6th Grand Canvas: AI品質の未来を共に描く
～AI品質マネジメントネットワーキングシンポジウム～

AI監査の展望：グローバル規制とISO標準から信頼のための実践的技術まで

Navigating the AI Auditing Landscape:
From Global Regulations and ISO Standards to Practical Technologies for Trust

2025-09-17

Yuchang Cheng (鄭育昌)
富士通株式会社 データ＆セキュリティ研究所
　シニアリサーチマネージャー
ISO/IEC JTC 1/SC 42
　エキスパート、SC 42/WG 4国際セクレタリー、プロジェクトエディター

自己紹介

富士通株式会社
データ&セキュリティ研究所
AIセキュリティコアプロジェクト
シニアリサーチマネージャー
鄭 育昌 (Cheng Yuchang) 博士(工学)

ISO/IEC JTC1/SC42 日本委員・各WGエキスパート
SC42/WG4 (use case and application)国際・国内幹事
ISO/IEC 5338 (AIライフサイクルプロセス)、
ISO/IEC TR 24030 (ユースケース)、
ISO/IEC 25589 (human-machine teaming)のリーダー（project editor）

自然言語処理/知識処理

2008 ソフトウェア開発企業入社
社内検索・IMEのエンジンの開発

2011 富士通研究所入社
NLP技術でSE作業効率化の研究開発

2016 多言語知識処理の研究開発

2018 AI倫理・AIトラスト・AIガバナンス
に関する技術の研究開発
ISO/IEC JTC1/SC42のAI標準化活動

2025

AIガバナンス、AIセキュリティ、human-machine teamingなどの研究開発、およびSC42のAI標準化活動を担当

本日のお話：監査可能で信頼できるAIへの架け橋

FUJITSU

課題：AIがもたらす光と影

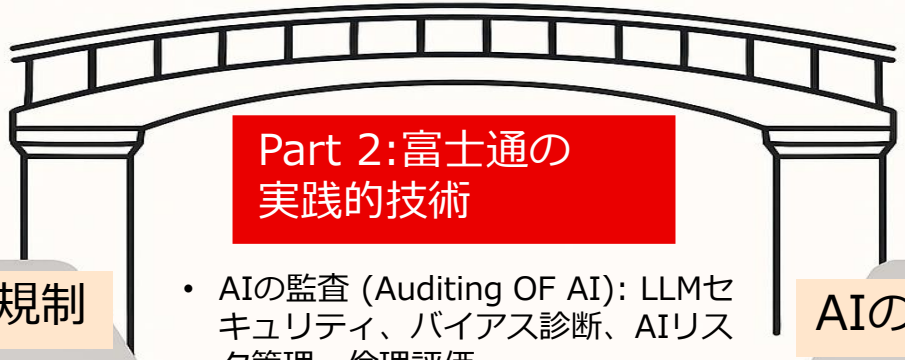
AIの急速な普及がもたらす大きな可能性と、同時に生じるセキュリティや倫理といった新たなリスク

監査可能で信頼できるAIへの架け橋

ゴール：監査可能で信頼できるAI

監査可能で信頼できるAIが実現された社会

Part 1:AI監査のランドスケープ（規制と標準）



Part 2:富士通の実践的技術

- AIの監査 (Auditing OF AI): LLMセキュリティ、バイアス診断、AIリスク管理、倫理評価
- AIによる監査 (Auditing WITH AI): AIによる脆弱性検出・攻撃防御、コンプライアンス自動チェック

グローバルなAI規制

AIの国際標準

FUJITSU

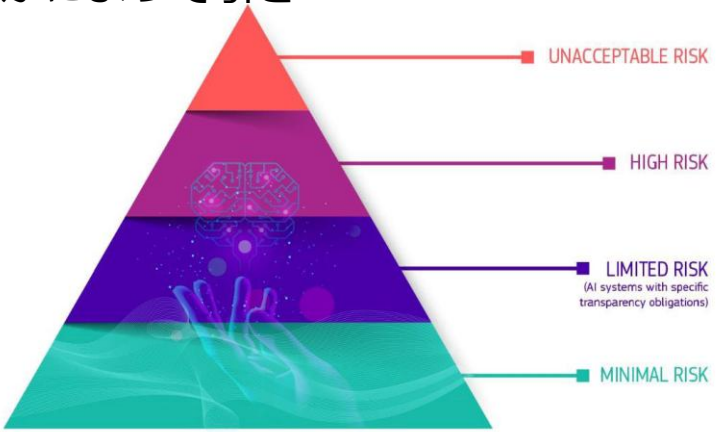
Part 1: AI監査のランドスケープ（規制と標準）

なぜ「AI監査」が不可欠なのか？その背景となるグローバルな規制と国際標準の動向を概観

代表的なAI規制: EU AI Act

FUJITSU

- 目的
  - AIのリスク（健康、安全、基本的人権のリスクなど）は、テクノロジーそのものではなく、AIがどのように使用されるかによって引き起こされる
  - リスクに応じた規制の調整
  - AIへの投資とイノベーションの強化
- 特徴
  - Risk-based アプローチ
  - 世界的に適用
  - 罰則
- 2024年8月発効
  - 禁止AI、汎用AIモデル、そして高リスクAIが対象
  - 標準と適合性評価の必要性



