Citadel AI

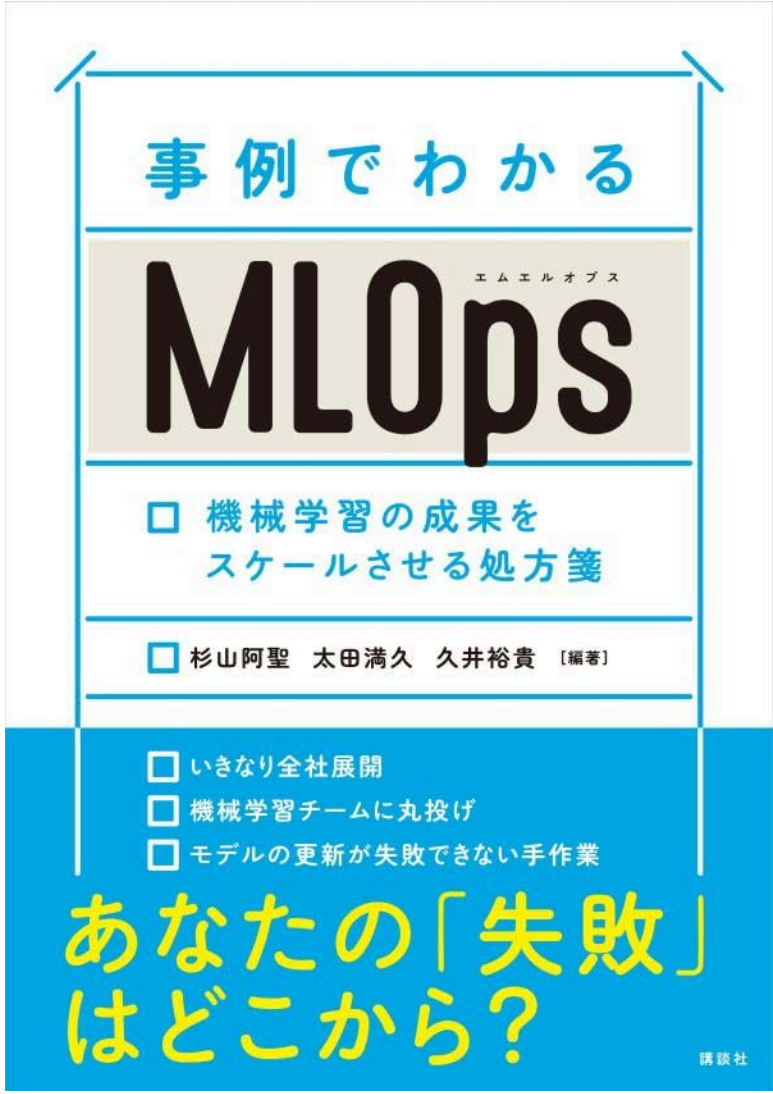
事例で見る AIセーフティと活用の両立

株式会社 Citadel AI 杉山阿聖

© 2021-2025 Citadel AI Inc.

自己紹介

- 杉山 阿聖
- **Research Lead @ Citadel AI**
- Google Developer Expert @ Cloud AI
- MLSE 主査
- 機械学習図鑑 共著
- 事例でわかる MLOps 共著



TOC

- **プロジェクト概要 <-**
- 生成 AI 実践ガイドと企業事例集概要
- 最後に

プロジェクト概要

- 当初の狙い
- ヒアリング
- 当初の狙いとのギャップ
- 発見
- 生成 AI 実践ガイドと企業事例集
- 結果

当初の狙い

- 各種ガイド・ガイドラインを各企業は気にしているだろう
- 生成 AI を活用する企業は AI セーフティに気を使っているはず
- さまざまな企業にヒアリングし、共通して行われている AI セーフティの観点をまとめる
- その観点で評価するための指標・ベンチマークを作成して公開すると多くの企業が使うだろう
- それを各種ガイドラインに反映し国際標準化、認証制度の設立を目指す

ヒアリング

- 生成 AI を活用する企業 (20 社超) にヒアリング

謝辞

本資料の作成にあたってはさまざまな企業・団体にご協力いただきました。ヒアリングにご協力いただいた次の企業に感謝申し上げます (五十音順、株式会社などの法人格を除く)。

