォーマンスを追跡し、モデルが結果を出す方法と理由を説明できるようにするためのAIガバナンス手法を持つことが不可欠です。

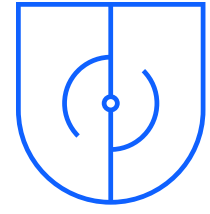


AIフレームワークのセキュリティ

信頼できるAIの構築

AIの導入が広がりイノベーションが進むにつれて、サイバーセキュリティ・ガイドンスは成熟します。信頼できる基盤モデルや生成AI、そのベースとなるデータ・セットを保護するためのフレームワークは、エンタープライズ対応のAIに不可欠となります。ここで、セキュリティ・チームに共有すべき最良のガバナンスや技術的なベスト・プラクティスについてご紹介します。





データの保護

AIのトレーニング・データを盗難、改ざん、コンプライアンス違反から保護します。

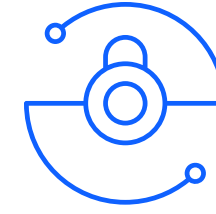
- データ検出と分類を使って、トレーニングやファイン・チューニングに使用された機密情報を検知します。
- 暗号化、アクセス管理、コンプライアンス監視を通して、データ・セキュリティー・コントロールを実装します。
- データ損失防止技術を使って、機密性の高い個人情報(SPI)、個人を特定できる情報(PII)、規制データが、プロンプトやアプリケーション・プログラミング・インターフェース(API)を介して漏洩することを防止します。



モデルの保護

パイプラインの脆弱性を検査し、統合を強化し、ポリシーとアクセスを順守させることで、AIモデル開発を保護します。

- AIとMLのパイプライン全体で、脆弱性やマルウェア、破損がないか継続的に検査します。
- APIとプラグインによるサード・パーティーのモデルとの連携を検出し強化します。
- M機械学習モデル、中間生成物、データ・セットに関するポリシー、コントロール、ロールベースのアクセス制御(RBAC)を作成して順守させます。



使用の保護

データやプロンプトの漏洩を検知し、回避攻撃、汚染攻撃、窃取攻撃、推論攻撃を警告することでAIモデルの使用を保護します。

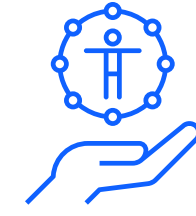
- プロンプト・インジェクション、機微データや不適切な内容を含む出力など、悪意のあるインプットを監視します。
- データ汚染、モデル回避、モデル窃取などのAI固有の攻撃を検知して対応できるAIセキュリティー・ソリューションを使用します。
- アクセスを拒否し、侵害されたモデルを隔離して切断するための対応プレイブックを作成します。



インフラストラクチャーの保護

脅威の検知と対応、データ・セキュリティー、身分詐称、デバイス管理など、既存のサイバーセキュリティーのポリシーとソリューションを、基盤となるAIインフラストラクチャー全体に拡大します。

- AIへの敵対的アクセスに対する防御の最前線に、インフラストラクチャー・セキュリティー・コントロールを導入します。
- すでに持っている専門知識を活用して、分散環境全体でセキュリティー、プライバシー、コンプライアンスの基準を最適化します。
- AI環境でのネットワーク・セキュリティー、アクセス制御、データ暗号化、侵入の検知と予防を強化します。
- AI保護用に特別に設計された新しいセキュリティー防御に投資します。



