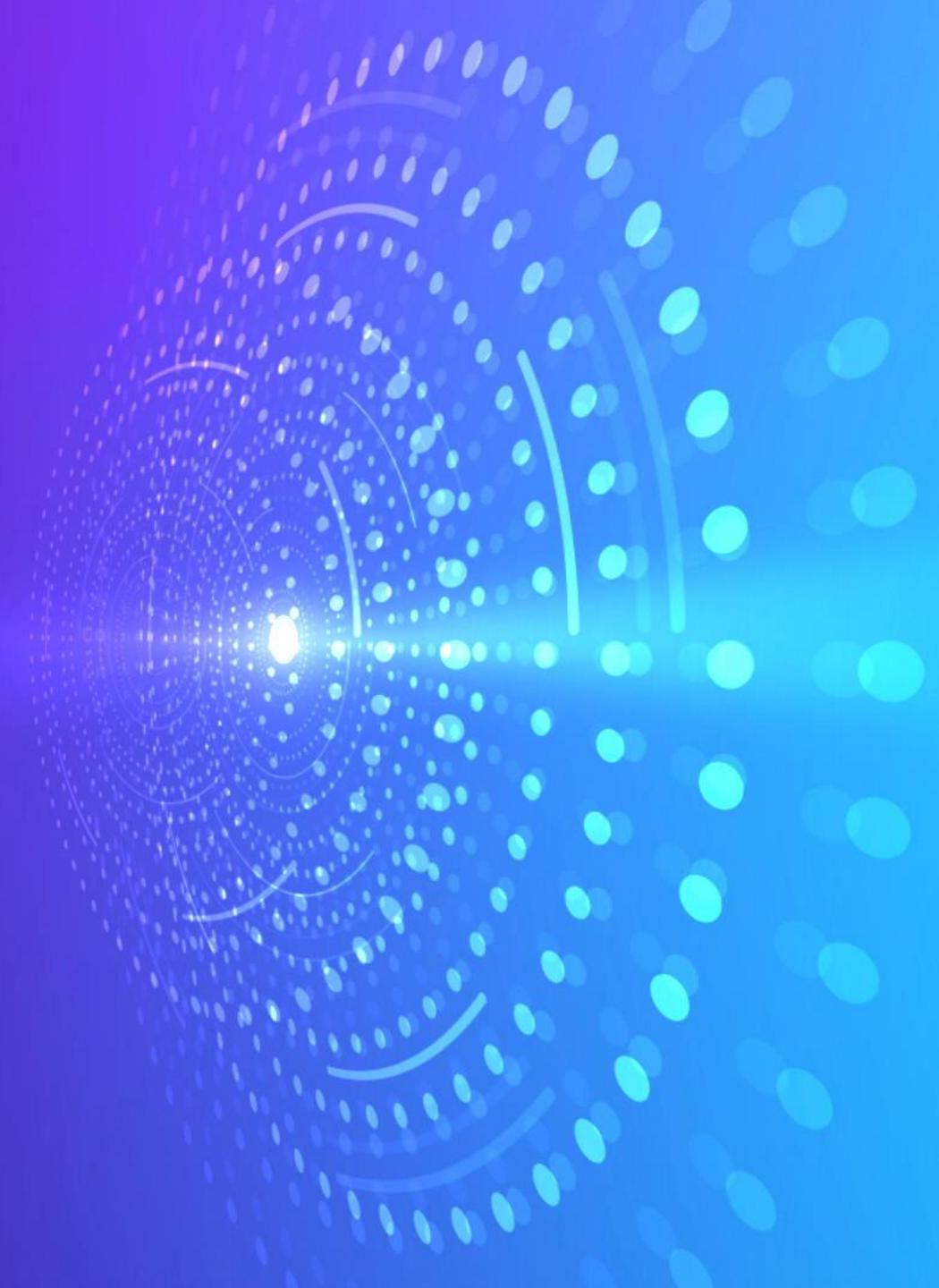




AI Day

Tokyo

～AWSのテクノロジーで加速する
生成AIのプロダクション活用～





AI-T1-03

イノベーション加速のための 実践！“責任ある AI”