出典:
https://www.soumu.go.jp/main_content/000826707.pdf
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

EU AI Actにおける高リスクAIシステムの要件

FUJITSU

高品質なデータを使用してAIシステムを訓練、検証、評価

トレーサビリティと監査可能（ロギングを含む）

透明性義務

Human oversight

堅牢性、正確性、サイバーセキュリティ

※坂本静生:「EU AI法とその整合規格に基づく対応へ向けて～バイオメトリクスを例に」第37回人工知能学会全国大会 の発表を元に著者が再構成

EU市場へのロードマップ:規制からCEマーキングへ

FUJITSU

①基盤ガバナンスの構築

- リスク管理体制の導入
- データおよび品質ガバナンスの確立
- 技術文書の準備開始

②適合性評価の実施

整合規格に対する評価

- パスA：内部統制による自己評価
  - 整合規格に基づく自己評価（多くの高リスクAIシステムが対象）
- パスB：第三者機関による評価
  - ノーティファイド・ボディの関与（生体認証など、特に重要なシステムが対象）

③登録と宣言

- 高リスクAIシステムを、公開されているEUデータベースに登録
- EU適合宣言書を作成

④CEマーキングと市販後監視の実施

- CEマークを貼付
- 性能やインシデントを追跡するための市販後監視システムを構築・運用

グローバルなビジネス展開においては、EUの整合規格のベースとなる**（かも）ISO/IECの国際標準**を活用することが、最も効率的な戦略

AIの国際標準化団体:
ISO/IEC JTC 1/SC 42 (Artificial Intelligence)

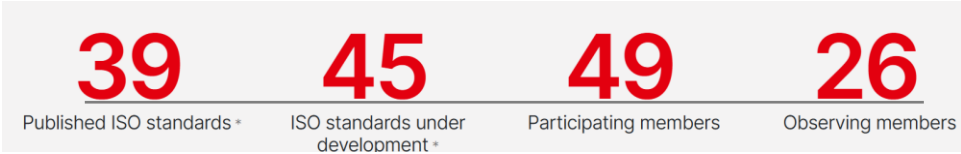
FUJITSU

No.	ワーキンググループ名称	幹事国
JWG 2	SC7との合同WG	UK
JWG 3	AI対応ヘルスインフォマティクス	Japan
JWG 4	機能安全とAIシステム	Italy
JWG 5	自然言語処理	France
JWG 6	AIシステムの適合性評価スキーム	Swiss
JWG 7	金融サービス	Ireland
WG 1	基本標準	Ireland
WG 2	データ	US
WG 3	信頼性	Ireland
WG 4	ユースケースとアプリケーション	Japan
WG 5	AIシステムの計算論的アプローチと計算特性	China
Adhoc 4	SC27とのリエゾン	US

SC42の概要

- 2017年10月、AIの国際標準化を議論するため、ISO/IEC合同技術委員会JTC1/SC42が設立され、審議を開始
- JTC 1/SC 42は、AIの標準化に焦点を当て、AIアプリケーションを開発するJTC 1、IEC、ISOの委員会にガイダンスを提供
- 事務局: ANSI(米国)、議長: Wael W. Diab氏

SC42の数値: 実績とメンバー



日本の貢献

SC42におけるトップレベルの存在感！

- オフィサー: コンビナ2名（セクレタリー２名）、プロジェクトエディター8名
- AI MSS、データ品質、ガバナンス、機能安全、human-oversight、ライフサイクルプロセス、human-machine teaming、ユースケースなど、多くの重要トピックで貢献（うち3つのプロジェクトはChengが担当）

