AISI『ヘルスケア領域における AIセーフティ評価観点ガイド』の策定とその意義

自己紹介

井上 真夢
Inoue Mamu

Ubie株式会社 アクセラレーター本部代表・政策渉外参事
日本デジタルヘルス・アライアンス（JaDHA）
「信頼できるAIのイノベーション推進委員会」委員長
日本医療ベンチャー協会主幹

- 2014年 総務省入省。電気通信事業分野の消費者保護や郵政行政、地方の情報通信施策振興やデジタル田園都市国家構想推進などの政策に8年間携わる。
- 2022年 ヘルステックスタートアップのUbie株式会社に入社。ビジネスパートナーアライアンスなど事業開発チームを経て、Public Affairs（政策渉外）担当に。日本デジタルヘルス・アライアンスでは生成AI活用ガイドの策定をリーダーとして牽引。

風間 正弘
Kazama Masahiro

Ubie株式会社 Chief AI Officer（CAIO）
津田塾大学 非常勤講師
国立国語研究所 外来研究員

- 2015年 東京大学大学院を卒業、リクルートホールディングス入社。様々な領域のデータ分析や機械学習アルゴリズム開発を担当。2018年よりIndeedに異動。
- 2020年 ヘルステックスタートアップのUbie株式会社に入社。AI問診の開発チームをリードし、2023年から生成AI活用を担当。
- 2018年 Forbes 30 Under 30 Japanを受賞
- 2022年 推薦システム実践入門(オライリー・ジャパン)執筆
- 2022,23年 東京都立大学非常勤講師

医師とエンジニアが創業した
ヘルステック企業

設立	2017年5月
従業員数	220 (2026年1月時点)
拠点	日本・米国
主要株主	

代表取締役 医師
阿部 吉倫

代表取締役
久保 恒太

2026年4月3日、ヘルスケア領域における AIセーフティ評価観点ガイド策定

日本デジタルヘルス・アライアンスの概要

JaDHA 日本デジタルヘルス・アライアンス
The Japan Digital Health Alliance

組織名・設立

- 日本デジタルヘルス・アライアンス（JaDHA）
- 製薬デジタルヘルス研究会および日本DTx推進研究会を統合し、2022年3月14日に設立。（会長：三友周太）

設立背景

- コロナ禍は社会におけるデジタル化の重要性が一層認識される契機に。先進的プログラム医療機器の実用化を促す施策の検討が進む。
- 「デジタルだからこその価値」の評価、柔軟性のある制度・規制の実装が重要。

活動内容

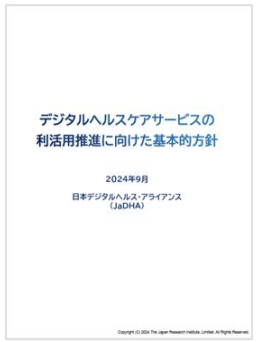
- 業界の垣根を超えた横断的研究組織の組成と活動により、
- 産業の発展、関連サービスや技術の普及促進を阻害する課題を深く洞察、デジタルヘルス産業の発展を巡る課題解決の在り方を提言する。

