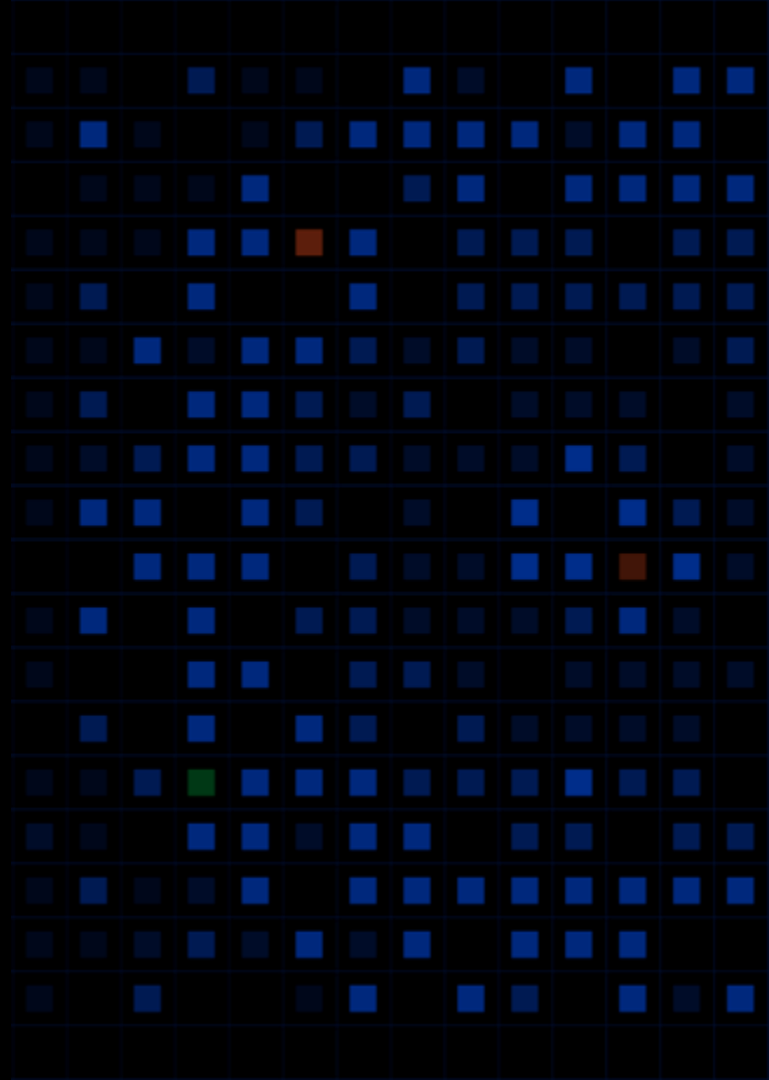


AIセーフティ強化に関する 研究開発プロジェクト速報

About Citadel AI

Citadel AI のご紹介



A woman with dark hair and glasses is looking at a computer screen. She is holding a yellow pen in her mouth. The screen displays a complex interface with various data tables and code snippets. The overall lighting is dim, with a blue and purple glow from the screen.