- アイレット株式会社
- 株式会社IVRy
- 株式会社Algomatic
- 伊藤忠テクノソリューションズ株式会社（CTC）
- 株式会社インテージ
- KDDI株式会社
- 株式会社サイバーエージェント
- サイボウズ株式会社
- サントリーホールディングス株式会社
- 株式会社SmartHR
- 株式会社すかいらーくホールディングス（すかいらーくグループ）
- セガサミーグループ
- SpiralAI株式会社
- ダイキン工業株式会社
- DataRobot, Inc.
- 合同会社デロイト トーマツ
- 日本生命保険相互会社
- PharmaX株式会社
- 株式会社Preferred Networks (PFN)
- 株式会社メルカリ
- Ubie株式会社 (ユビー)
- LINEヤフー株式会社
- 株式会社LayerX

当初の狙いとのギャップ

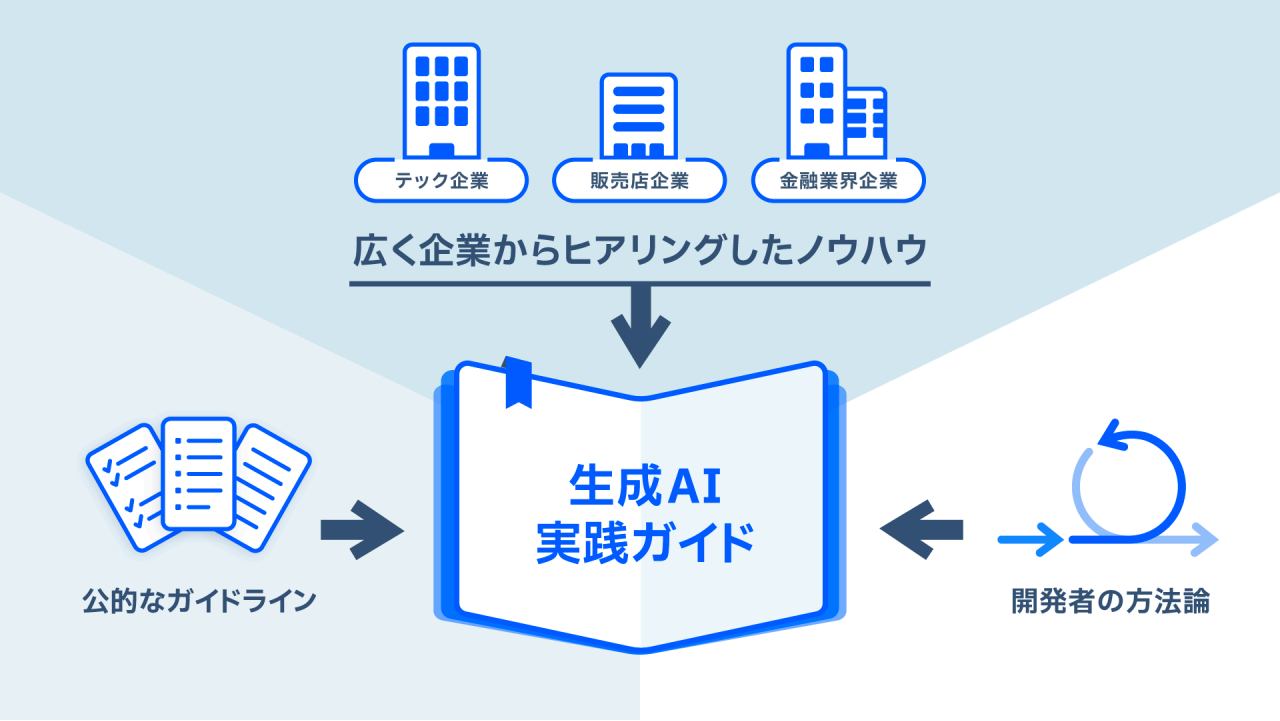
- どこの企業もベンチマークは使ってない
- 対策すべきリスクはバラバラで、共通点はほぼない
- 各企業は AI セーフティを確立してから本番導入するのではなく、利用を拡大しながら段階的に AI セーフティを確立している

発見

- 開発手法はアジャイルに基づく継続的改善
- AI セーフティもアジャイルに基づく学習と改善
- AI ガバナンスは「AI を活用しないリスク」への対応
- アジャイルな AI セーフティを行っていると、国際標準 (ISO/IEC 42001) に基づく認証が取得できる

生成 AI 実践ガイドと企業事例集

- DevOps の継続的改善をベースとするアプローチ
- DevOps と AI セーフティ・AI ガバナンスを統合
- 良い使い方を追求することがセーフティやガバナンスの達成につながる



NEDO事業「AIの安全性確保に関する研究開発・検証等の推進事業／AIセーフティ強化に関する研究開発」の成果報告について - Citadel AI
https://citadel-ai.com/ja/news/2026/04/03/nedo-citadel-ai-2025/

