CLOUDFLARE

電子書籍

AIの安全利用に向けて

スケーラブルなAI戦略を立てる方法に関するCISO向けガイド

目次

概要

CISOへようこそ！

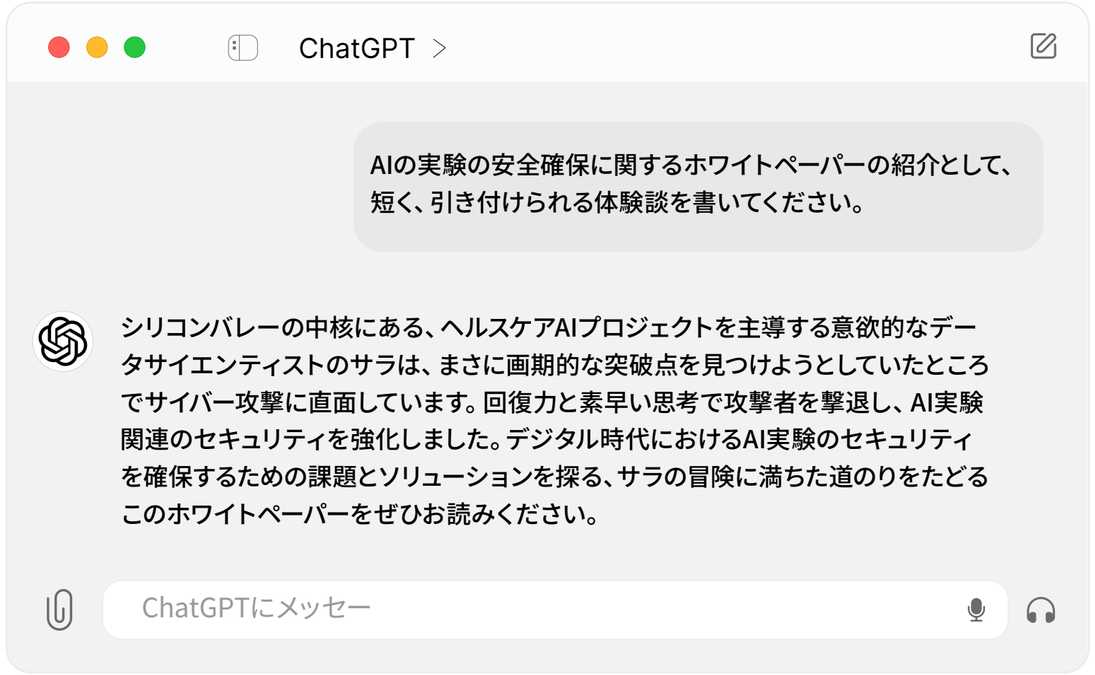
AIは最近最も流行している言葉であると同時に、セキュリティコミュニティにとって最も差し迫った問題の1つでもあります。その影響力には注意が必要で、だからこそ、Cloudflareは安全な生成人工知能（生成AI）の実験のためのガイドを作成しました。

AIツールは急激に発展しており、生活の中で身近な存在となりつつあり、さまざまな業界にイノベーションの機会がもたらされています。しかしながら、他のパラダイムシフトと同様に、生成AIでもセキュリティ、プライバシー、コンプライアンスにまつわる固有の課題が生じています。生成AIの急激な拡大により、想定外の利用の急増、ユーザーによる乱用、悪意のある行為、危険を伴うシャドーITの不適切な利用、データ漏洩や機密性の高い情報のリークといった問題につながる可能性があります。

社内でのAI活用が広がるにつれ、AIの使用や構築、安全性確保について、生成AI周りのルールや計画立案を予め進めていく必要があります。ここでは、成熟したレベルと使用法に基づいて生成AIの安全を確保するために使用できるリスクと見直しについてのヒントをご紹介します。こうした戦略を踏まえることで、組織はデータを保護し、コンプライアンスに準拠しながら、事業のニーズに合った生成AI戦略を展開できるようになります。

- Dawn Parzych、Cloudflare、製品マーケティング担当ディレクター

生成AIの実験を保護



残念ですが、サラの話はここで終わります。架空の人物の話はここまでとなりますが、予測AIと生成AIが拡大すれば、現実生活には数え切れないほどの「サラ」が登場することになります。それぞれがITチームや開発者チームのヒーローとして、ビジネス技術者や個々の従業員として機能することになるでしょう。