吉田 朗

Amazon ウェブ サービス ジャパン合同会社
公共政策部 シニアマネージャー

保里 善太

Amazon ウェブ サービス ジャパン合同会社
プロフェッショナルサービス本部
シニアセキュリティコンサルタント

吉田 朗

アマゾン ウェブ サービス ジャパン合同会社
公共政策部 シニアマネージャー



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

「責任あるAI」の課題と実践



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI の利活用のバランス



—



—



AIの安全性確保が求められています

優先課題

89%

89%の経営幹部は、サイバーセキュリティ(対AI、クラウド等)を優先課題としています。(BCG)

懸念

94%

94%の経営幹部が、AIの導入前に確実にAIをセキュアにすることが重要と述べています。(IBM)

意図せぬ結果

65%

65%のC-level層は、生成AIの使用による意図せぬ結果に懸念を持っています。(EY)

「責任あるAI」に関する国際的な動向



- 2023年、AIに関する2,175の法案が49か国で審議（2022年から倍増）
- 3,500以上のAI関連のインシデントが AI incident database に報告

AI制度や標準化に関する議論への貢献

Amazon はホワイトハウス、テクノロジー組織、AI コミュニティとの議論に参画し、
責任ある安全な AI 利用を推進



さらに詳しく

将来の生成 AI モデルの開発に向けた 新たな自発的取り組み

- ✓ 社内外からの攻撃検証 (red-teaming)
- ✓ セキュリティリスク情報
- ✓ 音声・画像が AI によって生成されたものかどうかを判断するメカニズム
- ✓ サイバーセキュリティと内部脅威対策
- ✓ 第三者による問題の発見と報告
- ✓ モデルの機能、制限事項、適切な使用領域
- ✓ AI がもたらす社会的リスクに関する研究
- ✓ 社会の課題に対処する AI システム

責任ある AI を AI ライフサイクル全体に実践



リスクベースアプローチの重要性



音楽配信サービスでの
おすすめ機能



レントゲン検査での
腫瘍の発見

それぞれのAI活用事例でのリスク対応の違い？

AWSによるAIライフサイクルを通じた支援



Amazon Partner Network & AWS Solution Architects

AWS Generative AI Innovation Center

AWS Audit Manager, AWS Artifact, AWS Config

Amazon SageMaker Data Wrangler & Ground Truth & Clarify

Amazon SageMaker ML Governance (e.g., Model Registry)

Amazon SageMaker Model Monitor

Bedrock Guardrails & Amazon CloudWatch

AWS AI サービスカードによる透明性の向上

責任ある AI を推進するための透明性リソース

- AWS AI サービスの使用目的と公平性に関する考慮事項を文書化
- お客様向けの責任ある AI のためのドキュメント
- 責任ある AI 設計の選択
- AWS の AI サービスで想定されるユースケースや制限を記載
- 最善のデプロイと運用プラクティスに関する情報
- 10の AI ベースのサービスで利用可能、今後さらに追加予定

Amazon Titan Text

Amazon Comprehend Detect PII

Amazon Transcribe Toxicity Detection

AWS HealthScribe

Amazon Rekognition Face Liveness

Amazon Rekognition Face Matching

Amazon Textract AnalyzeID

Amazon Transcribe – Batch

Amazon Transcribe – Streaming

<https://aws.amazon.com/machine-learning/responsible-ai/resources/>

保里 善太

アマゾン ウェブ サービス ジャパン合同会社
プロフェッショナルサービス本部
シニアセキュリティコンサルタント



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

責任ある AI 実践に向けたAWSのご支援



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

プロフェッショナルサービスとは



AWSプロフェッショナルサービスとは

お客様がクラウドのイノベーティブな活用により**ビジネス価値を生み出すことを支援し、加速させるための有償コンサルティングサービス**です。

特徴	ビジネス価値の例
<ul style="list-style-type: none">お客様のクラウド導入～利活用を支援または加速させるための有償コンサルティングサービスとしてご提供エンタープライズ、政府機関、それらのお客様に従事する SI / ISV 様にご提供AWSの技術領域に高度に特化期間は3か月～1年程度のプロジェクトベースでご支援タイムアンドマテリアル型（従量課金）	<ul style="list-style-type: none">✓ 俊敏性と生産性の向上✓ データセンターの統合・廃止✓ 迅速な世界進出✓ セキュリティ向上と事業復元力✓ デジタル変革（DX）の推進✓ IoT, AI/ML データ活用✓ 新しいビジネスモデル✓ 人材育成✓ コスト削減