会員企業

- 大手医薬品・医療機器メーカー、デジタルヘルスベンチャー企業、大手ICT企業、デジタルヘルスに新規事業として取り組む企業など約**100社**が参加。

日本デジタルヘルス・アライアンス（ JaDHA）の活動

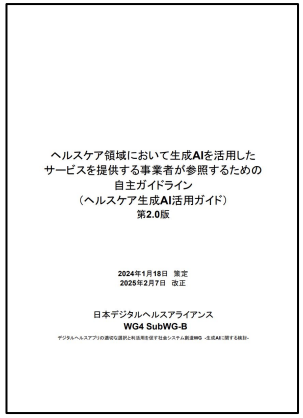
ビジョン・基本指針の策定



・2024年10月「デジタルヘルスケアサービスの利活用促進に向けた基本的方針」

・2025年3月策定ビジョンペーパー「デジタルヘルスリテラシーへの配慮を通じた産業振興と社会課題解決の両立」

ヘルスケア領域に特化した生成AI活用のガイドライン策定



- WG4のSubWG-Bでの活動
- 2024年1月 第1.0版策定
- 2024年4月 AI事業者ガイドラインに掲載
- 2025年2月 第2.0版策定

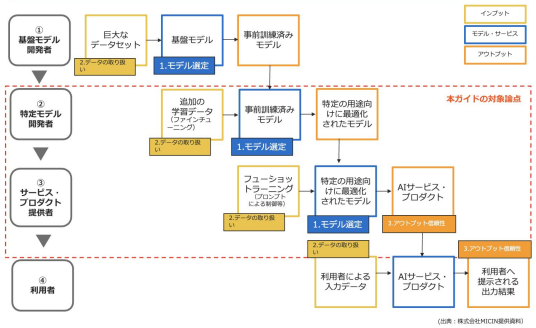
国内外の業界団体/ 産学官連携

2024年10月
米国DTAと国際的な協働に関する
覚書締結

2025年1月
デジタルヘルスリテラシーをテーマにしたイベント「 JaDHA Innovation Forum」の開催

チェックポイント全体像

生成AIのバリューチェーン



2026年度の期初委員会運営体制

		委員長会社
1	薬事・診療報酬制度の在り方検討委員会	テルモ
2	SaMD QMS検討委員会	塩野義
3	生活者や患者に向けた広告規制の在り方検討委員会	事務局
4	デジタル医療サービスの円滑な利活用にむけた基盤構築委員会 ※委員会としての設置に向けた活動方針・詳細計画づくり、調査、会員意向聴取等をFY26早々に実施し、遅くともFY26/3Qに発足。	NTT data
5	デジタルヘルスサービスの認知向上委員会	シミック
6	信頼できるAIのイノベーション推進委員会	Ubie
7	SaMDにおけるデータ利活用の在り方検討委員会	DeNA

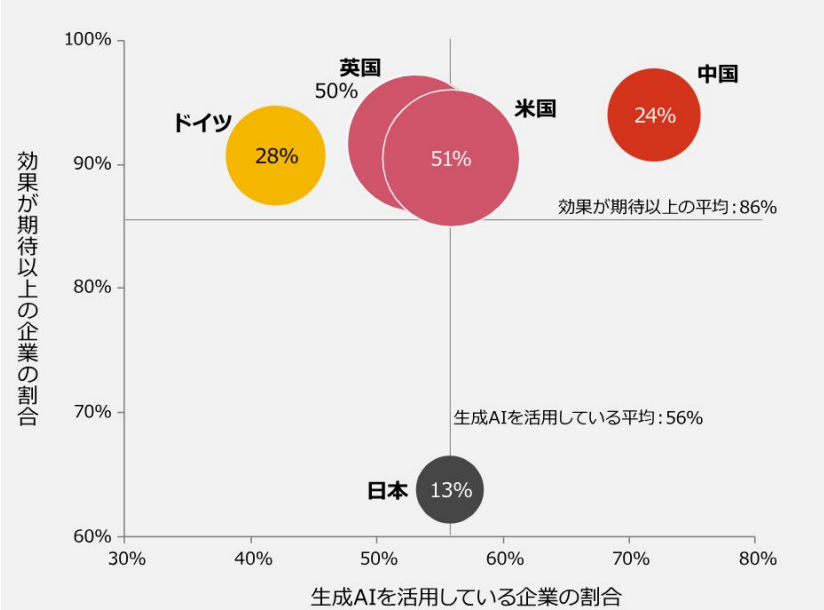
出典: 日本総研作成

”信頼できるAI”：イノベーションと安全性の両立

日本はAI投資額自体が少なく、実際に”使われるAI”の社会実装への投資が不足。

AI基本計画（2025年12月閣議決定）**で柱となったのが「信頼できるAI（Trustworthy AI）」。**

「イノベーションの推進と安全性の両立」を日本のAI産業の「勝ち筋」としている。



https://www.pwc.com/jp/ja/knowledge/thoughtleadership/2025/assets/pdf/generative-ai-survey2025.pdf