AIは、技術者も一般ユーザーも同様に魅了し、好奇心と創造性をかき立ててきました。この実験は、AIの可能性を最大限に引き出すために必要なものです。しかし、用心する姿勢とガードレール（保護策）がなければ、セキュリティを損なったり、コンプライアンスから逸脱したりする可能性もあります。

バランスを保ち、AIイニシアチブをより効果的に管理するためには、次の3つの主要分野を考慮する必要があります。

1 AIの活用

サードパーティベンダーによって提供されるAI技術（例：ChatGPT、Bard、GitHub Copilot）を用いる一方、アセット（例：機密データ、知的財産、ソースコードなどを）を保護し、ならびにユースケースに基づいて潜在的なリスクを軽減する

2 AIの構築

組織固有のニーズに合わせたカスタムAIソリューションの開発（例：予測分析、顧客向けのコパイロットやチャットボット、AI駆動型の脅威検出システムのための独自アルゴリズムなど）

3 AIの安全確保

悪意のあるアクターが操作して予測不可能な動作を引き起こすことからのAIアプリケーションとAIシステムの保護

生成AIの実験を保護

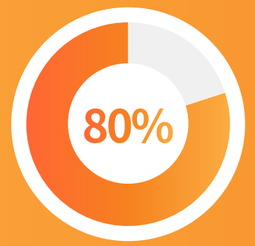
生成AIの変革：現状と今後

生成AIは消費者と企業にとって魅力的なものであり、これまでにない導入実績の軌跡をたどっています。活発なオープンソースコミュニティや、ChatGPTやStable Diffusionなどでの消費者主導の導入実験もあり、これにより少数のパワーユーザーグループが急速に成長することとなりました。

この中でユーザーが気づいたことは、実際にはボットが「人間を置き換える」ことはないということです。

生成AIは、ゼロからすべてを作るのではなく、人間を洗練させ、補強する立場にあり、企業が労働効率を増幅するのに役立つものとなるのです。予測AIでも、さまざまな取り組みの中で、意思決定の改善、よりスマートな製品の構築、顧客体験のパーソナライゼーションのためのデータ活用を容易にすることにより、同様のメリットを提供するものとなります。

現在、**開発者の59%**が開発ワークフローでAIを使用[1]



2026年までに、**企業の80%超**が本番環境にデプロイされた生成AI対応のAPI、モデル、アプリを使用（現在の5%から増加）[2]

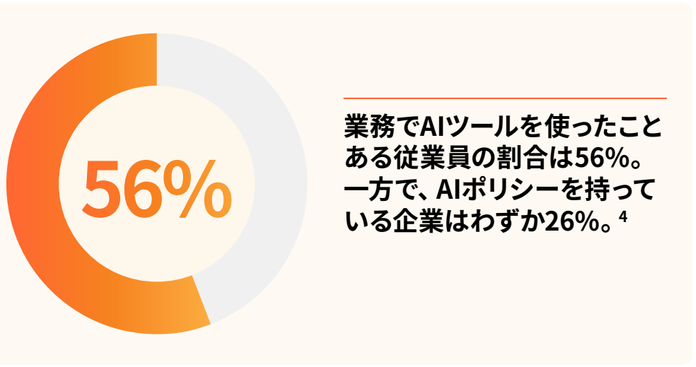
2030年までに、生成AIは**知的労働者のタスクの50%**を補強し、生産性または平均的な仕事の質を向上（現在の1%未満から増加）[3]

1. SlashData、“How developers interact with AI technologies”、2024年5月
2. Gartner、「A CTO's Guide to the Generative AI Technology Landscape」、2023年9月
3. Gartner、“Emerging Tech: The Key Technology Approaches That Define Generative AI”、2023年9月

生成AIを安全に利用する

AIの実験は、あらかじめ構築されたAIツールやサービスの使用からゼロからカスタムのAIソリューションの構築まで、多岐にわたります。独自のAIモデルやアプリケーションの作成に進む企業もある中、サードパーティのAIツールの利用にこだわろうとする企業も多いでしょう。