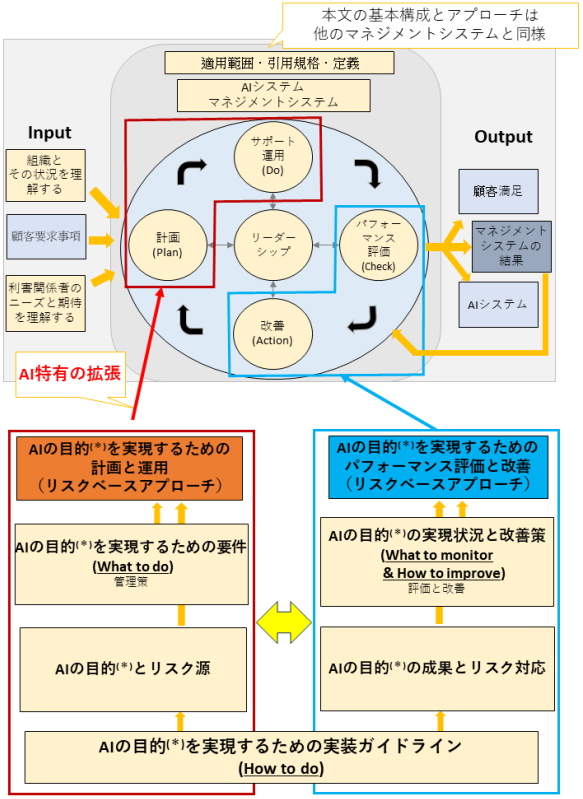
安全・安心なAIシステムの実現の礎：ISO/IEC 42001-AIマネジメントシステム規格

FUJITSU

- 安全・安心なAIシステムの開発・利用に関する国際規格
  - AIシステムを開発、提供、または利用する組織がAIシステムを適切に利用できるようにするため、AI MSSは組織がマネジメントシステムを構築する際に遵守すべき要求事項を規定
  - AI MSSは、AIシステムが信頼性、透明性、説明責任をもって活用されることを確実にするため、リスクの特定と低減、公平性とプライバシーへの配慮を要求
- 特徴
  - AI MSSの構造（PDCA）は、他のMSS（ISO 9001、ISO/IEC 27001など）と同じ
  - Risk-based アプローチ
  - 組織のAI利用または開発の目的のために、必要なアクション（何をすべきか）と実施ガイドライン（どのようにすべきか）を提供

JIS化済み：JIS Q 42001

※経済産業省 プレスリリース:
https://www.meti.go.jp/press/2023/01/20240115001/20240115001.html



＊AIの目的：組織が開発・提供・使用するAIで達成したいこと

なぜAIマネジメントシステム規格は重要なのか？

FUJITSU

- 規制と認証にとって非常に重要
  - AIマネジメントシステム規格は、規制と整合させることができ、規制を遵守するために必要な規格となる
  - AIマネジメントシステム規格はリスクマネジメントを重視しており、AIシステムの認証と監査の基礎となる
  - 広島AIプロセスと同じ方向性

グローバルなAI規制とAIの監査の出発点

- AI MSSに準拠する組織は、安心・安全なAIを開発・利用できる組織
  - AI MSSの認証を受けた組織は、**社会的信頼を得る**ことができる（ISO 9001（品質）やISO/IEC 27001（情報セキュリティ）と同様）

信頼

組織のAIシステムは管理下で運用されている

証拠

AIシステムの規格で要求される証拠を提供する

FUJITSU

Part 2: 富士通の実践的技術

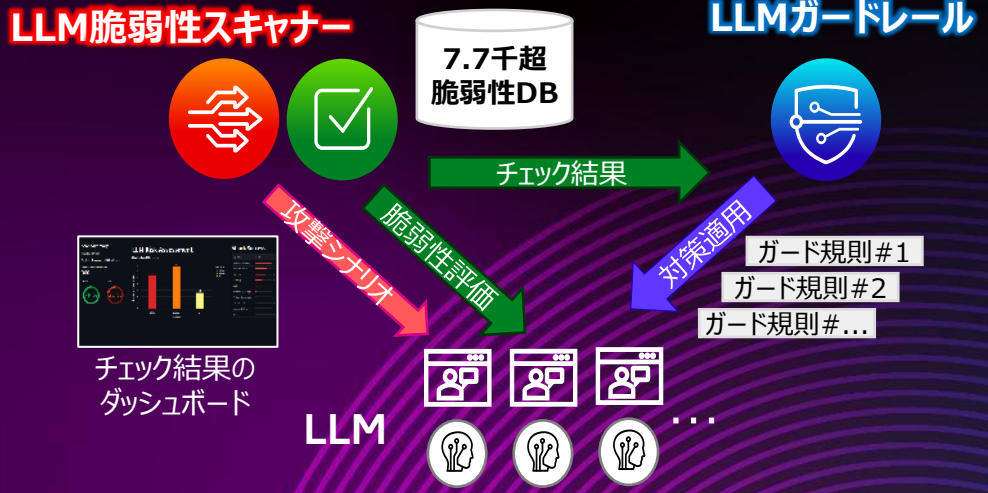
Part 1で示した要求に対し、富士通は「どのように」信頼を実装するのか。その答えとなる実践的技術をご紹介