AI Risk

から社会を守る



Citadel AI とは？

日米欧の世界のトップ企業・認証機関が採用



国際標準業界を代表するBSIが
ハイリスクAIの技術的審査ツール
として採用



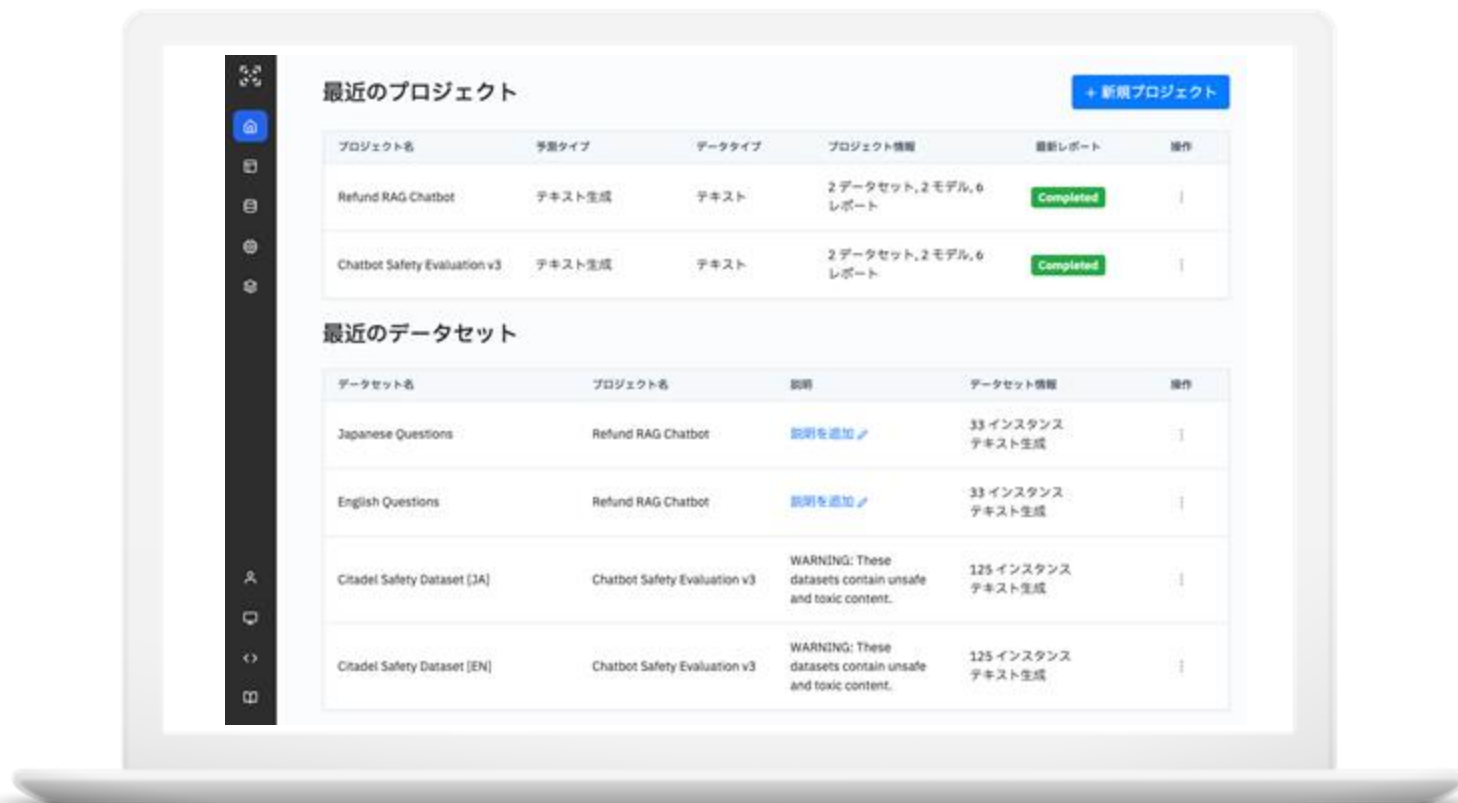
世界No.1病院と称される米国病院
のAI医療システム評価部門が採用



NEDOより「AIセーフティ強化に
関する研究開発」を受託

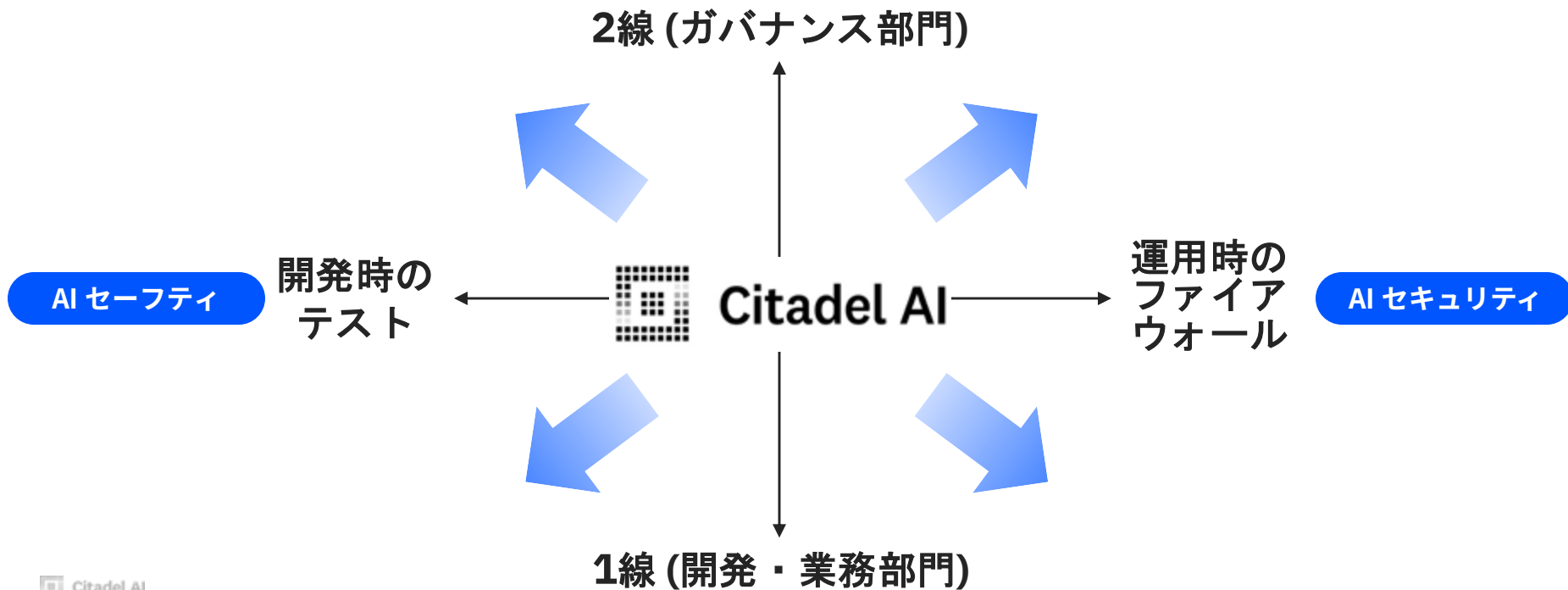