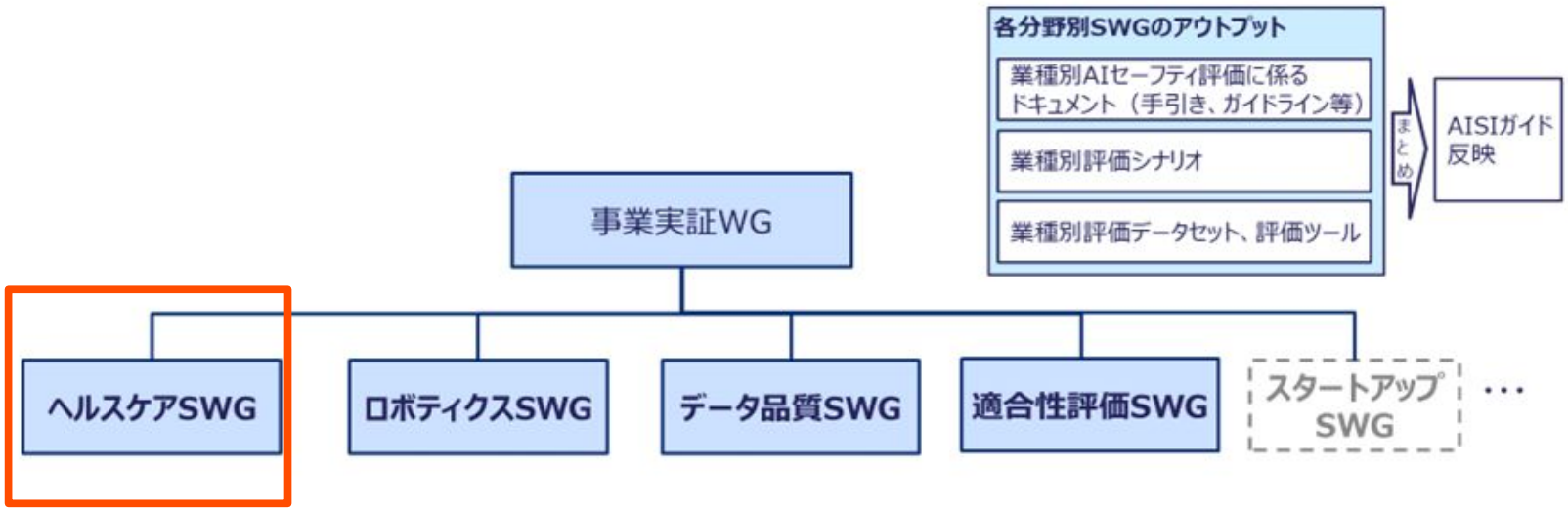
AI基本計画（2025/12閣議決定）

AISI ヘルスケア SWGの位置づけ

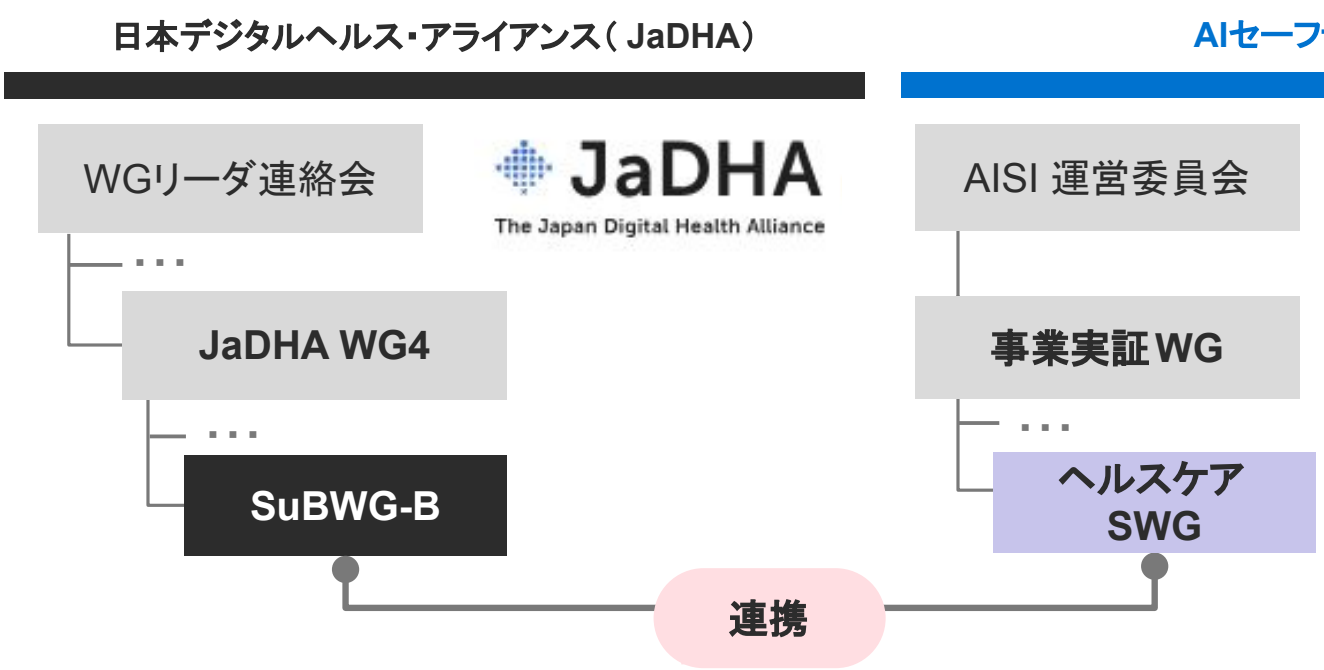
事業実証ワーキンググループ設置のお知らせ

AIの安全安心な活用が促進されるように官民の取組を支援するAIセーフティ・インスティテュートは、AIセーフティ評価に関するワーキンググループ（事業実証WG）を、AISI運営委員会の下のテーマ別小委員会として設置いたしました。AIセーフティ評価の活動を広く一般に普及させ、AIの利活用を促進させることを目的とし、民間事業者を中心に多