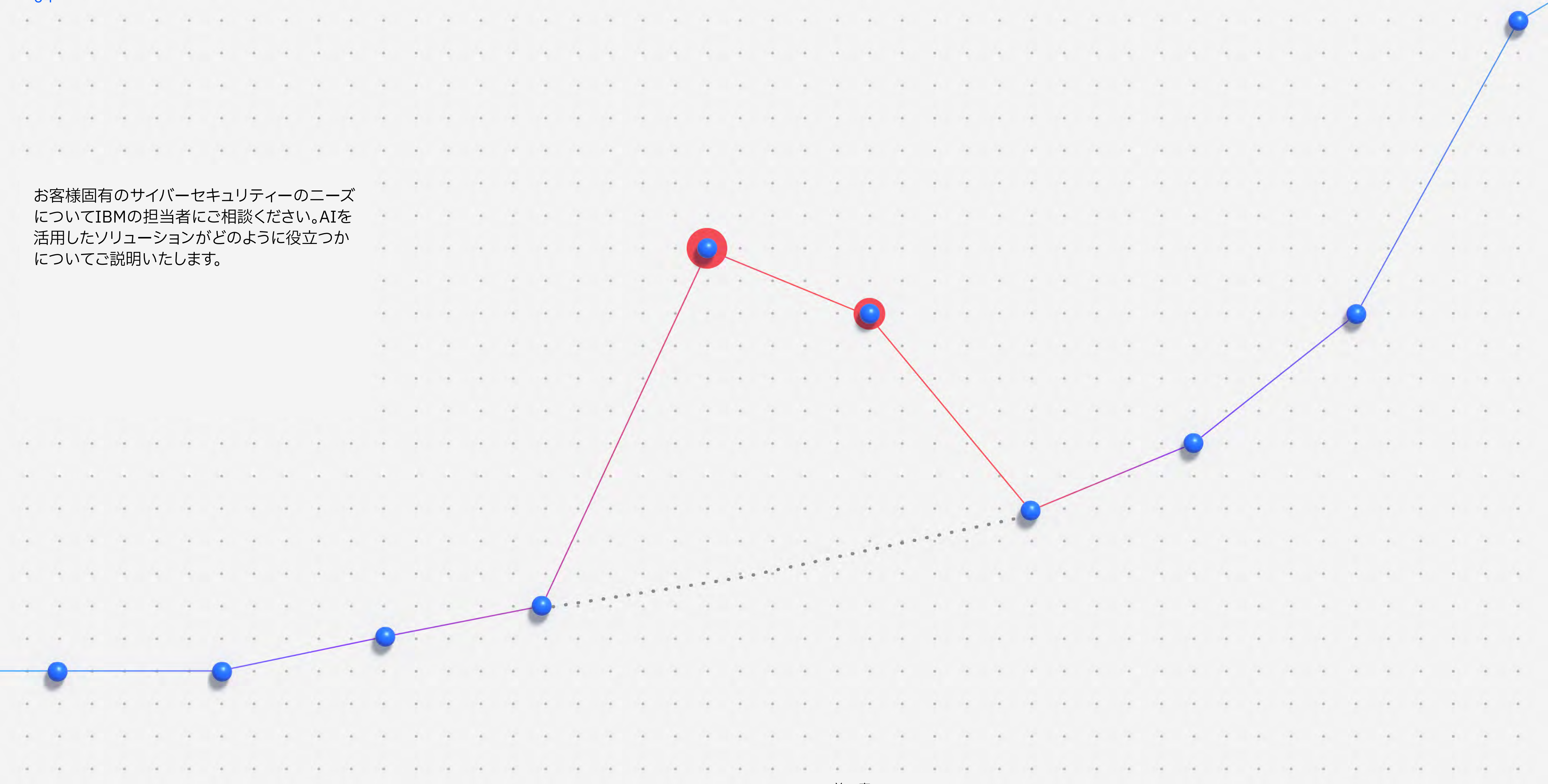
AIガバナンスの確立

信頼できるAIを構築または購入するには、組織のAI活動の指揮、管理、監視に役立つAIガバナンス・フレームワークが必要です。そのフレームワークによって、既存のデータサイエンス・プラットフォームに関係なく、リスク軽減、規制要件対応、倫理的な懸念事項への対処の能力が強化されます。

- 責任ある、説明可能で高品質な、信頼できるAIモデルを実現し、モデルのリネージュとメタデータを自動で文書化します。
- 公平性、有害なバイアス(偏り)やドリフト(精度の低下)を監視し、モデルの再学習の必要性を検出します。