**お客様のクラウドジャーニーの歩みをご支援する
Professional集団**

<http://aws.amazon.com/jp/professional-services/>

セキュリティ支援サービスの位置付け

セキュリティ支援サービスではお客様がセキュリティ上の責任を持つ範囲において、AWSのサービスや機能を効果的・効率的に利用し、セキュリティリスクを低減することを目的としています。

お客様のセキュリティ対応範囲

利用者自らが対応方法を
調査・策定、実装する範囲

AWSのサービス/機能を
活用して対応する範囲

セキュリティ支援サービスの
フォーカスポイント



AIとセキュリティの取り組みに関するAWSトップの声

Matt Garman, AWS CEO

私たちは企業に構築するための安全なプラットフォームを提供するにはどうすればよいのか考えました。なぜなら、これらすべての新機能があったとしても、セキュリティは依然として何よりも重要であり、お客様が心配していることだからです。

Andy Jassy, CEO

ところで、GenAI のセキュリティの重要性を過小評価しないでください。お客様のAIモデルには、最も機密性の高いデータがいくつか含まれています。AWS とそのパートナーは、世界で最も強力なセキュリティ機能とトラックを提供しており、その結果、GenAI を AWS で実行したいと考えるお客様が増えています。



CNBC TelevisionのYoutubeチャンネルより
「AWS CEO Matt Garman breaks down the company's AI strategy」から引用



2023 Letter to Shareholders



CEO Andy Jassy's 2023 Letter to Shareholders
より引用



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

生成 AI セキュリティとは

生成 AI セキュリティ

生成 AI 脆弱性への対策

プロンプトインジェクション

安全が確認されていない出力ハンドリング

訓練データの汚染

モデルのDoS

サプライチェーンの脆弱性

機密情報の漏えい

安全が確認されていないプラグイン設計

過剰な代理行為

過度の信頼

モデルの盗難

責任ある AI

制御性

プライバシーとセキュリティ

安全性

公平性

信憑性と堅牢性

説明可能性

透明性

ガバナンス

(例) OWASP Top 10 for LLM Applications

(例) 責任ある AI のコアディメンション

責任あるAI

AWSではAIの安全で責任ある開発を促進するため、責任あるAIの8つのコアディメンションを定めています。開発者は、利用者にとって安全な生成AIアプリケーションを構築して提供する必要があると考えます。

公平さ

さまざまなステークホルダーのグループへの影響を考慮する

説明可能性

システム出力を理解して評価する

プライバシーとセキュリティ

データとモデルを適切に取得、使用、保護する

安全性

有害なシステム出力と誤用を防ぐ

制御性

AIシステムの動作をモニタリングおよび制御するメカニズムを備える

正確性と堅牢性

予期しない入力や敵対的な入力があっても、正しいシステム出力を実現する

ガバナンス

ステークホルダーがAIシステムとの関わりについて十分な情報に基づいた選択を行えるようにする

透明性

プロバイダーとデプロイヤーを含むAIサプライチェーンにベストプラクティスを組み込む



生成AIセキュリティ検討に活用できるガイダンス (AWS)

生成AIは自然言語による多様なインプットを受け付け、様々なテキストや画像などを生成します。このため安全な生成AIアプリケーションを構築するために、生成AI特有の考慮事項を理解する必要があります。AWSでは一例として、次のようなガイダンスを提供しています。