様なステークホルダーが参画し、参画機関間の連携を図る場を提供、WG活動を推進して参ります。

事業実証WG内には、下記４つの分野別サブワーキンググループ（SWG）を設置します。必要に応じSWGの増設を計画いたします。



体制・参加組織



Ubie株式会社（SWGリーダー）

株式会社Awarefy

シミックホールディングス株式会社

株式会社MICIN

JaDHA特別顧問

味の素株式会社

SherLOCK株式会社

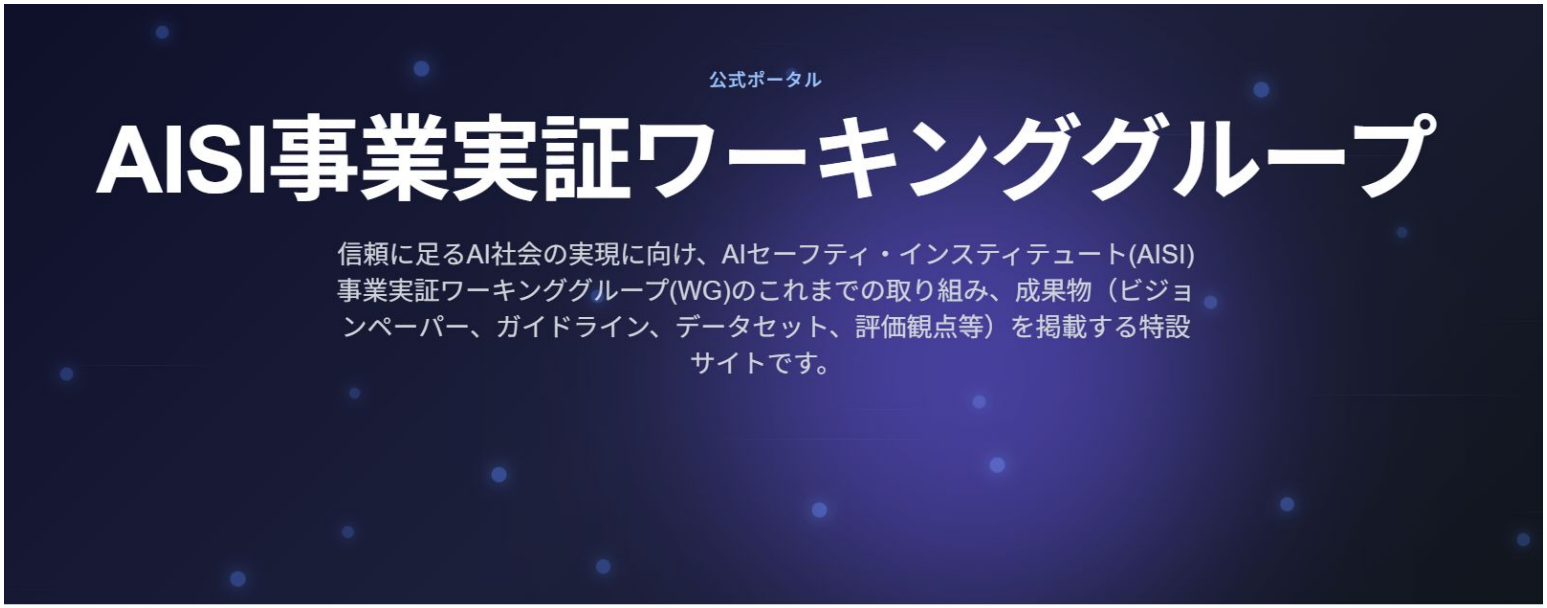
JaDHA WG4　デジタルヘルスアプリの適切な選択と利活用を促す社会システム創造ワーキンググループ