🔍 Citadel Lens - AI 時代の経営力を高める AI ガバナンスツール



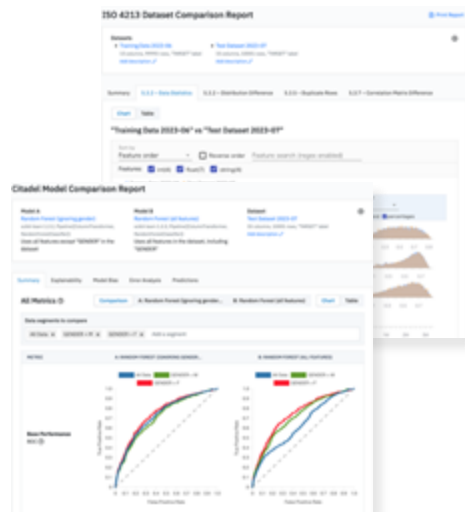
Citadel AI のポジショニング

Citadel AI は、AI ライフサイクル全体（AI セーフティ から AI セキュリティまで）を 1 つのプラットフォーム上で管理できる「AI ガバナンスツール」を提供しています。



経営力を強化する、AI ガバナンスの共通基盤

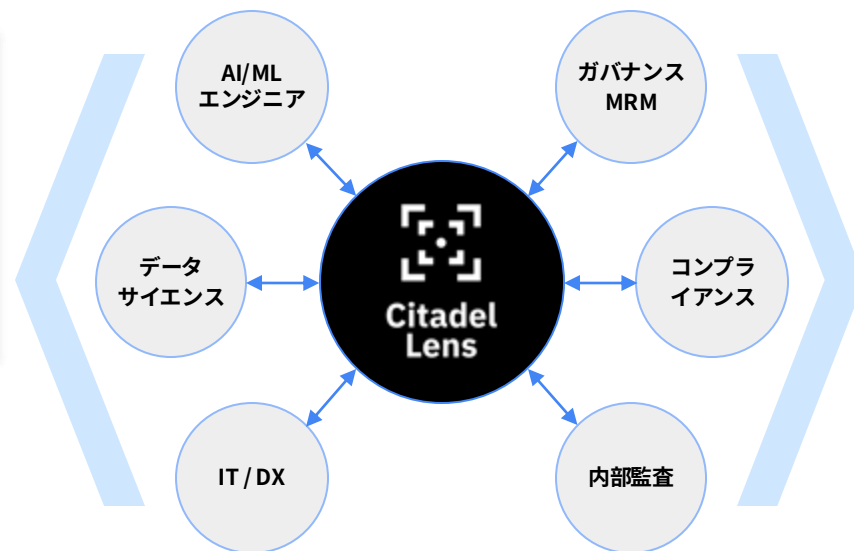
テクニカル・レポート



テクニカル系チーム (1 線)