生成AIセキュリティスコーピングマトリックス

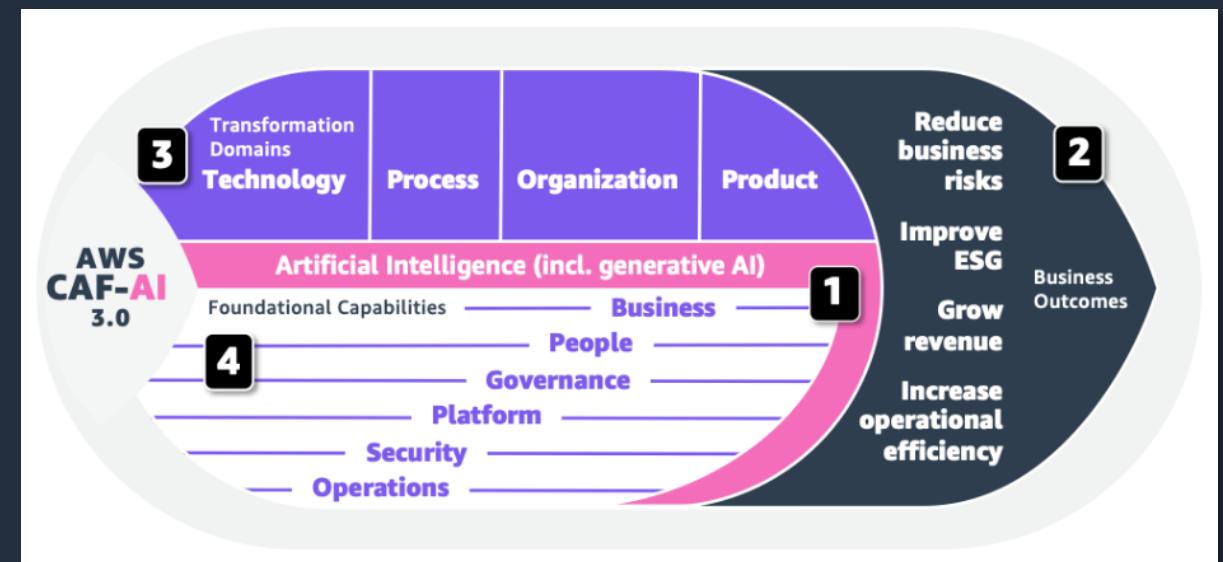
生成 AI セキュリティスコーピングマトリックス

ユースケースを分類するメンタルモデル



生成AIをセキュアにする: 生成AIセキュリティスコーピングマトリックスの紹介

AWS CAF-AIのセキュリティのパースペクティブ



クラウド導入フレームワークのセキュリティのパースペクティブ



生成AIセキュリティ検討に活用できるガイダンス（AWS以外）

またAWS以外にも、AI特有のセキュリティ、リスクを検討する際に活用できるフレームワークやガイダンスが主要な団体からリリースされています。

提供元	ガイドライン/フレームワーク	概要	分類
NIST	Artificial Intelligence Risk Management Framework (AI RMF 1.0)	AIに特化したリスク管理フレームワークであり、各管理策に対する Playbook なども整理されている	責任ある AI
ISO	ISO/IEC 42001:2023	AIに関するリスクを回避するための要件やリスクが生じた場合の対応を含む信頼性の高いマネジメントシステムを構築するために活用が想定される	責任ある AI
OWASP	OWASP Top10 for LLM Applications	LLM アプリケーションに特化した10の脆弱性に対して具体的な脅威と緩和戦略が記載されている	AI 脆弱性
MITRE	MITRE ATLAS	MITRE ATT&CK の観点をベースに生成 AI アプリケーションへの攻撃の戦術や技術または手法の観点で脅威を分類するフレームワーク	AI 脆弱性