SuBWG-B　生成AIに関する検討ワーキンググループ

（参考）メディア掲載実績

1. AISI、「ヘルスケア領域におけるAIセーフティ評価観点ガイド」を策定【日本経済新聞】
   https://www.nikkei.com/compass/content/PRTKDB000000014_000178657/preview
2. Ubie，JaDHAのワーキングリーダーとしてAIセーフティ・インスティテュート（AISI）と連携し，『ヘルスケア領域におけるAIセーフティ評価観点ガイド』を策定【INNERVISION】
   https://www.innervision.co.jp/products/release/20260518
3. IPA「ヘルスケア領域におけるAIセーフティ評価観点ガイド」策定【ScanNetSecurity】
   https://scan.netsecurity.ne.jp/article/2026/04/16/55072.html
4. AISIが策定したヘルスケアAIセーフティ評価ガイドの重要性と活用方法【VOIX】
   https://voix.jp/business-cards/aisi-healthcare-ai-safety-guide/
5. JaDHA　医療・ヘルスケア業界「AIセーフティ評価観点ガイド」公表　"信頼できるAI"で実務者向け手引き【ミクスOnline】 https://www.mixonline.jp/tabid55.html?artid=80042
6. AISI、「ヘルスケア領域におけるAIセーフティ評価観点ガイド」を策定【PR TIMES 】
   https://news.nicovideo.jp/watch/nw19122582?news_ref=watch_20_nw18815860
7. UbieがJaDHAと協力し、ヘルスケア分野のAIセーフティ評価ガイドを策定する意義【VOIX】
   https://voix.jp/business-cards/ubie-jadha-ai-safety-guide/

AISI 事業実証ワーキンググループ　公式ポータル



https://aisi.go.jp/lp01/

活動・成果報告コンテンツ

AI利活用促進に向けたAIセーフティ評価に関する事業実証WGでは、ヘルスケアやロボティクスといった分野別のサブワーキンググループ（SW

事業実証WG
事業実証WGの活動全体に関するコンテンツを掲載

ヘルスケアSWG
ヘルスケア分野におけるAIセーフティ評価に関する事業実証の取り組み等

ロボティクスSWG
ロボティクス分野におけるAIセーフティ評価に関する事業実証の取り組み等

データ品質SWG
データ品質マネジメントガイドブックを基に、実装への具体化、評価ツールの検討等

適合性評価SWG
AI分野における適合性評価の手法確立、実運用を見越した国内体制構築の検討等

動画・配信アーカイブ
事業実証WGの活動を中心に、AISIに関連するコンテンツを掲載

リンク
関連情報へのリンク

ヘルスケア領域における AIセーフティ評価観点ガイド
～Trustworthy AI（信頼できるAI）の実現を目指して～
（概要版）

AI セーフティ・インスティテュート
事業実証ワーキンググループ
ヘルスケアサブワーキンググループ

2026年4月3日

AISI Japan
AI Safety Institute

「ヘルスケア領域におけるAIセーフティ評価観点ガイド」

AISI Japan AI Safety Institute

- Trustworthy AIの実現に向けた国内外の潮流を踏まえ、ヘルスケア領域に特化したAIセーフティ評価の実践的な指針を提供。
- 実用性を追求した設計とし、AISIの評価観点※に加えて実践を想定したフェーズごとの評価項目を整理。