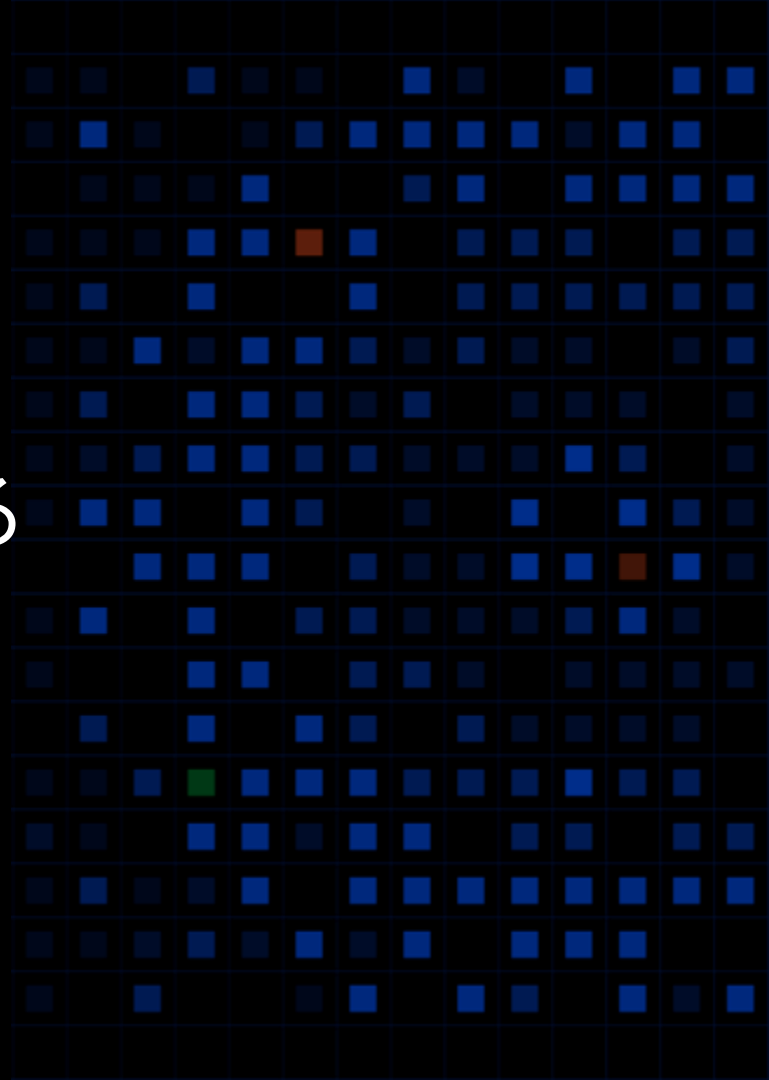
ガバナンス系チーム (2 線)

