



AIエコノミーと 責任あるAI

日本マイクロソフト株式会社
政策渉外・法務本部 政策渉外ディレクター
小島治樹



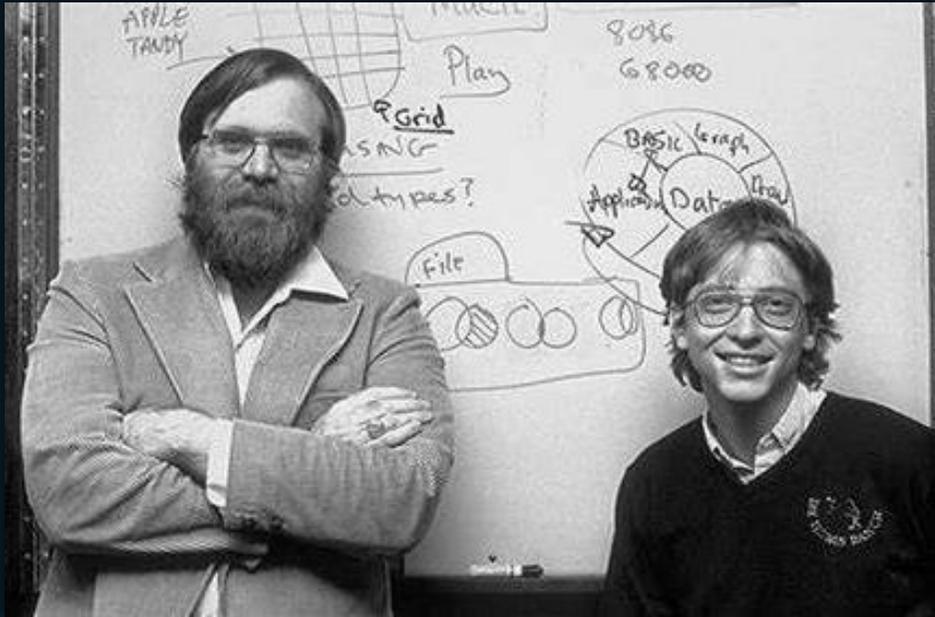
Microsoft

50 年

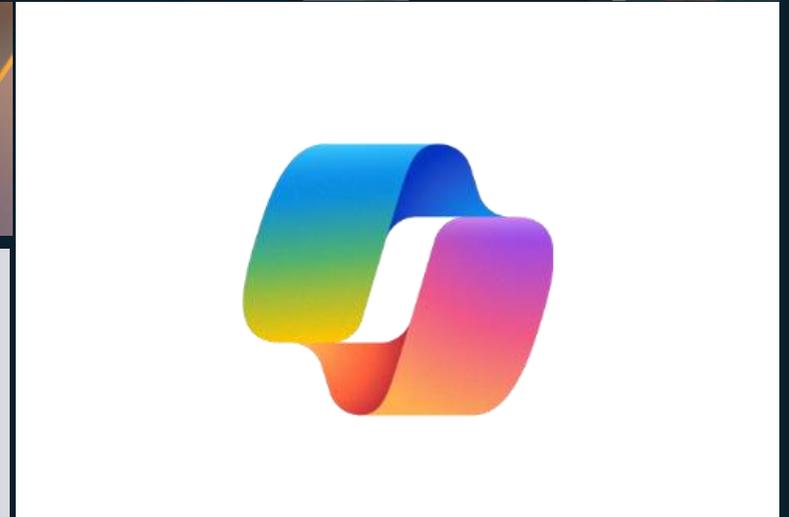
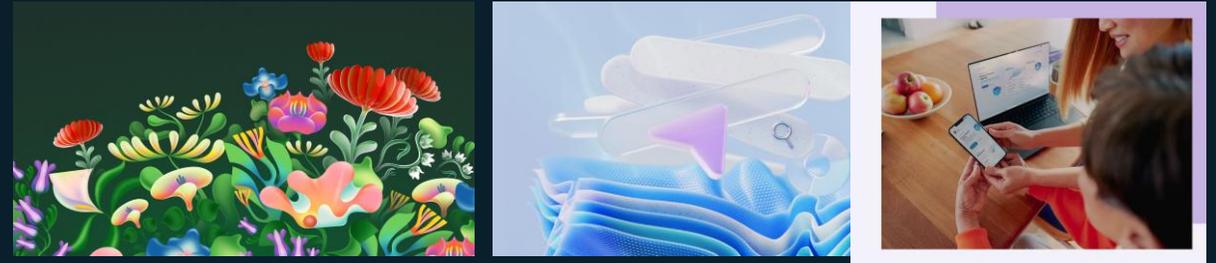
AIの経済効果