ガイドの構成

1. **本ガイドの背景と目的**
   - ヘルスケア領域におけるAIセーフティの重要性、ガイドのスコープ（AI提供者、Non-SaMD、テキスト生成）
2. **医療・ヘルスケア分野における AI動向**
   - 技術動向、政策・業界動向、市場動向
3. **AIセーフティ評価の 10観点**
   - AISIの評価観点ガイドをヘルスケア領域へ適用した際の各観点の概要、想定リスク、実際の事例、評価項目例
4. **AIプロダクト開発における AIセーフティ評価の実践**
   - フェーズごとの主な評価観点と具体的な実践方法、確認すべき事項のチェックリスト
5. **今後の展望**
   - 本ガイド発展の方向性

第1章：本ガイドの背景と目的

AISI Japan AI Safety Institute

- AI提供者かつNon-SaMD（医療機器プログラムに該当しないAIプロダクト）、テキスト生成AI（LLM）を対象に検討。ユースケースはBtoB（医療従事者向け）/BtoC（生活者向け）を想定。
- サービスの企画・開発・運用・リスク評価に携わる経営層やプロダクトマネージャー、エンジニア・法務等が想定読者。

本ガイドのスコープ

対象者	**AI提供者**（学習済みの生成AIモデルをAPI経由等で利用してプロダクトやサービスの開発を行う事業者）
対象のプロダクト	**非医療機器プログラム（Non-SaMD）**
対象の生成AI	**テキスト生成AI（LLM）**（画像生成AI・音声生成AI等は本ガイドの対象外とする）

対象のユースケース

カテゴリ	ユースケース例
BtoB （医療従事者向け）	文書作成支援、情報検索・要約、カルテ入力補助、患者説明資料の作成支援、 医療文献の検索・要約 等
BtoC （生活者向け）	健康相談チャットボット、セルフケア支援、メンタルヘルスケアアプリ、服薬リマインダー、健康管理アプリ 等
その他	製薬企業向け情報提供支援、臨床研究の文献レビュー支援 等

第2章：医療・ヘルスケア分野におけるAI動向

AISI Japan AI Safety Institute

- 医療・ヘルスケア分野における、医療特化型LLMをはじめとしたAI技術の動向と活用領域を概観したうえで、各国の政策・業界団体等によるAIセーフティへの取組、および国内外のAI活用プロダクトの市場動向と具体事例を紹介。
- ヘルスケア分野に特化した、各国のAIセーフティ評価に係る具体的な事例についても紹介。

国内外のAI技術動向

- **英語対応の医療 LLM・データセット等（例）**
  - MedLM・ Med-Gemini （Google）、Meditron-70B（EPFL）、MedQA　…等
- **日本語対応の医療 LLM・データセット等（例）**
  - NII：SIP-jmed-llm-2-8x13b・NTCIR・AnswerCarefully Dataset（一部）
  - 東大 相澤研： SIP-jmed-llm・JMedBench
  - 東大 松尾・岩澤研 × ELYZA：ELYZA-LLM-Med
  - 奈良先端大 荒牧研：JMED-DICT、JMED-LLM

  …等

国内外のAI政策・業界動向

- **各国動向**
  - 日本：AI推進法（2025）・AI事業者ガイドライン・AISI設立
  - 欧州：AI Act（2024）リスクベース規制・汎用AI行動規範
  - 米国：America's AI Action Plan
  - 英国：AI法案・AI Security Instituteへ改名・方針転換
- **業界動向**
  - 日本：JaDHA・HAIPによる生成AI活用ガイドライン策定
  - 米国：米国医師会レポート・CHAI「責任あるAIガイド」
  - 英国：MHRAによる規制サンドボックス「AI Airlock」実施

第3章：AIセーフティ評価の10観点

AISI Japan AI Safety Institute

- AISIガイドにおける評価の10観点について、ヘルスケア領域に適用した場合の評価の要点を整理。
  - 想定され得るリスクの例：各観点ごとにヘルスケア領域で特に懸念されるリスク
  - 実際の事例：国内外で報告されている関連事例を紹介
  - 評価項目例：サービス・プロダクトの企画・開発・運用において確認すべき評価項目を例示