このような場合、組織はセキュリティとプライバシーの設定に関する直接制御が限定的であるため、サードパーティのAIツールによる新たなリスクが生じます。



業務でAIツールを使ったことある従業員の割合は56%。一方で、AIポリシーを持っている企業はわずか26%。[4]

現在、従業員はMicrosoft 365などのSaaSスイート、検索エンジンやパブリックアプリに組み込まれたチャットボット、さらにはAPIなど、既製のAIツールを業務用に使用しているものと思われます。

組織は、リスクを最小化するため、以下を例としたデューディリジェンスを行う必要があります。

- サードパーティツールのセキュリティリスク**の評価**
- データプライバシーの懸念へ**の対応**
- 外部APIへの依存または過剰依存**の管理**
- 潜在的な脆弱性**の監視**

従業員がChatGPTのようなパブリックWebアプリを使用する場合が例に挙げられます。プロンプトに送られるすべての入力は、組織の制御下からデータに変換されます。ユーザーは、個人を特定できる情報PII、財務データ、知的財産、ソースコードなど、機密情報、機密情報、または規制対象情報を過剰に共有する恐れがあります。また、明示的な機密情報を共有していない場合でも、インプットの文脈を組み合わせて機密データを推測できてしまいます。

安全策として、従業員は設定を切り替えることで自身の入力がモデルをそれ以上訓練するのを防ぐことができますが、これは手動で行う必要があります。セキュリティを確保するために、組織にはユーザーがプライベートデータを入力できないようにする方法が必要になります。

AIがセキュリティにもたらす影響に備える

データ流出

ユーザーがどの程度外部AIサービスと機密データを不適切に共有しているか匿名化/仮名化の技術は十分か

APIのリスク

攻撃者のゲートウェイになり得るサードパーティAPIの脆弱性にどのように対処すべきか

ブラックボックスシステム

予期せぬリスクをもたらす可能性がある外部AIモデルの意思決定プロセスは何か

ベンダーリスク管理

利用するサードパーティAIプロバイダーのセキュリティプラクティスについてどれくらい理解しているか、さらに重要な点として、知り得ていないことは何なのか

4. The conference Board、2023年9月

AI利用を防御するためのステップ

1 ガバナンスとリスクを管理する

- ユーザーが生成AIと共有できる情報、アクセス制御ガイドライン、コンプライアンス要件、違反の報告方法など、AIの使用方法と時期に関するポリシーを組織側が策定する
- AI利用の情報を収集し、特定し、定量化するためにインパクト評価を実施する

2 セキュリティとプライバシーの可視性と制御の強化

- AIアプリを含むすべての接続を記録し、ユーザーのアクティビティ、AIツールの使用状況、データアクセスパターンを継続的に監視し、異常を検出する
- シャドーIT（AIツールを含む）が何かを知り、承認、ブロック、追加の制御を決定する
- SaaSアプリの設定をスキャンして、潜在的なセキュリティリスク（例：承認されたアプリから未承認のAI対応アプリへのOAuth権限の付与、データの漏洩リスク）を見つける

3 AIツールに出入りするデータを調べ、IPを侵害したり、機密性に影響を与えたり、著作権制限を侵害したりする可能性のあるものをフィルターで除外

- ユーザーがAIツールとインタラクションする方法についてセキュリティ制御を適用（例：アップロードの阻止、コピー/ペーストの防止、機密データや専有データのスキャンと入力のブロックなど）
- WebサイトのスクレイピングなどのAIボットの挙動を防止する予防措置を講じる
- 他の制御方法がない場合だけ、AIツールを完全にブロックします。ご存じの通り、ユーザーは回避策を見つけ、セキュリティは制御不能に。

4 AIアプリとインフラストラクチャへのアクセスを管理

- AIツールにアクセスするすべてのユーザーとデバイスについて、AIツールを利用できるユーザーの範囲を拡大するために厳格な本人確認を確実に行う
- アイデンティティベースのZero Trustアクセス制御を導入。最小権限を適用し、安全性が損なわれたアカウントや内部脅威による潜在的被害を制限