TOC

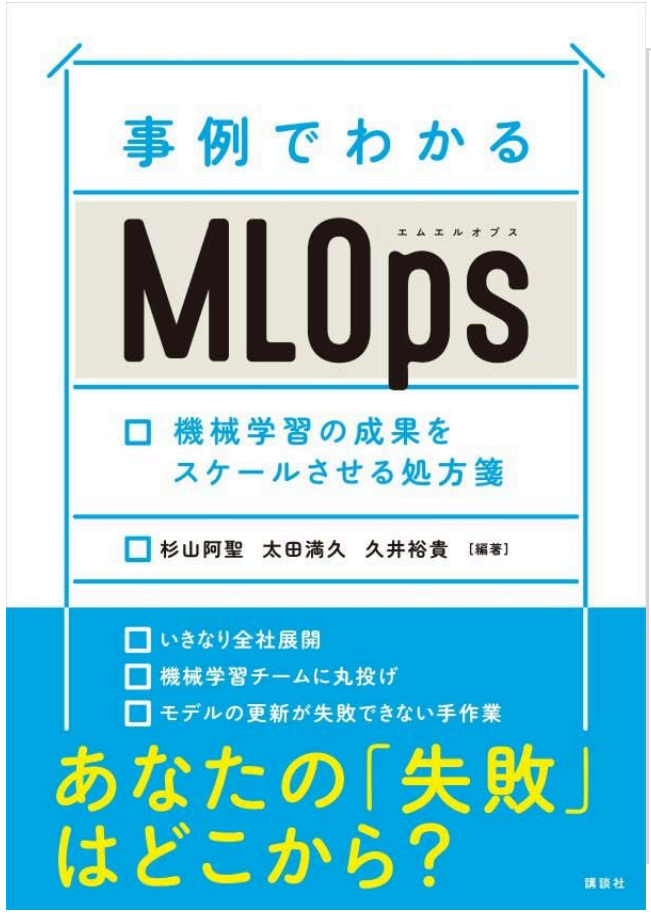
- プロジェクト概要
- **生成 AI 実践ガイドと企業事例集概要 <-**
- まとめ

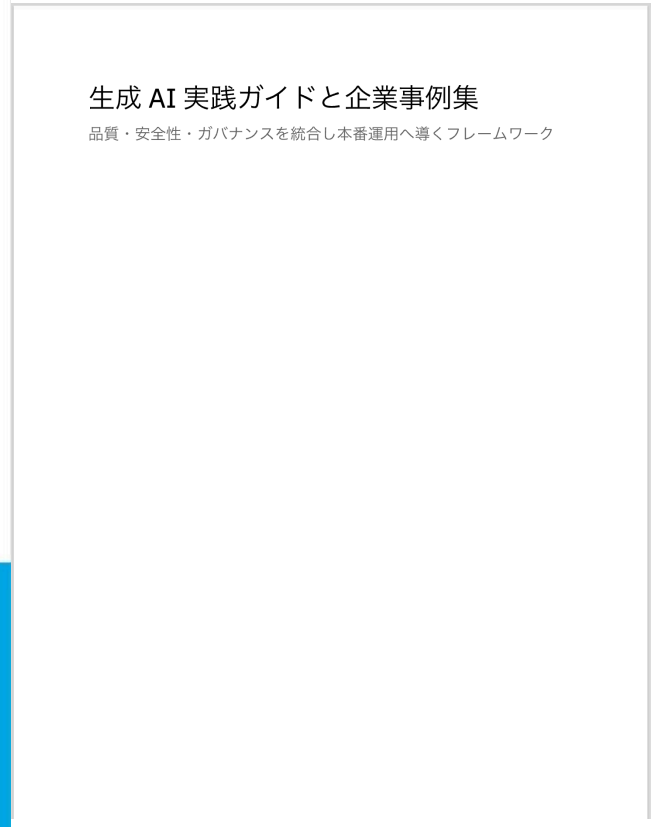
生成 AI 実践ガイドと企業事例集概要

- 全体像
- 技術
- プロセス
- 組織

全体像

- 技術・プロセス・組織で、生成 AI に関する取り組みを整理
- 事例でわかる MLOps の経験を活かした





事例でわかるMLOps　機械学習の成果をスケールさせる処方箋 | 書籍情報 | 株式会社 講談社サイエンティフィク
https://www.kspub.co.jp/book/detail/5369562.html