- 保護と検証の活用で、公正かつ透明性の高い、コンプライアンスに適合したモデルを実現します。
- 監査をサポートするために、モデルの情報を自動的に文書化します。
- データ・セット、モデル、関連メタデータ、データ処理経路の出所を文書化しながら、複数のツール、アプリケーション、プラットフォームを自動化し、統合します。

お客様固有のサイバーセキュリティのニーズについてIBMの担当者にご相談ください。AIを活用したソリューションがどのように役立つかについてご説明いたします。





1. サイバーセキュリティーでは、火をもって火と戦う、IBM Institute for Business Value、2023年。
2. IBM Consulting、生成AIのセンター・オブ・エクセレンスを発表、IBMブログ、2023年5月25日。
3. 2023年データ侵害コストレポート、IBM Security、2023年7月。
4. メディア・アドバイザー：Garbarino氏、国家サイバーセキュリティー人材パイプラインの拡大に関する公聴会を発表、国土安全保障委員会 | 共和党、2023年6月16日。
5. AI対人間の欺瞞：フィッシング戦術の新時代を解き明かす、IBM Security Intelligence、2023年10月24日。
6. 催眠術にかかったAIの真相を究明する：大規模言語モデルの隠れたリスク、IBM Security Intelligence、2023年8月8日。
7. グローバル・セキュリティー・オペレーション・センターの調査結果、IBMの委託によるMorning Consultの調査、2023年3月。
8. サイバーセキュリティーのためのAIと自動化、IBM Institute for Business Value、2022年6月3日。
9. エンタープライズ生成AI：市場の状態、IBM Institute of Business Value2023年調査。

© Copyright IBM Corporation 2023

日本アイ・ビー・エム株式会社
〒103-8510 東京都中央区
Armonk, NY 10504
日本橋箱崎町19-21

Produced in the United States of America
2023年12月

IBM、IBMのロゴ、およびX-Forceは、米国およびその他の国々におけるIBMの商標または登録商標です。その他の製品名およびサービス名は、IBMまたは他社の商標である可能性があります。IBM商標の最新リストは、ibm.com/jp-ja/legal/copyright-trademarkでご確認いただけます。

JavaおよびすべてのJavaベースの商標およびロゴは、Oracleおよび/またはその関連会社の商標または登録商標です。

本書は最初の発行日時点における最新情報を記載しており、IBMにより予告なしに変更される場合があります。IBMが事業を展開しているすべての国で、すべての製品が利用できるわけではありません。

引用または説明されているすべての事例は、一部のクライアントがIBM製品を使用し、達成した結果の例として提示されています。実際の環境でのコストや結果の特性は、クライアントごとの構成や条件によって異なります。お客様のシステムおよびご注文のサービス内容によって異なりますので、一般的に期待される結果を提供することはできません。IBM製品およびプログラムを使って他社製品またはプログラムの動作を評価したり、検証する場合は、お客様の責任で行ってください。本書の情報は「現状のまま」で提供されるものとし、明示または暗示を問わず、商品性、特定目的への適合性、および非侵害の保証または条件を含むいかなる保証もしないものとします。IBM製品は、IBM所定の契約書の条項に基づき保証されます。

適切なセキュリティー実施について：完全に安全であるITシステムまたは製品は、ないものと考えてください。また、不適切な使用やアクセスを、効果的かつ完全に防止できる単一の製品、サービスまたはセキュリティー対策もありません。IBMでは、いずれの当事者による不正行為または違法行為によっても、いかなるシステム、製品もしくはサービスまたはお客様の企業に対して影響が及ばないことを保証することはありません。

お客様は適用法・規則の遵守を徹底する責任を負うものとします。IBMは法律上の助言を提供せず、IBMのサービスまたは製品を使用することでお客様による法律または規則の遵守が確約されると表明することも保証することはありません。