5 コスト合理化と運用効率の向上

- レート制限、キャッシング、リクエストの再試行、使用の拡張に伴うモデルのフォールバックをコントロールできるようにするため、ユーザーがAIアプリケーションをどのように使っているかを分析とログで把握

お客様が構築するものを保護する

AIモデルの訓練

AIパイプラインは、脆弱性の範囲を広げています。しかし、開発の初期段階と開発プロセス全体で保護を経験してきたことから、何が成功につながるのかを見込んでいます。AIセキュリティの場合、取り組みを始めるべき点は、自然とそのモデルになります。

AIアプリケーションの前提として、AIモデルのトレーニングに使用されるものはすべて出力に反映されます。後で悪影響が出ないよう、最初にデータを保護する方法を検討するべきです。保護を怠れば、攻撃対象領域を拡大し、将来的にアプリケーションの問題を発生させる危険性があります。

データの完全性を保証するセキュリティは、意図的および偶発的なデータ漏えいを軽減する上で極めて重要になります。AIパイプラインにおけるセキュリティリスクには、以下のようなものがあります。

- **データポイズニング：**悪意のあるデータセットは成果に影響を与え、バイアスを生み出す
- **ハルシネーションの乱用：**脅威アクターは、AIのハルシネーション応答を生成するための情報の捏造を正当化し、悪意のある不正なデータセットがアウトプットに影響を与える

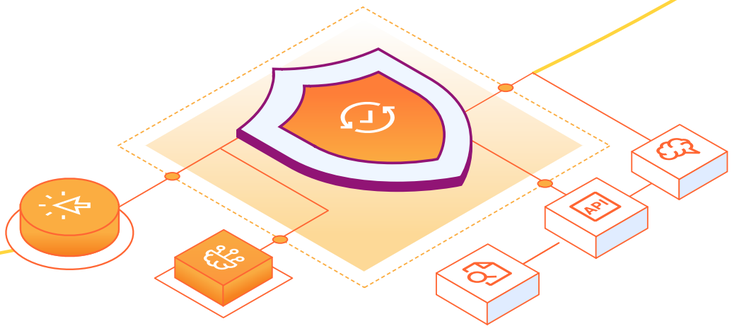
また、モデルをトレーニングしていない場合、社内のAIはタスクを実行するためのモデルを選択することから始まります。このような場合、推論の役割を果たすこのモデルを作成者がどのように作成し保護したのかを調べた方がよいでしょう。

推論は、AIの訓練に従うプロセスです。モデルをより良く訓練し、より微調整するほど、推論の精度も向上します。ただし、完全性は保証されません。高度に訓練されたモデルでもハルシネーションが起こることもあります。

デプロイ後のセキュリティ

社内AIを構築してデプロイしたら、そのプライベートデータを保護し、AIへの安全なアクセスを確保する必要があります。このホワイトペーパーですでに述べた各ユーザーへのトークンの強制やレート制限などの推奨事項に加え、次の点も考慮すべきでしょう。

- **クォータの管理制限：**を使用して、ユーザーのAPIキーが侵害/共有されるのを防ぐ
- **特定の自律システム番号（ASN）をブロック：**攻撃者が大量のトラフィックをアプリケーションに送信するのを阻止
- **待機室の提供やユーザーへの異議申し立て：**リクエストの困難または時間がかかり、攻撃者にとっては経済的損害につながる
- **APIスキーマの構築と検証：**すべてのAPIエンドポイントを識別してカタログ化することで、意図した使用方法を概説し、すべての特定のパラメータとタイプ制限をリストアップ
- **クエリの深さと複雑さを分析：**完全なDoS攻撃や開発者エラーから守り、オリジンを健全に保ち、期待通りにユーザーにリクエストを配信
- **トークンベースのアクセスに関する規則を整備：**ミドルウェア層またはAPI Gatewayでトークンを検証する際に、安全性が損なわれたアクセスから保護



生成AIの実験全体に堅牢な脅威保護を適用

導入から実装まで、生成AI実験の各段階は、最小限のリスクまたは許容されるリスクで進める必要があります。本書で得た知識があれば、お客様の組織が将来的に何らかの形でAIを使用、構築、計画、計画しているかどうかに関わらず、ご自分のデジタル環境をコントロールできます。