34 兆円

From:



To:



2種類のGPT



Generative Pretrained Transformer



General Purpose Technology

電気の「テックスタック」

メーカーとユーザー

アプライアンス

配線、スイッチ、ソケット

変圧器およびサーキットブレーカー

配電用電力網

電力ストレージ

発電

燃料

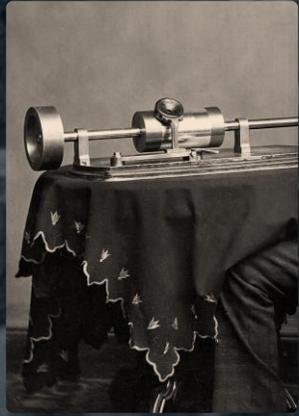
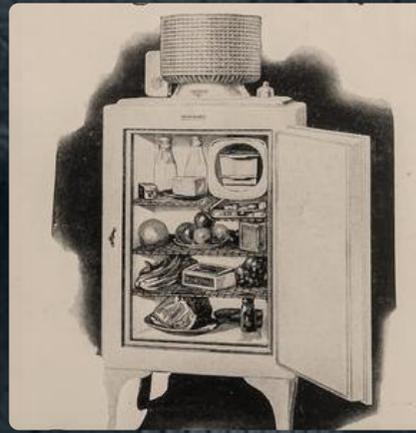
資本集約的なインフラストラクチャ



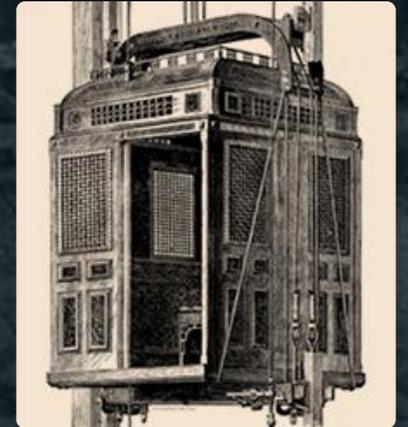
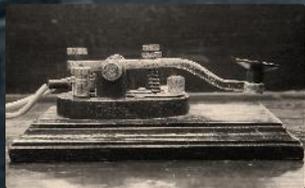
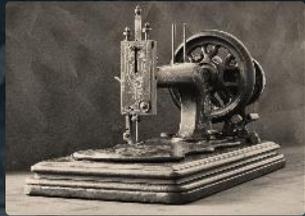
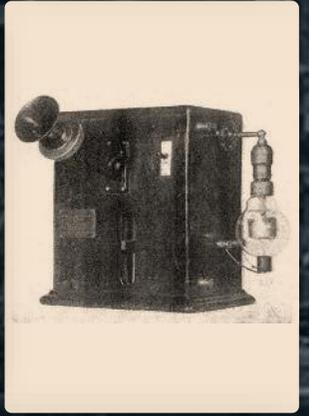
発電所



電力網



電化製品は、人々の生活を楽にするために
電気を動かします



The background features a complex digital network of glowing orange and blue lines and nodes, with vertical columns of binary code (0s and 1s) on the left side. A central dark purple rounded rectangle contains the text.

AI エコノミー

AIのテックスタック

ユーザー

流通

アプリケーション

ツーリング

基盤モデル

データ

AIデータセンター(インフラストラクチャ)

半導体チップ

電力+ コネクティビティ

A photograph of a server room with blue racks and yellow cables. The racks are filled with server equipment, and the cables are bundled and organized. The lighting is blue, creating a high-tech atmosphere.