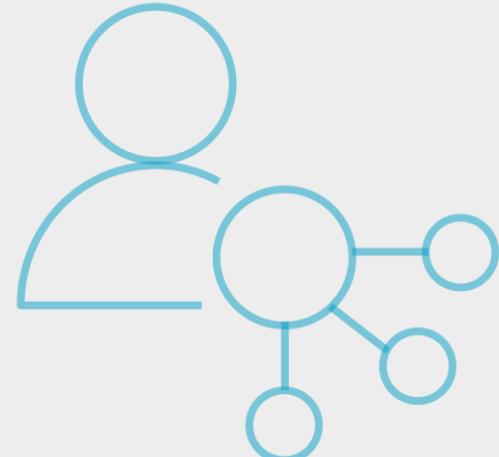
AWSでのAIシステムの保護の例

AWSでは、人、プロセス、テクノロジーの3つの観点でAIシステムを保護しています



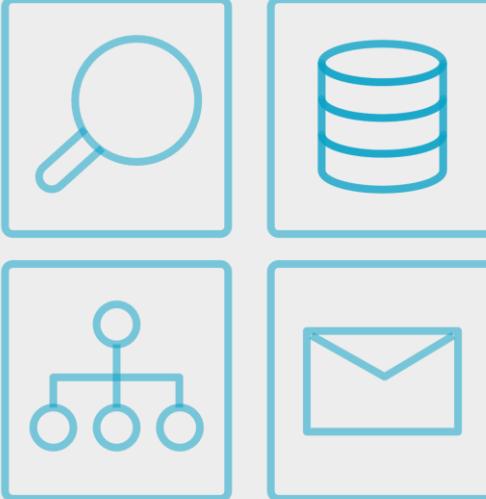
人

- トレーニングと啓発
- アクセス制御
- セキュリティ文化



プロセス

- リスク管理
- インシデント対応
- コンプライアンス



テクノロジー

- データ暗号化
- ネットワークセキュリティ
- 脅威検知

生成AIセキュリティに関する ご支援内容の紹介



想定される課題とご支援内容の概要

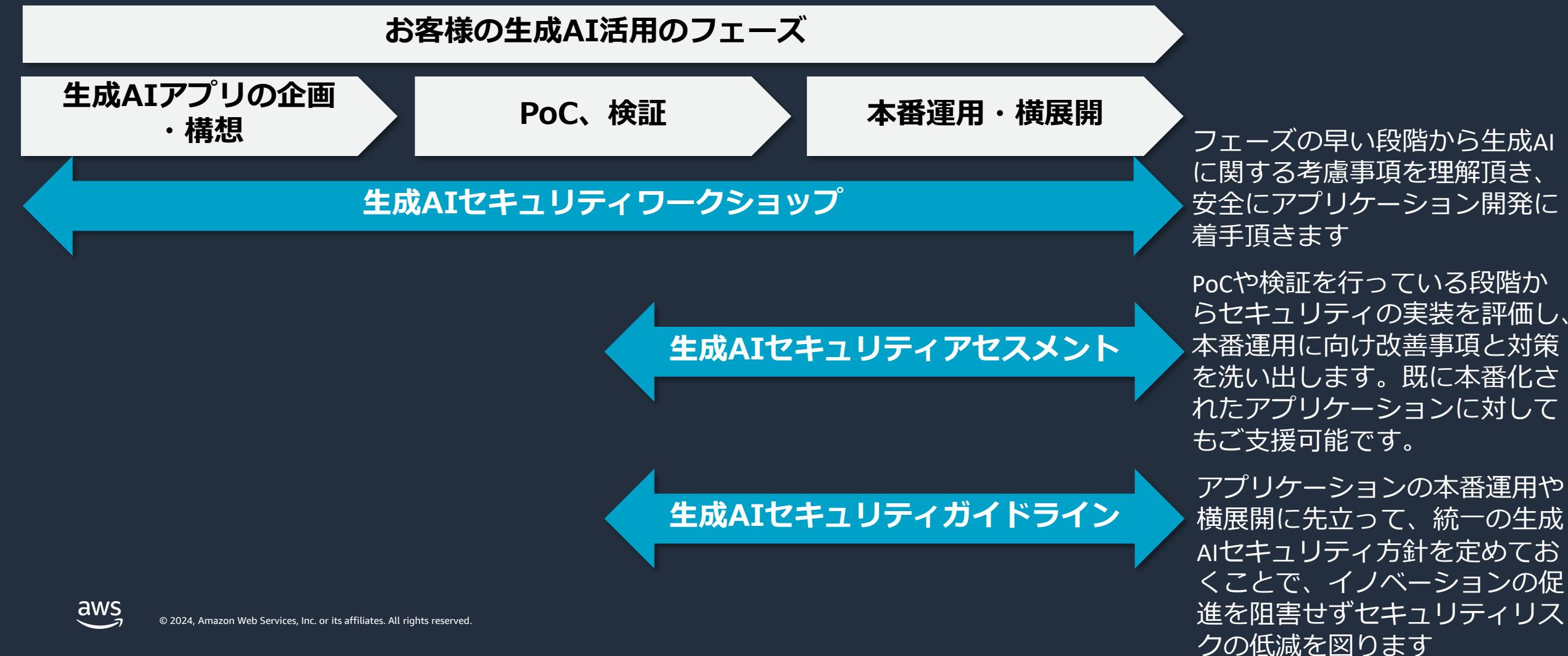
Professional Service セキュリティコンサルタントは、「人」「テクノロジー」「プロセス」に対して次のようなご支援を通して、お客様の安全な生成AIアプリケーションの実装をご支援します。

