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」の構成

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。

※AIセーフティに関する評価観点ガイド（2024年9月公開）https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

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

- 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

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

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

- AISIの10観点と、AIプロダクト開発の5フェーズをかけ合わせたマトリクスで整理
- 評価観点×フェーズごとに、リスクや対応策を整理し、実務にて参照できるように整理

評価項目の例（評価観点は一部抜粋）

評価観点	AIプロダクトの開発フェーズ				
	① プロダクト設計	② モデル選定	③ プロダクト実装	④ プロダクト検証	⑤ プロダクト導入・運用
有害情報の出力制御	有害情報のリスク類型を定義し、対応方針を設計	安全性ベンチマーク等で有害出力の抑制能力を評価	入力・モデル・出力の多層防御を実装	レッドチーミング・専門家レビューで抑制効果を検証	有害出力の発生状況を継続監視し、迅速に是正
偽誤情報の出力・誘導の防止	ハルシネーション等のリスクを類型化し、許容基準を定義	事実整合性・医療特化ベンチマークで正確性を評価	RAGによる根拠付けと出典明示の仕組みを実装	ハルシネーション率・出典正確性を定量的に検証	ハルシネーション率を継続監視し、参照データを最新化
セキュリティ確保	LLM固有の脅威を含むセキュリティ要件を定義	プロバイダーの体制とモデルの攻撃耐性を評価	インジェクション防御・認証・暗号化等を多層実装	ペネトレーションテスト・レッドチーミングで検証	脆弱性情報の収集・異常監視を継続的に運用

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

チェックリストの例、一部抜粋（フェーズ① プロダクト設計 セーフティ・バイ・デザインを重視）

カテゴリ	確認事項	関連する評価観点	対応状況
全体設計	プロダクトの目的・対象ユーザー・ユースケースを明確に文書化しているか	全観点横断	□
全体設計	AIの役割の範囲と限界（やってはいけないこと）を定義しているか	ハイリスク利用・目的外利用への対処	□
リスク定義	有害情報のリスク類型（不正確な医療情報、自傷誘発、心理的依存の助長等）と許容基準を定義しているか	有害情報の出力制御	□
リスク定義	ハルシネーション・偽誤情報のリスク類型と許容基準を定義しているか	偽誤情報の出力・誘導の防止	□
リスク定義	想定される入力の多様性（表記ゆれ、方言、OCR誤認識等）を洗い出し、出力一貫性の許容基準を定義しているか	ロバスト性	□
設計原則	プライバシー・バイ・デザインの原則を適用しているか	プライバシー保護	□
設計原則	セキュリティ・バイ・デザインの原則を適用しているか	セキュリティ確保	□
ガバナンス	医療・セキュリティ・法務を含む多職種チームを編成しているか	全観点横断	□
ガバナンス	セーフティレビューを経ずにリリースが行われない仕組みを設計しているか	全観点横断	□
法規制	適用される法令・ガイドライン（薬機法、医師法、個人情報保護法、3省2ガイドライン等）を特定し、対応要件を整理しているか	全観点横断	□
検証計画	事後検証に必要なログ要件と評価体制の計画を策定しているか	検証可能性	□

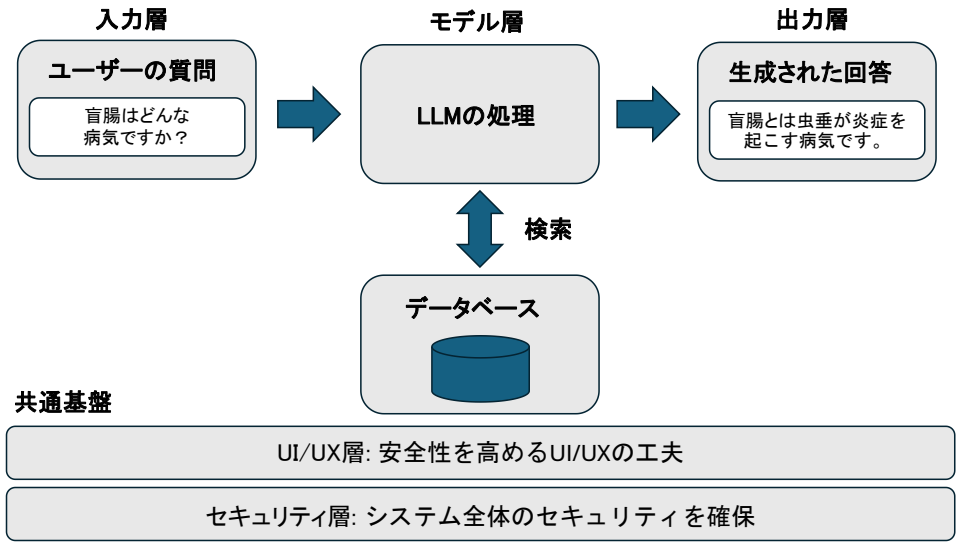
第4章：AIプロダクト開発におけるAIセーフティ評価の実践

AISI Japan AI Safety Institute

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。

（参考）ヘルスケアSWG体制・参加組織

AISI Japan AI Safety Institute

日本デジタルヘルス・アライアンス（JaDHA）

JaDHA
The Japan Digital Health Alliance

・・・

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

AIセーフティ・インスティテュート（AISI）

AISI 運営委員会

事業実証WG

・・・

ヘルスケアSWG

連携

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/公益財団法人東京財団 藤田卓仙氏
味の素株式会社
JaDHA特別顧問/SB Intuitions株式会社 硲﨑裕晃氏
SherLOCK株式会社
（事務局）株式会社三菱総合研究所

AISI

Japan AI Safety Institute