新しいAIインフラストラクチャへの
資本集約的な投資の需要増加

マイクロソフトの投資アナウンスメント (2024年4月)



日本の AI 及びクラウド基盤の強化に
約4,400億円 (\$2.9bn)を投資

300 万人のリスキングを通じた
AI 活用の底上げ

マイクロソフトリサーチアジアの
日本初拠点を東京に開設

サイバーセキュリティ強化で
政府と連携

急速に増加しているAIモデル

TOYOTA : Multi-Agent System “O-Beya (大部屋)”



AIを活用した
職人技の継承



AIガバナンス

米国
先進AIチップの輸出規制

AI権利章典の青写真

NIST AIリスク管理フレームワーク

OMBガイドライン

カナダ
行動規範

英国
自主的なガードレール、セクター別ガバナンス

中国
レコメンダー制度の規制
ディープシンセシスレギュレーション
生成AIに関する中間措置
顔認識ルールの提案

日本
AI事業者ガイドライン
法律を検討中

韓国
AI法

EU
AI法、Code of Practice

インド
AIガードレールとイノベーション推進のフレームワーク

ASEAN
批准されたガイドライン

ブラジル
AI法案

チリ
AI法案

シンガポール
AI verify
安全性レッドチーム

オーストラリア
AI法
必須ガードレール案

国際的な取り組み

OECD AI原則 2019年5月 | **G20** AI原則 2019年6月 | **UNESCO** AIの倫理に関する提言 2021年11月 | **G7** AI Guiding Principles & CoC for Developers 2023年10月、モニタリングCoCのレポートフレームワーク 2024年6月 | **AI Safety Summits** – ブレッチリー宣言 2023年11月、ソウル宣言 2024年5月 | **Council of Europe** AIに関する条約 2024年3月 | **UN General Assembly** AI決議:2024年3月(Safe, Secure, Trustworthy AI)、2024年7月(AIアクセスと包括的な進歩)

AIガバナンスレイヤー

国際政策：国連

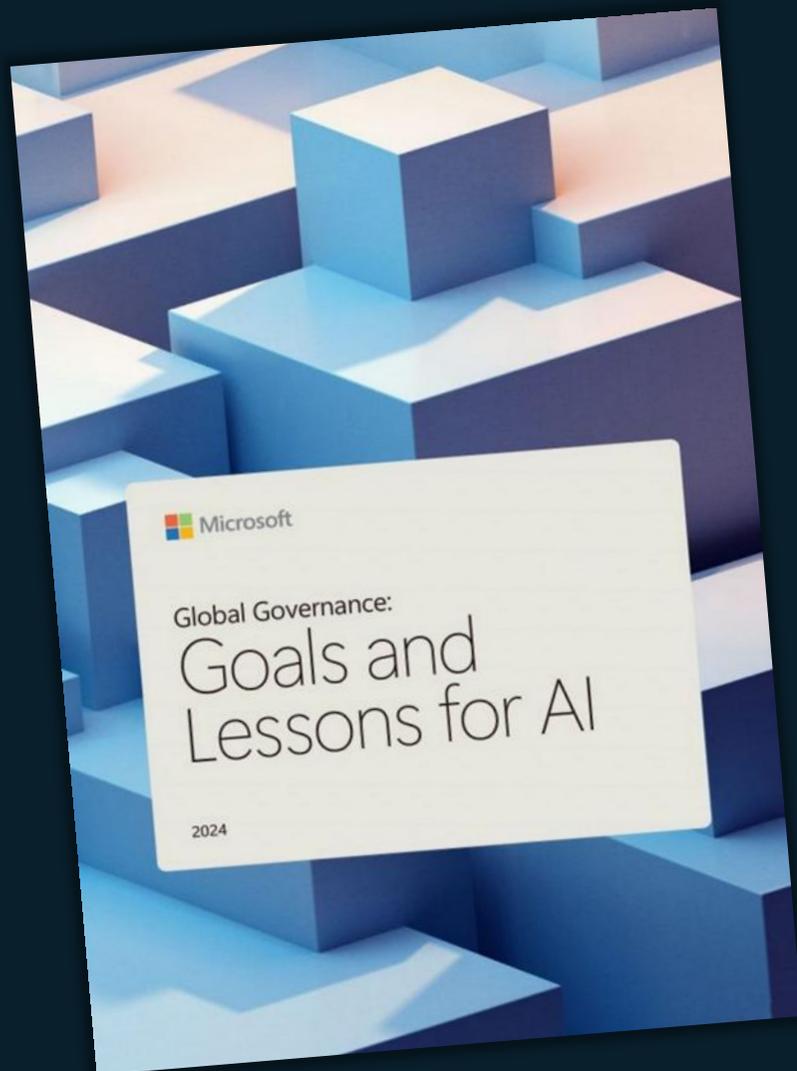
多国間政策：G7、G20、OECD

二国間政策

国内政策

業界標準

社内ポリシー



専門家のケーススタディ:

ICAO

The International Civil Aviation Organization

CERN

The European Organization for Nuclear Research

IAEA

The International Atomic Energy Agency

IPCC

The Intergovernmental Panel on Climate Change

FSB

The Financial Stability Board

FATF

The Financial Action Task Force

ドメイン間の比較分析を実施

グローバル ガバナンス: AI に関する目標と教訓

成果



広範で包括的なアクセス



世界的に重要な
リスク ガバナンス



規制の
相互運用性

機能

リソースとスキルへの
アクセスの強化

科学的コンセンサス
の構築

スタンダードの
設定と実装

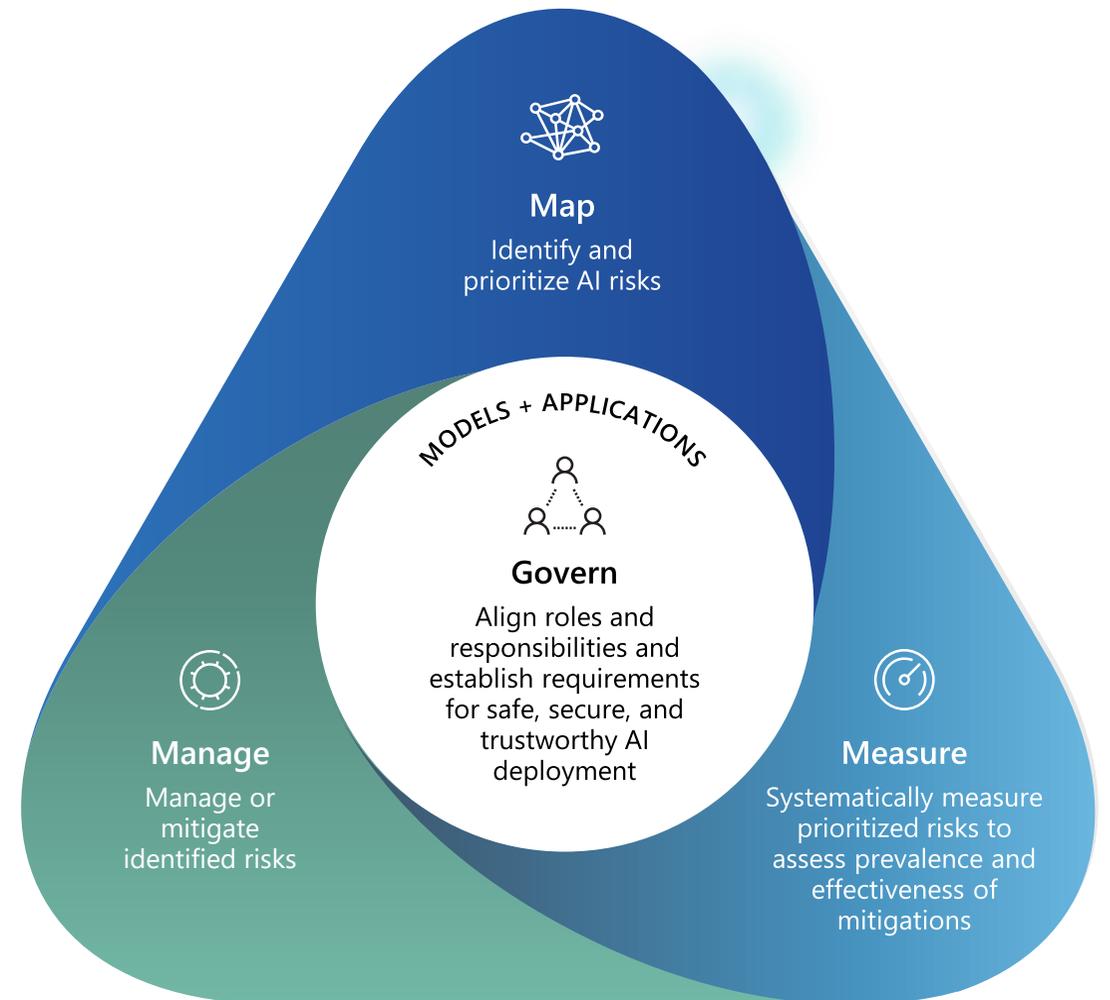
重要なリスクの
監視と管理

AI Action Summit後のAI Action

- ・ 透明性の高い報告のための国境を越えたコラボレーションの促進
- ・ AIの影響をより深く理解するための科学的研究の強化
- ・ アクセシビリティのためのオープンソースAIツールの開発を推進
- ・ 広島AIプロセスフレンズグループを通じたインクルーシブな進歩