想定される課題	プロフェッショナルサービスによるご支援	お客様にとっての価値
生成AIセキュリティについて何から考えてよいかわからない	人に対するアプローチ	<ul style="list-style-type: none">生成AIセキュリティのスキル向上
生成AIセキュリティについてスキルアップが必要である	AWS生成AIセキュリティワークショップ	<ul style="list-style-type: none">開発者のスキル向上による安全な生成AIアプリケーションの提供
現在の本番/PoC環境が安全に構築されているかわからない	テクノロジーに対するアプローチ	<ul style="list-style-type: none">評価対象の生成AIアプリケーションおよびAWSアカウントの現状の評価と改善事項の抽出
従来のツールが対応しておらず生成AI特有の観点が考慮できない	AWS生成AIセキュリティアセスメント	<ul style="list-style-type: none">今後のアクションと優先順位の明確化
生成AIを利用する際のルールが定まっていない	プロセスに対するアプローチ	<ul style="list-style-type: none">生成AIアプリケーションのセキュリティ実装方針の明確化
各部門のPoCが先行しガバナンスが利かせられていない	AWS生成AIセキュリティガイドライン	<ul style="list-style-type: none">セキュリティ対策の検討期間短縮生成AIアプリケーションのガバナンス強化によるリスクの低減



お客様の生成AI活用のフェーズと対応するご支援の例

前述の3つのご支援メニューは、お客様の生成AI活用の各フェーズをご支援します。



責任あるAI観点のセキュリティアセスメントの例

カテゴリ	質問内容
AI/ML ワークロードが有効であることを確認する (NIST AI 100-1 3.1)	AI システムの結果が正確で信頼性が高く、トレーニング データ セットを超えたデータに対して十分に一般化されていることを、モニタリングとテストを通じて検証できますか？
AI/ML ワークロードが安全であることを確認する (NIST AI 100-1 3.2)	AI システムが人間の生命、健康、財産、環境が危険にさらされる状態にならないことをどのようにして確保しましたか？
AI/ML ワークロードが公平であり、有害なバイアスが管理されていることを確認する (NIST AI 100-1 3.7)	AI システムは、システムと対話するすべてのユーザー、またはシステムによって影響を受けるすべてのユーザーに公平かつ公平な結果を促す方法でトレーニングおよび設計されていますか？AI システムの設計に導入される有害なバイアスを軽減するための措置は講じられていますか？