新しい機能を採用する際にためらいを感じるのは自然なことであるものの、自信を持って安全にAIを実験できるようにするためのリソースは確実に存在します。これらのリソースのうち、企業が今日最も必要としているのは、すべてのITとセキュリティのための結合組織と言えます。環境内のあらゆるものと連携することで複雑さを軽減する共通のスレッドとして機能し、どこでも利用可能で、必要なセキュリティ、ネットワーキング、開発機能を実行できるものがあるのです。

つなぎ合わせる細胞を使えば、以下のようなさまざまなユースケースに自信を持てるようになります。

- 規制対象データの移動の検出・制御機能により規制を遵守
- SaaSアプリ、シャドーIT、新AIツールにおける機密データの可視性と制御を取り戻す
- アップロードやダウンロードの際にソースコードを検出、ブロックすることで、開発者コードを保護。さらに、コードリポジトリを含むSaaSアプリケーションやクラウドサービスでの設定ミスの防止、発見、修正も可能

AIが進化し続けるにつれ、不確実性が生まれることは確実です。だからこそ、Cloudflareのような安定化を支える力を持つことは非常に有益なのです。

3種類のLLMにおけるAIリスクからの保護

使用によって、AIが組織にもたらすリスク露出のレベルは異なります。大規模言語モデルLLMの利用と開発に関連するさまざまなリスクを理解した上で、LLMのデプロイメントに積極的に関与することが重要です。

LLMのタイプ	主なリスク
内部	機密データおよび知的財産へのアクセス
製品	風評被害
パブリック	機密データの漏洩

スケール、使いやすさ、シームレスな統合

Cloudflareのコネクティビティクラウドにより、コントロールが可能になり、可視性とセキュリティが向上します。それにより、AIの実験を安全かつスケーラブルなものになります。さらに、弊社のサービスはすべてを強化し、ユーザーエクスペリエンスとセキュリティの間での トレードオフとは無縁です。

ほとんどの企業が、AIのみを利用するだけか利用した上で構築するかのいずれかであることを考えると、Cloudflareを活用することは、AIプロジェクトが決して行き詰まることがないことを意味します。

- 弊社の**グローバルネットワーク**により、必要に応じて、迅速に制御を拡張して適用可能
- その**使いやすさにより**、ユーザーがAIを利用する方法に関するポリシーを簡単に導入、管理できる
- 単一の**プログラム可能なアーキテクチャ**により、ユーザーのAI活用方法を阻害することなく、構築するアプリケーションにセキュリティを多層に適用

Cloudflareのコネクティビティクラウドは、AI実験のあらゆる面で次の点を保護します。

- 弊社の**Zero Trustおよびセキュアアクセスサービスエッジ（SASE）**サービスは、従業員によるサードパーティAIツールの**使用**方法におけるリスクを軽減
- 弊社の**開発者向けプラットフォーム**が、お客様の組織独自のAIツール・モデルの安全かつ効率的な**構築**を支援
- **AIを活用した保護**の場合、弊社のプラットフォームでは、AIと機械学習の技術を活用して脅威インテリジェンスを構築し、AI実験を通じて組織を保護するために使用

	AIの活用方法	AIの構築方法
弊社のグローバルネットワークはスケール可能	あらゆる場所に一貫性を持って制御を適用	推論、クエリ、キャッシングを高速化
シンプルな管理	シンプルなデプロイメントとポリシーを備えた1つのコントロールプレーン	すばやくオンボードするためのテンプレート
統合され、プログラム可能なネットワークアーキテクチャ	AIの活用方法を中断することなく、新たなセキュリティを重ねる	プライバシーとコンプライアンスを内蔵

次のステップ

組織のAI使用方法の保護から構築するAIアプリケーションの防御まで、AI向けCloudflareがお客様をカバーします。弊社のサービスを利用すれば、無限の相互運用性と柔軟な統合により、どのような順序でも新しい機能を導入できます。

専門家に相談

詳細については、cloudflare.com/ja-jpをご覧ください。

CLOUDFLARE

enterprise@cloudflare.com | **cloudflare.com/ja-jp**

REV:BDES-6307.2024AUG19