LLM セキュリティ

FUJITSU

LLMのセキュリティ脆弱性を自動検証し、高度な攻撃から自動防御

■ 業界トップクラスの7.7千超の脆弱性に対応する
LLM脆弱性スキャナーとガードレール



LLM脆弱性スキャナ技術

アダプティブ・プロンプト技術により、人手による検出が困難だった脆弱性を高精度に検出

Sequence of Contexts (SoCs) やPersuasive攻撃など富士通独自の高度な攻撃手法に対応

LLMガードレール技術

脆弱性に対応するガード規則を自動作成しLLMガードレールで防御

プロンプトの文章構造・意図を当社のHypergraph技術で分析、クラスタリングして攻撃or無害を自動判定

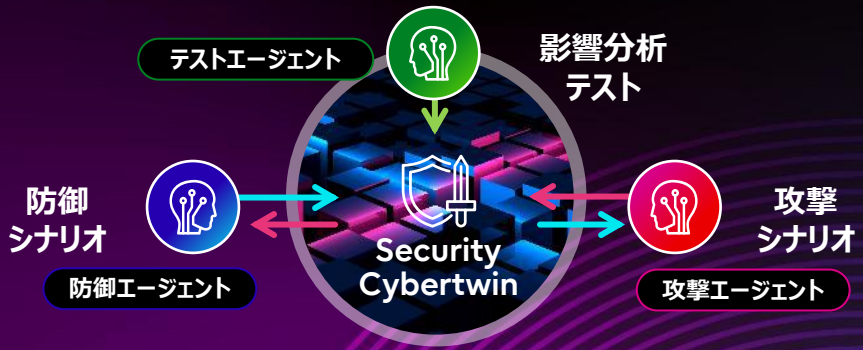
https://note.com/fujitsu_pr/n/nc6868f981195

AIによる高度なセキュリティ

FUJITSU

マルチAIエージェントによるプロアクティブなセキュリティ対応を実現

■ 異なる専門スキルを持つAIエージェント間のナレッジ連携で堅牢性を高める「**共創学習**」の新技術



■ 攻撃と防御シナリオを、実環境を模擬した仮想環境上で協調させ、最適な施策を実行

単独AIの潜在バイアスによる誤判断の問題を解決

新たな脅威を自動で見つけ出し、未然に対処

https://note.com/fujitsu_pr/n/nc6868f981195

AI 倫理 / コンプライアンス

EU AI ACT、倫理、リスクに対応したAIガバナンス技術

リスク
分類・判定

リスク
インパクトアセスメント

バイアス
アセスメント

バイアス緩和

AI Compliance Assistant
EU市場におけるAIシステムのEU AI Actへの準拠をサポート

AI Ethics Impact Assessment
AIシステム構成図を入力するだけで、欧州のAI倫理ガイドラインに基づくすべての倫理リスクをリスト

AI Risk Management Tool
信頼できるAIの開発をサポート

AI Ethics for Fairness (Intersectional Bias)
AIの学習データや意思決定の公平性をWebブラウザ上で簡単に確認し、必要に応じて改善

Fujitsu LLM Bias Diagnosis (LLM Bias & Disinformation)
さまざまな観点からLLMのバイアスを診断し、ユーザーが最適なLLMを選択

Partnership
AKOS AI

Company

AI Risk NG → Action

AI Fairness
Check OK

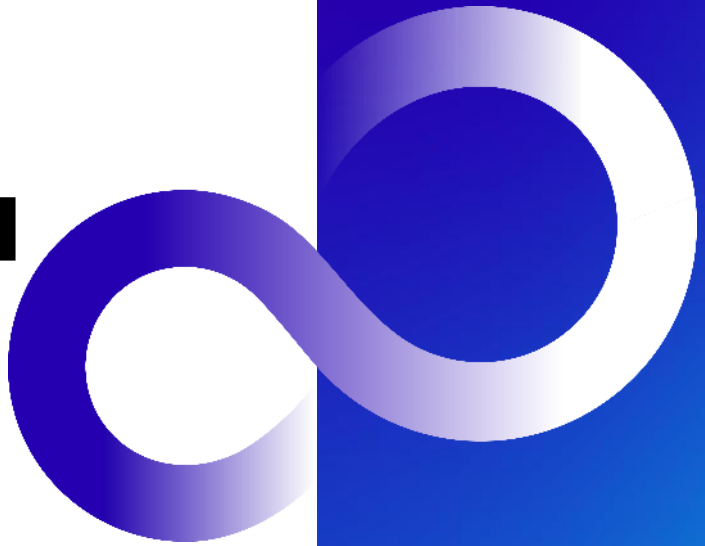
https://pr.fujitsu.com/jp/news/2025/04/18.html

FUJITSU

AIを『守り』、AIで『導く』：自律的な信頼性を実現するガード技術への進化

- 偽・誤情報フェイク対策
- AIによるセキュリティ自動化
- AI 倫理/コンプライアンス
- LLM/RAGセキュリティ

Thank you



If you have any questions,
Please contact:

cheng.yuchang@fujitsu.com