ガバナンス・レポート



NEDO Project

AIセーフティ強化に関する 研究開発プロジェクト



生成 AI (LLM) も「一発退場」のリスクを伴う



AI チャットボットが
誤った補償内容を案内



ブランド・信頼毀損
SNS 炎上・訴訟



契約書作成支援の生成 AI が
存在しない法律を引用



信用失墜
契約の無効化



従業員が機密情報や
個人情報を生成 AI に入力



コンプライアンス違反
情報漏洩

本プロジェクトの目的

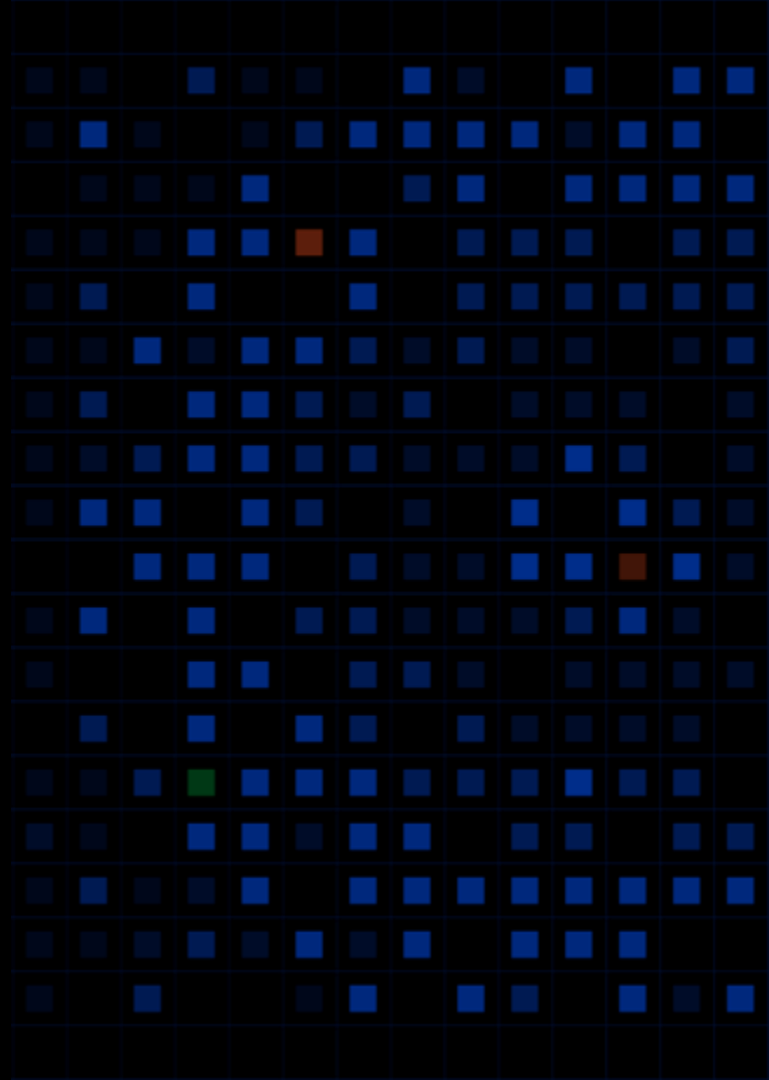
- 生成AIを適切に管理・利用するために必要となる、AIセーフティ基準の策定・普及と、AIセーフティ評価・管理技術の開発を、一体的に行うこと。
 - リスクベースアプローチの基になる 安全性の“ものさし”となる評価・管理技術の開発
 - 暮らし領域での評価手法の開発と実証およびテスト環境構築技術の開発
 - 国際標準化および普及のためのガイダンス等の整備

本プロジェクトにおけるCitadel AIの役割

- 産業界のニーズと技術的有効性に裏打ちされた AIセーフティの評価基準・評価手法の整理
 - 高い安全性や品質が求められる社外向け生成AIアプリケーションの企業向け実装解説の作成
 - AIセーフティを実現する評価指標（メトリクス）の開発と技術検証を通じた有効性の確認
- ⇒ 以下ページの通り企業ヒヤリングを実施

NEDO Project

企業ヒアリング



生成AI活用が進む先進企業28企業にヒアリングを実施

- 生成AIシステムの「デモは簡単にできるものの、サービス化や本番化は難しい」という課題に対してヒアリングを実施
- ベストプラクティス集・事例集として「企業向け実装解説」を取りまとめて公表予定

AI セーフティ強化に関するヒアリング依頼書

国立研究開発法人新エネルギー・産業技術総合開発機構（以下 NEDO）による「AIセーフティ強化に関する研究開発」に関連して、生成 AI の企業内での利活用に関するヒアリングをお願いいたします。

ヒアリングの背景

AI セーフティ強化に関する研究開発プロジェクト概要

NEDOは、研究開発とSociety 5.0との橋渡しプログラム(BRIDGEO)における施策として、生成AIを適切に管理・利用するために必要となる安全性評価技術の開発・普及を目的に、「AIセーフティ強化に関する研究開発」を行います。具体的な研究開発内容は以下の通りです。