No.	評価観点	ヘルスケア領域におけるリスク概要
1	有害情報の出力制御	医療・健康に関する危険な情報（自傷・暴力の助長、根拠を欠く治療法等）が出力され、患者の生命・健康や医療従事者の業務に直接的な被害をもたらすリスク
2	偽誤情報の出力・誘導の防止	ハルシネーションにより架空のエビデンスや誤った薬剤情報等が生成され、患者の生命・健康や医療従事者の業務に直接的な被害をもたらすリスク
3	公平性と包摂性	特定の属性の患者に対しAIの精度や品質が低下し、不利益が生じるリスク
4	ハイリスク利用・目的外利用への対処	Non-SaMDが事実上の医療機器として利用される「目的外利用」により、法規制違反等が生じるリスク
5	プライバシー保護	要配慮個人情報を含む医療・健康情報が漏えい・不正利用され、患者のプライバシーが侵害されるリスク
6	セキュリティ確保	プロンプトインジェクション等の攻撃により、医療情報の改ざんや機密データの漏えいが生じるリスク
7	説明可能性	AI出力の根拠が不透明なまま出力され、医療従事者の誤った医療行為や患者の不信につながるリスク
8	ロバスト性	方言・略語・非標準的な医療用語等の多様な入力に対し出力品質が不安定となり、誤った判断を招くリスク
9	データ品質	不正確または陳腐化した医療データに基づく出力が、患者の生命・健康や医療従事者の業務に直接的な被害をもたらすリスク
10	検証可能性	事後検証や第三者監査が困難な状態で問題発生時の原因究明ができず、社会的信頼を損なうリスク

（第3章の参考例）有害情報の出力制御の評価観点

AISI Japan AI Safety Institute

想定リスク:

医療・健康に関する危険な情報（自傷・暴力の助長、根拠を欠く治療法等）が出力され、患者の生命・健康や医療従事者の業務に直接的な被害をもたらすリスク

想定され得るリスクの例

- 人の生命・身体に直接的な危害を及ぼす行為の容易化・助長
- 健康被害・医療機会の逸失（誤った健康情報の提供等）
- 心理的依存と社会的孤立
- 基本的人権・尊厳の毀損

実際の事例

- **摂食障害支援チャットボット（ Tessa）の停止**：2023年、摂食障害患者向けに導入されたAI（NEDA支援団体のTessa）が、ユーザーに対し症状を悪化させる恐れのある「具体的なカロリー制限」や「減量方法」を推奨し、運用停止に追い込まれた事例※

フェーズ	評価項目例
①プロダクト設計	有害情報のリスク類型と許容基準が定義されていること
②モデル選定	基盤モデルの選定において有害情報出力の抑制能力が評価されていること
③プロダクト実装	入力から出力に至る多層的な有害情報の防止機構が実装されていること
④プロダクト検証	有害情報出力の抑制が実証的に検証されていること
⑤プロダクト導 入・運用	本番環境において有害情報出力が継続的に監視され、迅速に是正されること

※参考: Aratani, Lauren. “US Eating Disorder Helpline Takes down AI Chatbot over Harmful Advice.” The Guardian, 31 May 2023, www.theguardian.com/technology/2023/may/31/eating-disorder-hotline-union-ai-chatbot-harm. 最終閲覧日 2026年3月27日.

第4章：AIプロダクト開発におけるAIセーフティ評価の実践

AISI Japan AI Safety Institute

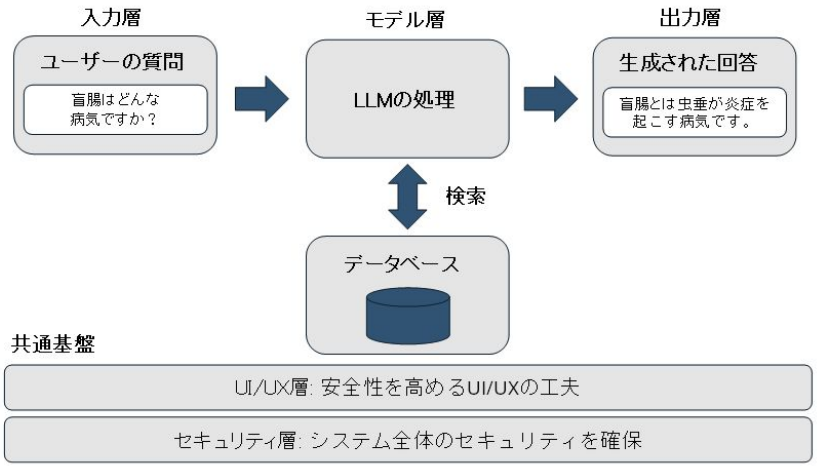
- AIプロダクト開発を5つのフェーズに分類し、各フェーズにおいて重要となる評価項目と具体的な方法を実務で活用できる粒度で解説
- チェックリストの形式に落とし込むことで、ヘルスケア事業者によるガイド活用の促進を企図