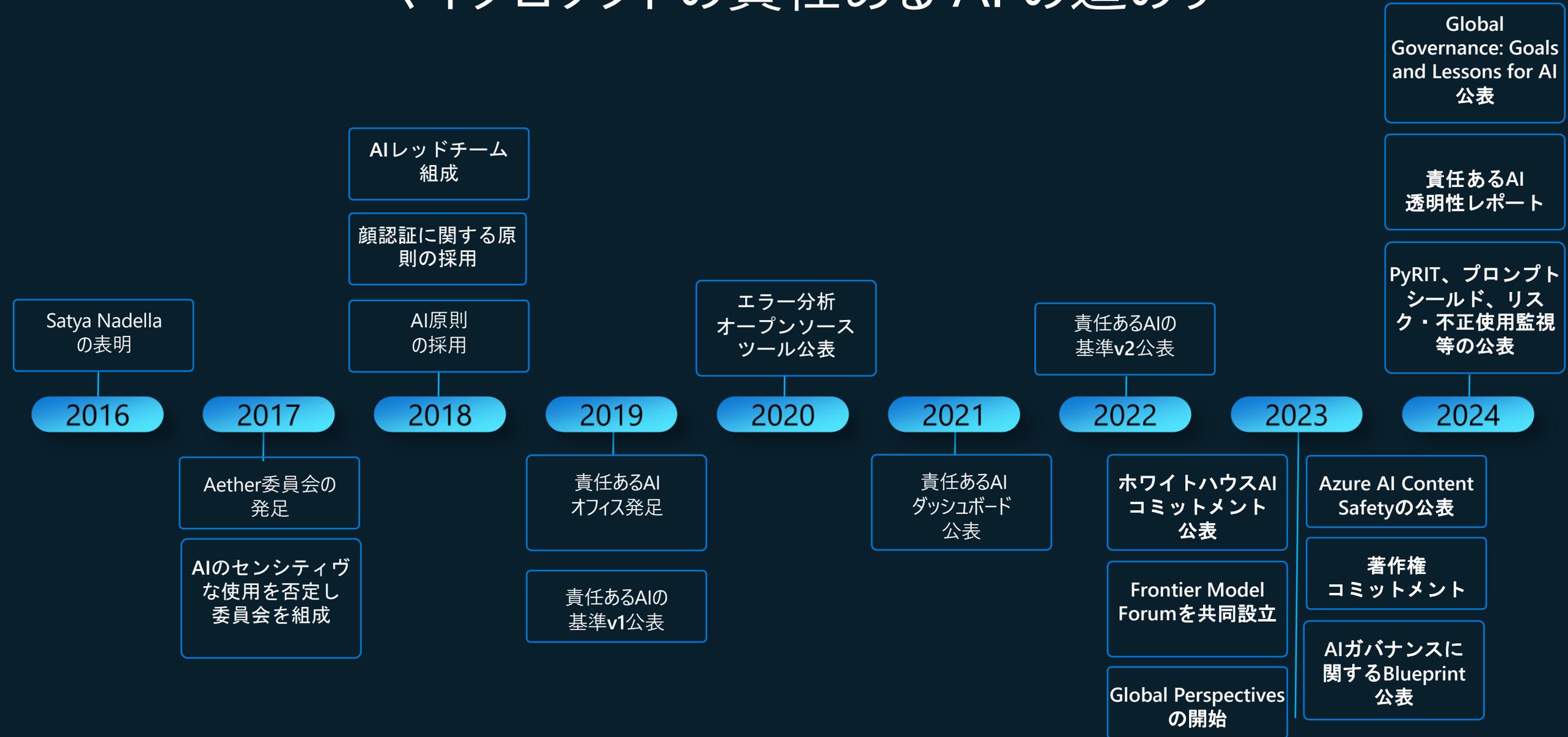


責任をもって生成AI を構築するフレー ムワーク



NIST AI Risk Management Framework

マイクロソフトの責任ある AI の道のり



責任ある AI に向けたマイクロソフトのアプローチ



原則

公平性
信頼性と安全性

プライバシーとセキュリティ
包括性

透明性
説明責任

企業 基準

目標
要件
プラクティス

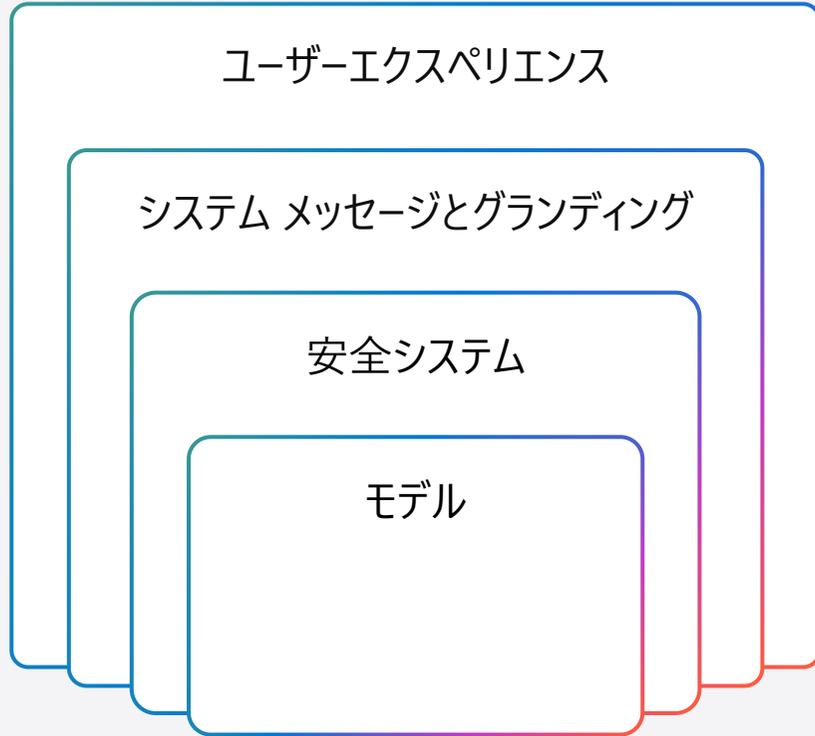
導入

プロセス
トレーニング
ツール

監督

監視
レポート
監査

リスク軽減策



責任ある人間と AI の対話のための設計

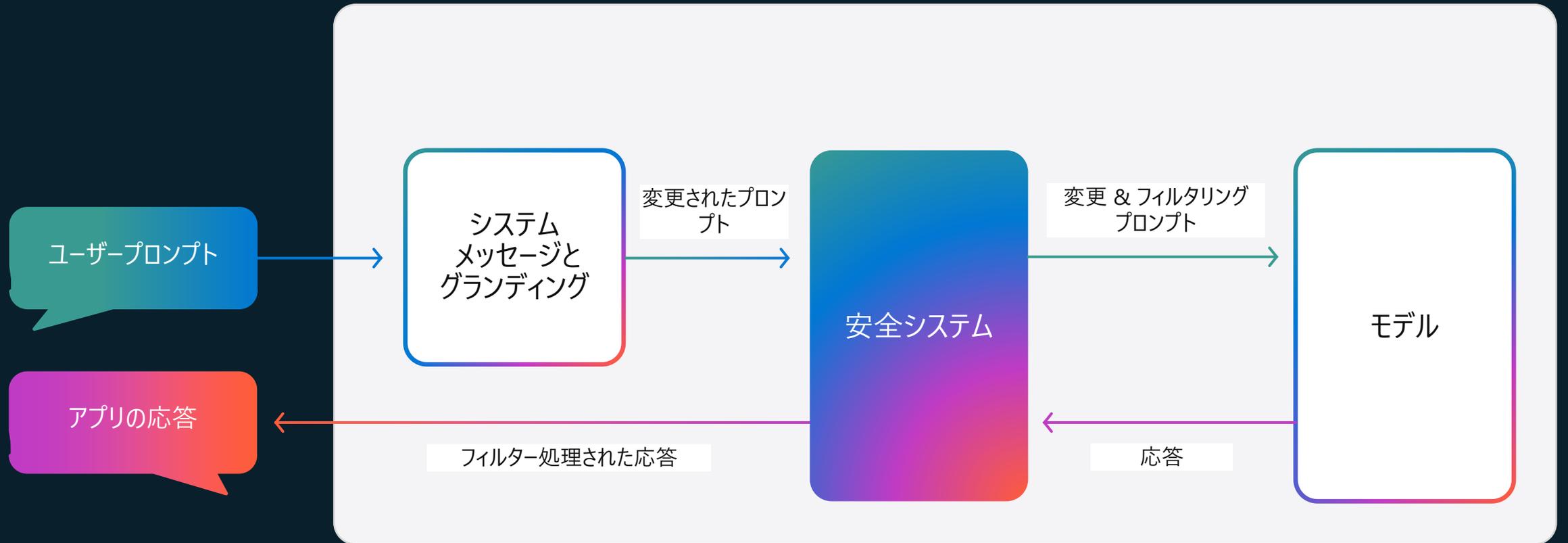
モデルをグラウンドし、その動作を指示する

モデルの入力と出力の監視と保護

ユースケースに適したモデルを選択

安全システムが組み込まれたモデル展開

自社のアプリケーション



Azure AI Content Safety

- センシティブなテキスト・画像に対する

コンテンツフィルター

- ジェイルブレイクに対する

プロンプトシールド

- テキストやコードを保護する

保護済み素材の検出

- ハルシネーションの検出（評価用）

根拠性検出