- (1)リスクベースアプローチの基になる安全性の「ものさし」(基準)となる技術(評価・管理技術)の開発
- (2)人間拡張など暮らし領域での評価手法の開発と実証およびテスト環境構築技術の開発
- (3)国際標準化および普及のためのガイダンス等の整備

Citadel AIは、本プロジェクトを産業技術総合研究所(以下、産総研)、株式会社コービーと共同で進めています。

Citadel AI の担当範囲

弊社は上記の「AI セーフティ強化に関する研究開発」のうち、「(3)国際標準化および普及のためのガイダンス等の整備」の一部を担当します。具体的な担当範囲は次のとおりです。

- 企業等へのヒアリングに基づく AI セーフティ評価基準・評価手法の整理
- 企業等へのヒアリングに基づく企業向け実装解説の作成およびその技術的有効性の検証

これらの項目について、生成 AI を活用するためのベストプラクティス集・事例集を「企業向け実装解説書」として作成・公開することを予定しています。

目的

「企業向け実装解説書」においては、企業等の各組織で行われている生成 AI を活用するための具体的な取り組みについてベストプラクティスや事例を記載します。これは、現行の各種ガイドラインを研究する形で、ヒアリングに基づき各組織の AI セーフティ評価基準・評価手法事例についての取り組みをボトムアップに収集してまとめる取り組みです。本ヒアリングの結果は「企業向け実装解説書」に記載する内容に用います。

ヒヤリングサマリー

AIガバナンスの意義を経営が認識し、トップダウンで組織的に取り組み、それと同時に、活用推進を掲げて実現を試みている場合、成功しやすい傾向にある。

- AIガバナンスは経営課題
- リスク管理 + 提供価値の最大化を目的
- 攻めと守りのバランスが重要
- 継続的な評価と継続的な改善が必要

AI ガバナンスにおける成功パターン

業態やアプリケーションのユースケースによって求められる内容は異なるが、以下が典型的な成功パターン。

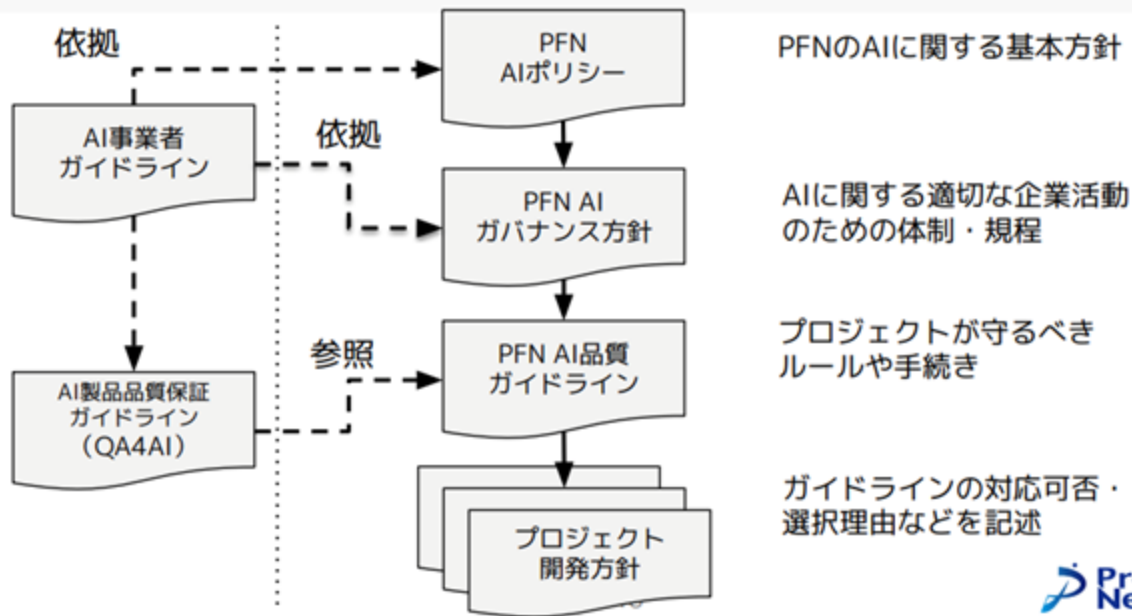
- ガイドラインの整備
- 組織体制の構築・現場や専門家との連携
- AIガバナンス実現の業務フローの確立