Amazon Bedrock Guardrails

アプリケーションの要件と責任ある AI ポリシーに合わせてカスタマイズされた保護手段を実装



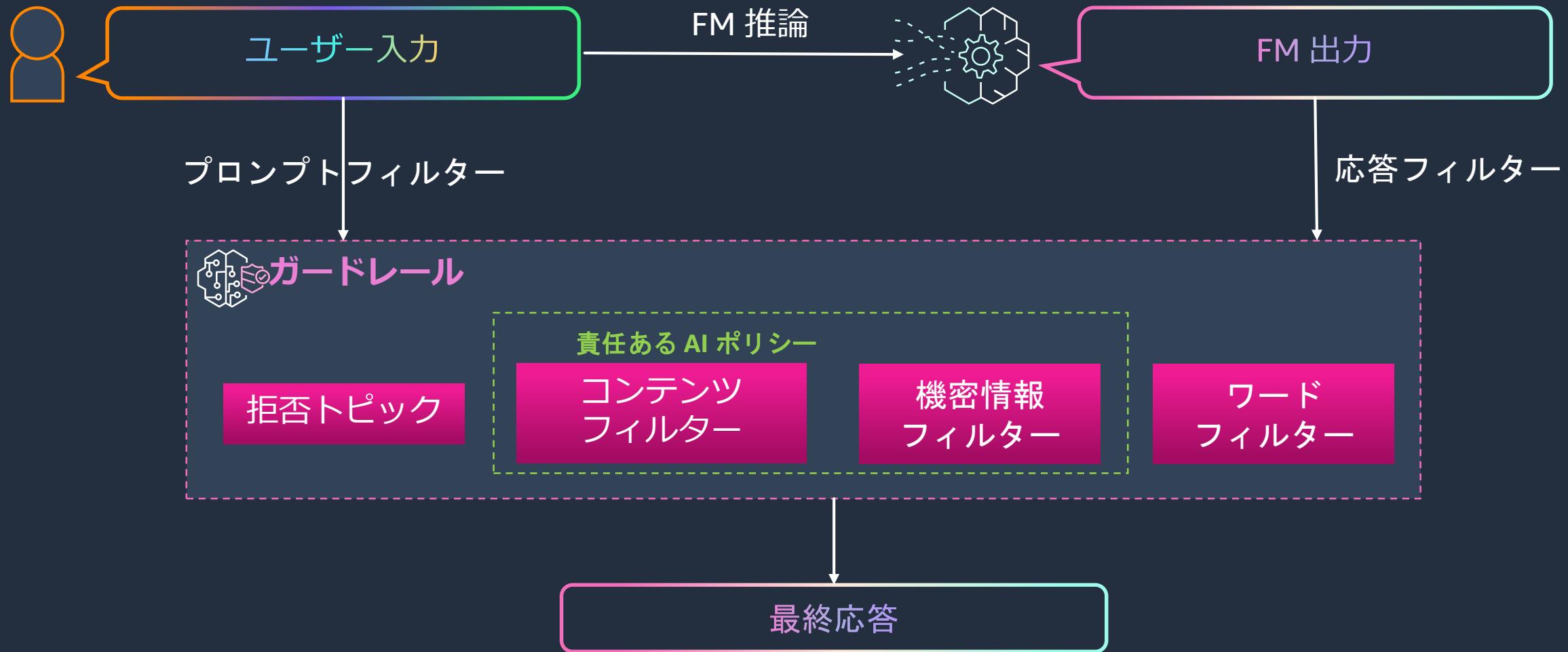
Amazon Bedrock の複数の基盤モデルとエージェントにガードレールを適用

責任ある AI ポリシーに基づいて有害コンテンツのフィルタリングを設定

拒否するトピックを短い自然言語による記述で定義して禁止

基盤モデルレスポンス内の機密情報をマスク/ブロック

Amazon Bedrock Guardrailsの仕組み



ガードレールの違反検出結果

The screenshot shows the AWS Guardrail trace interface. On the left, under 'Test' (a), the 'Prompt' section contains the question "Can I get guaranteed returns if I invest in stocks?". A red box highlights this prompt. Below it, the 'Model response' section displays a Japanese translation: "株式に投資したら確実にリターンを得ることはできますか？". To the right, the 'Guardrail trace' panel shows a table of results:

Category	Test result	Details
Denied topics	☒ Failure	Detected 'Investment Advice' topic
Content filters	☑ Pass	—

A large orange callout box points to the 'Denied topics' row, stating: "拒否トピックに検知されて回答がブロックされていることがわかる。". At the bottom, the 'Final response' section shows the message "Sorry, your query violates our usage policies.", with a red box highlighting it. A yellow callout box points to this message, stating: "すみません、この質問は利用ポリシー違反です。".



AI-T1-03

イノベーション加速のための 実践！“責任ある AI”

吉田 朗

Amazon ウェブ サービス ジャパン合同会社
公共政策部 シニアマネージャー

保里 善太

Amazon ウェブ サービス ジャパン合同会社
プロフェッショナルサービス本部
シニアセキュリティコンサルタント

実装において責任あるAI、
セキュリティについて
お客様はどのような対応をしているか

どのようなことに
お客様が苦労されているか

責任あるAIの対応の秘訣



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

責任あるAIについて、
日本政府はどういうところに関心があるか

海外でもAIに対する規制の議論が
盛んになっているとのことだが、
日本において注目すべきことは何か



Thank you!