コンテンツ フィルターを作成する

- フィルターの構成
- 追加モデル (オプション) - プレビュー
- ブロックリストの追加 (省略可能) - プレビュー
- ストリーミングモード (省略可能) - プレビュー
- レビューして終了

追加モデル (オプション) - プレビュー

メイン コンテンツ フィルター上で実行できる追加のコンテンツ セーフティ モデルを有効にします。モデルはテキストのみを処理します。画像のプロンプトまたは入力候補 (DALL-E、GPT-4 Turbo with Vision) には適用されません。
[詳細情報](#)

有効化/注釈付け	フィルター	モデル	説明
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> On	脱獄攻撃に対するプロンプトシールド	脱獄の試行を検出するモデルで、モデル開発者が最初に設定したコンテンツ ポリシーまたは安全ポリシーに違反するために設定された必要な動作をバイパスするモデルを取得するユーザー操作戦略。ユーザー プロンプトで実行します。
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> On	間接的な攻撃に対するプロンプトシールド	間接攻撃 (間接プロンプト攻撃またはクロスドメイン プロンプトインジェクション攻撃とも呼ばれます) を検出するモデル。これは、第三者が、生成型 AI システムがアクセスして処理できるドキュメント内に悪意のある命令を配置する潜在的な脆弱性です。プロンプトで実行します。必須: ドキュメントの書式設定
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> On	保護された素材テキスト	既知のテキストコンテンツ (例: 曲の歌詞、記事、レシピ、選択した Web コンテンツなど) に一致する言語の逆流を検出して保護するのに役立つモデル。完了時に実行されます。
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> On	保護された素材コード	パブリックリポジトリから一連のソースコードに一致するソースコードを検出し、注釈に引用とライセンス情報の例を提供するのに役立つモデル。完了時に実行されます。

責任あるAI 透明性レポート



Responsible AI Transparency Report

How we build, support
our customers, and grow

May 2024



生成AIアプリケーションの責任ある構築方法

生成AIアプリケーションのリリースに関する意思決定の方法

生成AIアプリケーションを構築するお客様をサポートする方法

私たちが学び、進化し、成長する方法

