AIパイプラインの監査可能性

Auditability of AI pipelines

運用中のAIの判断を、後から証明できる記録にする

株式会社コーピー　藤井 光

第2回AIセーフティシンポジウム — 2026年10月16日

株式会社コーピー — AIの品質保証に取り組む会社

ミッション — 先端AI技術で人命を救い、平等を拡張する

XAI（説明可能AI）と QA4AI（AIモデルそのものの品質保証）を強みに、ミッションクリティカル領域での AI の社会実装を進めています

研究チーム

・東大・ケンブリッジ大・INRIA（仏）出身の研究者
・XAI・QA4AI の専門家が在籍

プロダクト

・Confide — AIモデルそのものの品質を保証する基盤
・本日ご紹介する Confide Contents は、AIの運用を監査できる形で残す基盤

公的研究

・NEDO AIセーフティ事業（2025年度）
・産総研との共同研究（AIセーフティ）

昨年度：NEDO AIセーフティ事業で、安全性評価の手順を整備

NEDO「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」（2025年度）
産総研・Citadel AI・琉球大学とともに参画し、成果を2026年5月に公表

研究成果

・ISO/IEC 42001（AIマネジメントシステム規格）に整合した「AIマネジメントシステムに基づく生成AI安全性評価プロトコルとその実装ガイド」と評価用テンプレートを整備

・分析・テスト・報告の3フェーズで、生成AIの安全性評価を実務として実施する

・視覚言語モデルを用いた顧客サポートシステムを題材に、レッドチーミングなどによる評価を実施

・組織的なマネジメント要求と技術的な安全性評価とを接続する枠組み

今年度へのつながり

昨年度	評価する手順を整えた
▼	
残った課題	運用中のAIの判断を 後から証明できるか
▼	
今年度	**産総研との共同研究で 運用の記録を担う基盤へ**

出典：NEDOニュースリリース（2026年5月）／コーピーHP https://corpy.co.jp/jp/news/op46vus_

その日、AIは何をしたのか — 答えられますか

**半年前のある日、このAIは何を根拠にこう答え、
誰がそれを確認しましたか？**

1

何を見たか

質問・参照した文書・その時の設定

2

何を判断したか

回答・そのときの関連度と基準値

3

人は何を決めたか

誰が確認し、どう是正したか

普通のログでは、AIの判断を後から証明できない

	普通のログの穴	何が起きるか	必要なこと（Confide Contents の対応）
1	**ばらばらで、つながっていない** → 何を見たか	質問はアプリ、根拠の文書は検索、人の判断はチケットやメール。1つの回答について、後から寄せ集められない	**1つの回答を軸に、質問・引用した文書・関連度と基準値・プロンプトの版・人の判断を同じIDでつなぐ**
2	**後から変わってしまう** → 何を判断したか	回答を作り直すと別の文章になる。文書も更新・削除される。「そのとき何が表示されたか」が分からない	**表示されたとおりの質問と回答をそのまま保存し、改ざんを検知できる台帳にも複写する**
3	**人の判断が残らない** → 人は何を決めたか	誰が確認し、何を是正したかが残らない。残っても書き換えられる	**管理者の確認結果と是正措置を記録し、確認済みの判断は取り消せない形で残す**

もう一つの壁：記録はすぐに膨大になり、人がすべてを見ることはできない

Confide Contents — AIの運用を監査できる形で残す基盤

記録する

・処理と判断を、運用と証拠の2系統で記録
・証拠は業務処理と同じ単位で確定し、取りこぼさない
・改ざんを検知できる台帳にも複写

確かめる

・記録をさかのぼり、何を根拠に答えたかを確認
・回答ごとの関連度と基準値を、そのときのまま残す
・利用者の報告 → 管理者の確認 → 是正措置を記録

規格につなぐ

・記録を ISO/IEC 42001 の運用管理（箇条8.1）の証跡として分類
・重要度に応じて保持期間を区分
・監査ログを絞り込み、CSV／JSONで提出

産総研との共同研究（AIセーフティ）で、42001を実際の運用で成立させるための基盤として開発しています

デモ — システム管理者が、AIの回答をチェックする

1 気づく

確認待ちの報告に気づく

利用者から、問題のある回答の報告が確認待ちに届いている

2 さかのぼる

何を根拠に答えたかを確かめる

表示どおりの質問と回答、そのときの関連度と基準値、引用した文書のページ

3 判断する

人の判断を記録する

判断・確認メモ・是正措置を記録。確認済みの判断は取り消せない

4 証明する

一連の流れを監査ログで示す

「重大」で絞り込み、報告から管理者の判断までをたどる。保持区分を示し、CSVで出力

デモ動画
（当日上映のみ）

記録の7分類 — それぞれが「後から聞かれる問い」に答える

A	**構成と権限**	誰が何にアクセスでき、どの設定が有効だったか	●
B	**変更管理**	ある時点で、システムはどんな状態だったか	●
C	**運用と稼働**	その期間、システムは正常に動いていたか	●
D	**品質とリスク**	出力にどれだけ自信があり、基準の範囲内だったか	●
E	**逸脱と重大事象**	何が、いつ、基準を外れたか	●
F	**選択的フォレンジック**	重大事象の瞬間、システム全体はどんな状態だったか	○
G	**人の判断と是正**	人が何を、いつ決め、何をしたか	●

● 実装済み　○ 開発中

この7分類を、ISO/IEC 42001 の運用管理（箇条8.1）で求められる証跡として整理しています

ISO/IEC 42001 箇条8.1「運用の計画及び管理」に、記録で答える

箇条8.1が求めること：計画したとおりにAIを運用し、それを示せること

箇条8.1が求めること（要約）	答える記録	デモで見えたもの
運用の基準を決め、その基準どおりに管理する	A 構成と権限 B 変更管理	利用者と管理者の役割／回答の基準値
計画どおりに実施されたと確信できる記録を残す	C 運用と稼働 D 品質とリスク	回答ごとの関連度0.62・基準値0.45
意図した結果にならなかったとき、是正する	E 逸脱と重大事象 G 人の判断と是正	利用者の報告 → 管理者の確認・是正措置
変更を管理し、意図しない変更の影響を見直す	B 変更管理 F 選択的フォレンジック（開発中）	設定の変更履歴（デモでは非表示）

認証の審査では、記録を抜き取って運用を確かめる ── AIの運用記録を、そのまま8.1の証跡として出せる形にする

※ 規格本文の引用ではなく、要旨をまとめたものです

AIが一次監査し、人が判断する

研究開発中

記録

すべての処理と判断

▶

AIによる一次監査

高リスクの候補を選ぶ
確認が要る理由を要約
見落とし・滞留を検出

▶

人による判断

受け入れ／却下／修正
理由と是正措置を記録
最終判断・是正は人が行う

▶

証跡

監査に提出できる形で保持

AIの判断を、人が検証できる形にする

重大な記録はAIを通さず、必ず人へ

AIが「問題なし」とした記録も残し、抜き取りで人が確認

答えの分かっている記録を混ぜ、見逃し率を測り続ける

集めた記録をAIが後から一次監査する仕組みを、自社の運用で回し、実際の運用記録で有効性を検証していきます

まとめ

1

AIの安全性は、運用した後に説明できるかで決まる

ただし、記録を人がすべて見ることはできない

2

Confide Contents は、AIの判断を監査できる記録として残す

ISO/IEC 42001 の運用管理（箇条8.1）の証跡になる形で設計

3

次の一歩は、AIが一次監査し、人が判断すること

自社の運用で検証していく

品質マネジメントの土台は、監査できる記録です