技術

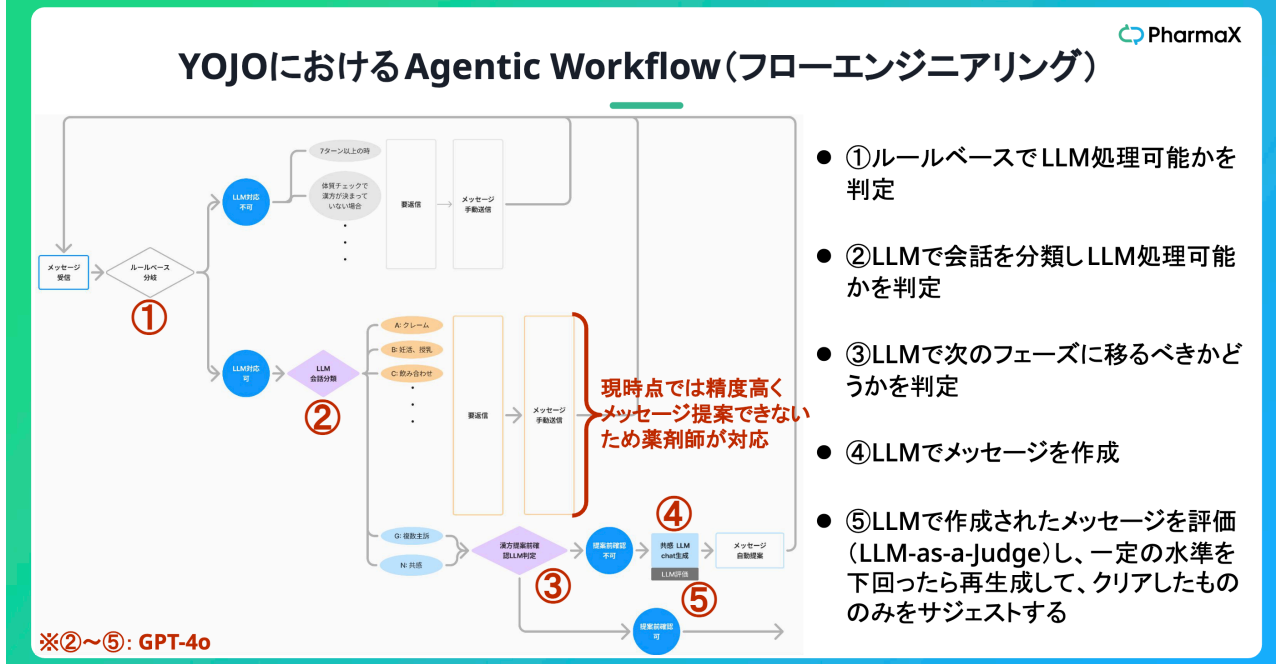
- 情報システムとしての AI エージェント
- **Human-in-the-loop <-**
- ガードレール
- AI エージェントのテスト
- モニタリング
- 技術選定

事例: PharmaX (YOJO)

- ユーザーが OTC 薬を購入できるオンライン薬局
- 問い合わせをルールと LLM で分類
- エージェントで対応不可能な場合、薬剤師にフォールバック

AIエージェントの継続的改善のためオブザーバビリティ
https://speakerdeck.com/pharma_x_tech/aiezientonoji-sok-de-gai-shan-notameobuzababiritei





プロセス

- AI エージェントの開発フェーズ
- 継続的な評価による継続的な改善
- **リスクマネジメント**
- サービスリリース後の迅速なフィードバックループ、モニタリング

リスクマネジメント

次の原則を導入し、リスクマネジメントとアジャイルを両立させる

- ユースケースを事前に特定し、ワークフローを細分化することで、リスクの低いユースケースに限定してAIエージェントを導入する
- 初期はリスクの低いユースケースに限定し、徐々にAIエージェントの対応可能な範囲を広げていく
- 顧客との積極的なコミュニケーションを通じて、リスクの発見やユーザーリテラシーの向上を図る