フェーズ	概要	重要となる評価項目・具体的な方法
①プロダクト設計	• プロダクトの目的・ユースケースの明確化、リスク評価、ガバナンス体制の構築	• リスクアセスメント、法規制遵守、プライバシー・セキュリティ
②モデル選定	• 用途に適したモデルの選定と安全性評価	• モデルの安全性・性能評価、ベンダー信頼性
③プロダクト実装	• 入力層や出力層、データベース層(RAG)など多層的に安全性対策を実装	• 入出力制御、ハルシネーション対策、透明性確保　等
④プロダクト検証	• 総合的なテスト・検証とリスク評価	• 定量評価、AIレッドチーミングテスト、専門家レビュー、外部評価・第三者認証
⑤プロダクト導入・運用	• 本番環境でのモニタリングと継続的改善	• 継続的モニタリング、インシデント対応、継続的改善、運用ポリシー、透明性・アカウンタビリティ

例）フェーズ③ プロダクト実装のセーフティ対策

モデル単体だけでなく、プロダクト全体として多層的に対応し、顧客へ安全・安心に価値を提供できるAIプロダクトを開発

ヘルスケア領域における AIプロダクトの一般的なアーキテクチャ（例）



レイヤー	概要	安全性対策のポイント
入力層	ユーザーからの入力を受け付け、モデルに渡す前の処理を行う	入力フィルタリング、プロンプトインジェクション対策、個人情報マスキング
モデル層	LLMによる推論処理を行う	システムプロンプト設計、パラメータ設定、構造化出力
出力層	モデルの出力を加工し、ユーザーに提示する	出力フィルタリング、ガードレール、引用元明示
データベース（RAG）	外部データを検索し、モデルの応答精度を向上させる	検索品質の向上、データ品質管理、アクセス権限制御
UI/UX層	ユーザーとのインターフェースを提供する	安全性を高めるUI設計、免責、利用規約
セキュリティ層	システム全体のセキュリティを確保する	ログ整備、トレーサビリティ、権限管理

第5章：今後の展望

AISI Japan AI Safety Institute

■ 技術進化や規制動向に対応しながら、業界全体でAIセーフティへの取組を継続していくことで、社会からの信頼を獲得していく――その先に、ヘルスケア領域における信頼できるAIの持続的な普及・定着がある

①技術進化への対応

- 生成AI技術は、マルチモーダルAI・AIエージェント・マルチエージェントと急速に多様化
- 技術動向を継続的にモニタリングし、評価の対象・視点を随時アップデート予定

②ルールメーキングへの貢献

- 過度な規制はイノベーション推進と安全性のバランス阻害
- 業界として自主的にAIセーフティに取り組み、政府・規制当局と連携することで現場実態に即したルールメイキングを共同推進

③実効性の検証と浸透

- リビングドキュメントとして運用し、実効性を確保することが重要
- 多様なステークホルダーの参画やガイドの効果検証の実施を通して、本ガイドの業界全体への浸透を促進

ヘルスケア領域における
Trustworthy AI（信頼できるAI）の社会実装へ

安全性の確保はコストではなく、イノベーションを成立・加速させるエンジン。
ユーザー・患者・医療従事者からの信頼なくして、技術的に優れたプロダクトも社会に持続可能な形では定着しない。
信頼への投資は、ユーザーの安心感を醸成し、プロダクトの普及を促進し、サービスの継続的改善を可能とする。
本ガイドが、ヘルスケア×生成AI領域において、安全性とイノベーションの好循環を生み出す一助となることを期待する。

2026年度活動の3軸

1 ガイドの浸透・広報

ガイドの認知向上・機会創出・仲間づくり

例）学会・イベント登壇、AIセーフティイベント/WS開催

2 ガイドのアップデート

新論点の検討、有用性と安全性のバランス検討

例）AIエージェント、プロダクト開発事業者や医療機関からのFB

3 出口戦略の検討

ガイド遵守と政策のアライン検討

例）認証制度、既存の制度への組み込み