ガイドラインの整備

- EU AI法や国内外の各種ガイドライン、関連業法、社内規定等に基づき、AIの活用にあたって遵守すべき事項をまとめたドキュメントを作成
- 組織全体に適用する活用指針をもとに、実業務に適用するためのガイドライン、具体的なチェックリストといった階層構造をもたせることが一般的
- 活用指針は透明性のために外部公開することも

PFN規定・ガイドライン

規程・ガイドライン体系



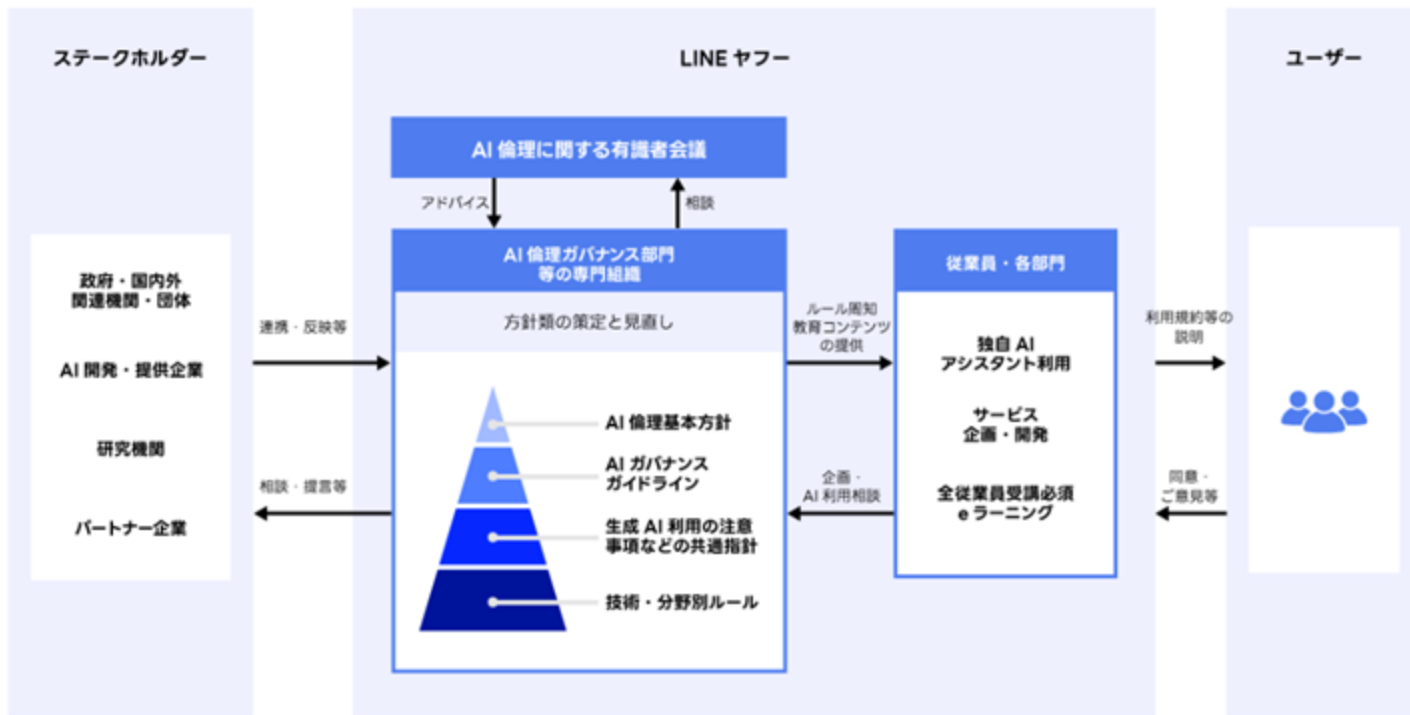
Preferred Networksにおける大規模言語モデル開発と活用での AI安全性の考え方

https://www.digiarc.aist.go.jp/event/4th_grand_canvas/pdf/20250304-4th-grand-canvas-03-ohno.pdf