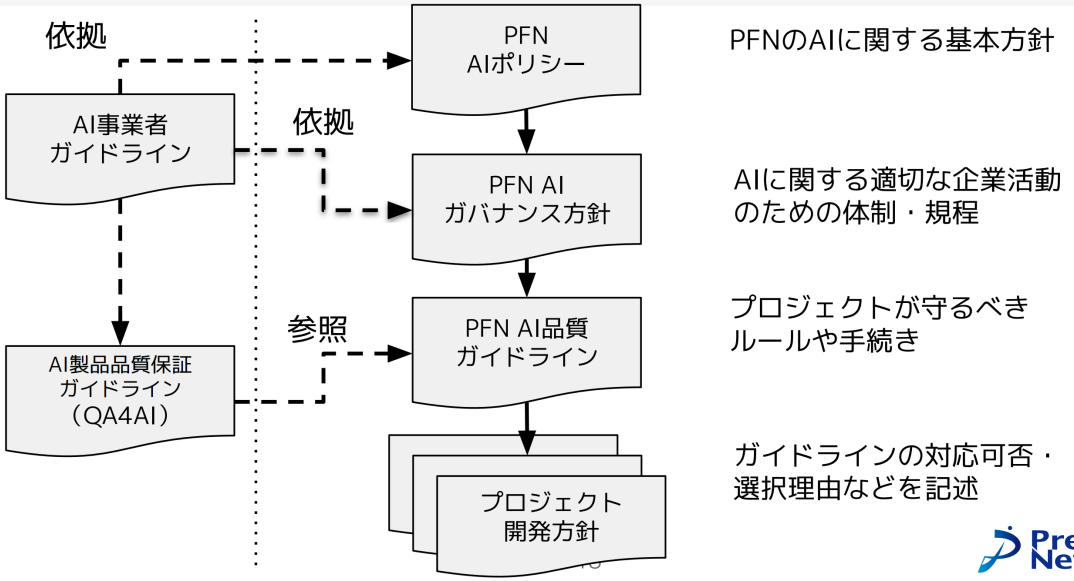
組織

- 組織体制
- **ガイドラインの整備**
- 統合的な AI 相談窓口
- AI 活用支援
- 教育
- ISO 42001 による認証
- トップのコミット

PFN 規定・ガイドライン体系

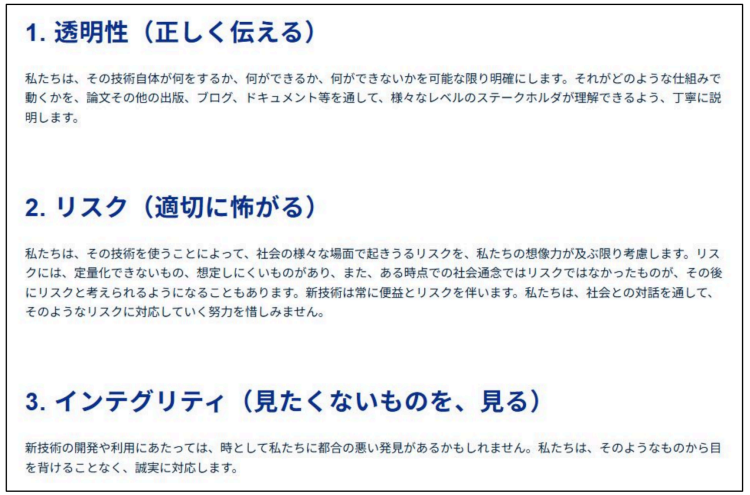
- 階層的なドキュメントを用意し、段階的に個々のプロジェクトにあわせて具体化
- 最終的には「プロジェクト開発方針」を個々のプロジェクトで策定

規程・ガイドライン体系



AI Policy

https://www.preferred.jp/ja/company/aipolicy/



1. 透明性（正しく伝える）

私たちは、その技術自体が何をするか、何ができるか、何ができないかを可能な限り明確にします。それがどのような仕組みで動くかを、論文その他の出版、ブログ、ドキュメント等を通して、様々なレベルのステークホルダが理解できるよう、丁寧に説明します。

2. リスク（適切に怖がる）

私たちは、その技術を使うことによって、社会の様々な場面で起きうるリスクを、私たちの想像力が及ぶ限り考慮します。リスクには、定量化できないもの、想定しにくいものがあり、また、ある時点での社会通念ではリスクではなかったものが、その後にリスクと考えられるようになることもあります。新技術は常に便益とリスクを伴います。私たちは、社会との対話を通して、そのようなリスクに対応していく努力を惜しみません。

3. インテグリティ（見たくないものを、見る）

新技術の開発や利用にあたっては、時として私たちに都合の悪い発見があるかもしれません。私たちは、そのようなものから目を背けることなく、誠実に対応します。

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

TOC

- プロジェクト概要
- 生成 AI 実践ガイドと企業事例集概要
- **最後に <-**

プロジェクトの成果

- 公開したガイドについて直接感想をいただいた
- 産総研やプロジェクトの関係者に実践的な取り組み方を共有、AI セーフティや AI ガバナンスにおける「AI を育てていく」アプローチが浸透しつつある
- 中間報告の内容を実践したとのご報告も

まとめ

- 「AIの安全性確保に関する研究開発・検証等の推進事業／AIセーフティ強化に関する研究開発」に取り組んだ
- 20社を超える、さまざまな業界、規模の企業の AI 活用事例を集約
- 成果を「生成 AI 活用ガイドと企業事例集」として公表
- DevOps と AI セーフティ・AI ガバナンスを GenAIOps として統合