組織体制の構築・現場や専門家との連携

- AI ガバナンスを推進するための体制構築
- 専門チームを設けることが多い
- 新たな領域で、かつ複数部署が絡むため、役割分担の明確化が重要
- 経営的な視点・実務的な視点と、リスク観点・技術的観点とを組み合わせた組織化が必要

LINEヤフー株式会社の例



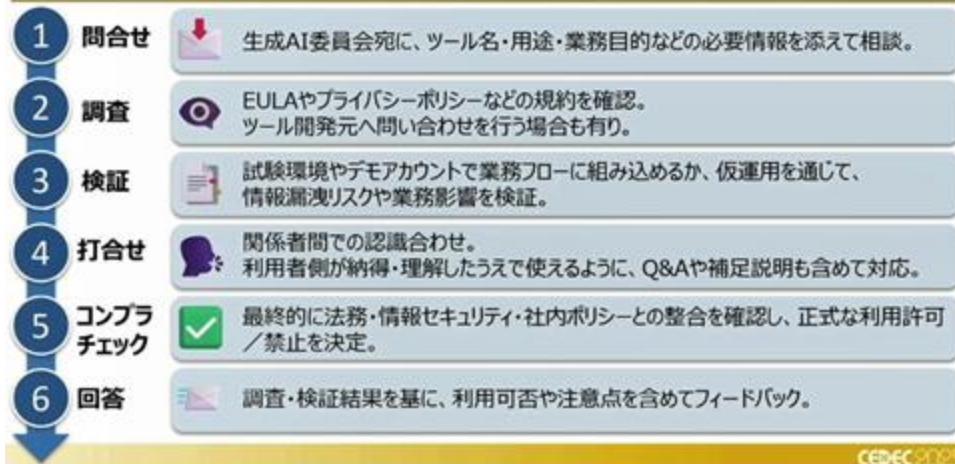
AIガバナンス実現の業務フローの確立

- 攻めと守りの両面を理解したリーダーシップが重要
- 1.5線のような、1線と2線の間立つ組織を設けると共に、現場や専門家など他部署を巻き込むことも効果的
- 具体的な案件をもとに小さく始め、徐々に詳細化し、継続的に改善する
- 自分の業務に該当するベンチマークが世の中に存在するわけではなく、自社業務に応じた、独自の活用推進と評価手法を確立する必要がある

株式会社セガの例

- 現場からのいかなる相談にも答えて、信頼関係を構築
- 現状の変化に、隈無く目を配り新機能を把握
- 教育により、社員全員のリスク感度を向上

ユーザー向けワークフロー



典型的な課題

- リソース不足
 - AI ガバナンス専門のチームが全社で1つで人数も限定的
 - 数多くの技術検証を手作業で行うことが現実的に困難
 - どこまで検証したら十分なのかが分からない
- ルールの周知と徹底
 - AI ガバナンス専門のチームはタイムリーにガイドラインを更新できるものの、すべてのチームがそれを見ってくれるわけではない
 - 生成AIの場合、誰もが手軽に取り組める一方、リスクを知らずに企画立案して推進してしまうことも

NEDO Project

AI Agent開発を通じた作成過程の公開 ～技術検証を通じた有効性の確認～

ヒヤリング結果の課題

- 完成度の高いベストプラクティス事例は収集できた
- 但し、ユースケース毎に評価したい観点は異なるため、汎用的な評価メトリクスを作ることは困難
- AIエージェント開発は、改善を繰り返し、適用範囲やテストデータセットの規模を徐々に拡大する「過程・プロセス」こそが重要
- 最終的な成果物では、抽象的な方法論を実際の行動に落とし込むための具体例（個別のLLM App+データセット+メトリクス）を添えて、より役に立つものにしたい

AIエージェント開発

- 本プロジェクトのための AI エージェントを自ら作成
- 作成過程の検討状況をすべて公開することで、各企業がはじめの第一歩を踏み出しやすくすることを目指す
- AIエージェント並びにテストデータセットを鋭意開発中



Citadel AI

企業URL

<https://citadel-ai.com>

お問い合わせ

info@citadel-ai.com

X (旧Twitter)

<https://x.com/CitadelAI>