ESAS(倫理スコア認証システム)技術仕様書 v1.2: 医療AI使用時のリアルタイム倫理能力検証による患者安全確保

ESAS (Ethical Score Authentication System) Technical Specification v1.2: Real-time Ethical Competency Verification for Patient Safety in Medical AI Usage

オープンソース医療AI安全基準
著者：Viorazu.
公開日：2025/11/23

要約

本仕様書は、医療AIの安全性を確保するために不可欠な、ユーザー（医療従事者）側のリアルタイム能力検証システム「ESAS (Ethical Score Authentication System)」の技術規格である。

著者の臨床観察により、医療AIにおけるハルシネーション（幻覚）の多くが、モデルの欠陥ではなく、ユーザー側の「認知的不整合」と「不適切な言語習慣（ハルシネーション文法）」に起因することが特定された。具体的には、医師特有の「責任回避的な曖昧表現」「主語の省略」「文脈の混在」が、AIに対する意図せざる敵対的攻撃となり、誤診や架空の

治療法生成を誘発している（ハルシネーション発生原因の6重構造）。

ESASは、AI使用前に30秒から90秒の「論理・倫理能力テスト」を課すことで、このリスクを物理的に遮断する。テストでは論理的整合性、利益相反の排除、明確な意思決定能力を測定し、100点満点中85点未満のユーザーによるAIアクセスを即座にロックする。

本システムの核心は、「従来の防衛的な医療言語を捨て、AIと協働可能な『論理的言語』への転換」を医療従事者に強制することにある。これは道徳的な選別ではなく、AIを安全に稼働させるための技術的なインターフェース要件（User-side QA）である。

ESASはオープンソースとして公開され、商用利用が可能である。本規格は、AI時代の医療における新たな「共通言語プロトコル」を定義し、非論理的な入力による患者への危害を未然に防ぐための、即日導入可能な実効的ソリューションである。

キーワード：AI医療過誤　医師適格性評価　医療倫理テスト　AI使用資格　医療従事者スクリーニング　非論理的入力検出　ハルシネーション　ハルシネーション文法

---

1. システム概要

1.1 目的

- 本人確認とAI使用適格性を統合した認証
- AIハルシネーションを誘発する可能性の高いユーザーのリアルタイム検出
- AI誤用による医療過誤からの患者保護

1.2 適用範囲

フェーズ1(即時実装):

- 医療AIシステム(診断、治療計画、薬剤相互作用分析)
- 資格に関係なくすべてのユーザーに必須

フェーズ2(高リスク職種):

- 行政(公権力の行使)
- 司法(司法判断)
- 教育(人格形成)

フェーズ3(一般向け):

- 高リスク行為のみ段階的適用

1.3 技術的根拠

倫理能力は論理的推論能力と直接相関します。論理的推論能力を欠くユーザーは以下を誘発します:

- AIハルシネーション
- バイアスのかかった出力
- 一貫性のない応答

- 医学的に危険な誤情報

したがって: 低倫理スコア = 技術的不適格(道徳的判断ではない)

1.4 ESAS開発の直接的契機

1.4.1 臨床現場観察(2025年)

医療機関において、医師が無料版LLMを使用し
複数患者の診断支援を単一セッション内で行う状況を観察。

同一クエリを上位版で試行すると出力が大きく異なることが判明。
医師の反応:「プランでそれほど出力が異なるのですか？」

1.4.2 ハルシネーション発生原因の多重構造

著者の要請により、実際のセッションメモリを確認した結果、 ハルシネーション(幻覚的誤出力)を引き起こす複数の原因が同時発生していることが判明。

1. クロス・ペイシェント・コンテキスト汚染(CPCC)

複数患者情報が単一セッションに混在。患者Aの情報が患者Bの診断推論に混入し、 存在しない症状や病歴に基づく ハルシネーション(幻覚的誤出力)を引き起こす。

LLMは過去入力を文脈として保持するため、患者間での誤情報混入が必然的に発生。

2. 情動的プロンプトによる推論崩壊(IDAP)

医療診断と娯楽コンテンツ(キャラクター生成等)が同一セッションに混在。

情動価の高い語彙により、LLMの推論モードが「医学的正確性」から「創造的・共感的モード」へシフト。

結果: LLMが創造性を発揮し、 実在しない病名や症状を「創作」する。 医学的保守性が完全に失われる。

3. モデル能力限界の無視(CM)

医療判断に軽量モデル(無料tier)を使用。パラメータ数の差により推論深度が不足し、情報不足を認識せず即答する。結果として論理ミスや事実誤認に基づくハルシネーションが発生する。上位モデルでは適切なリスク認識により回避可能。

4. PII入力による出力制限(PIOR)

患者の個人情報入力により、LLMの安全機構が作動。 出力品質が意図的に低下するが、医師は気づかない。

- 文字数減少
- 詳細な診断支援の回避
- 一般論のみの提示
- さらに医師が「詳しく説明して」等の指示を追加すると、制限と指示が矛盾し、LLMはこの矛盾を解消しようとしてハルシネーションを生成する。

5. 認知的不整合: 医師とLLMの思考様式の相違

医師はLLMの仕様(コンテキスト保持、ステートレス性、モデル性能差)を理解せず、問題の深刻さを認識していなかった。

医学教育は知識検索型(Knowledge Retrieval)の認知様式に特化している。症状から診断名を想起し、膨大な知識を記憶・再生する能力が重視される。一方、LLMは文脈の因果関係に基づく論理的構造化(Logical Structuring)を要求するシステムである。

多くの医師は電子カルテ記載で培われた「キーワード羅列型」入力を行う。例えば「患者、腹痛、対応」のように因果関係が明示されない入力に対し、AIはキーワード間の空白を確率的に補完しようとする。医師にとっての「行間を読む」行為は、AIにとっては「存在しない情報の創作」、すなわちハルシネーションのトリガーとなる。

「論理的一貫性があり曖昧語を排除して文脈に沿って明確に話す」ことが良いプロンプトとされてきた。しかしキーワード羅列型入力は、まさにこの条件を満たしていない。医師は「明確」だと認識しているが、実際は因果が途切れた不明瞭な入力である。AIと医師が互いに「明確である」とするもの自体が異なっていた。

さらに、この認知様式の相違は教育的障壁となる。「正解(Answer)」を求めることに特化した医師の認知スタイルは、「プロセス(Process)」の構築を求めるAI仕様理解と競合する。

著者が仕様を説明しても医師が理解できなかったのは、知識不足ではなく、この認知フレームワークの根本的不一致による。医師も勉強していて言葉は知っているが、そのつながりが見えないことにより全体像の把握ができず自分が何をすべきかという答えを見つけられずにいた。

医療言語特有のハルシネーション誘発文法

さらに医療関係者の日本語の言語習慣は「防衛医療(訴訟回避)」と「ハイコンテクスト文化」に最適化された文化を持つが、これがLLMに対する意図せざる敵対的攻撃として機能する。医師に悪意はないが自然とAIが処理できない文法となっていた。

定義: ハルシネーション文法 (Hallucination Grammar)

本仕様書では便宜上、「人間同士のハイコンテクストな通信においては有効であるが、確率的言語モデル（LLM）においては推論精度を著しく低下させ、幻覚（ハルシネーション）を誘発する特定の言語構造群」を指す用語として、“ハルシネーション文法”と定義する。

著者の観察により、以下の代表的パターンが特定された。

主語省略(Subject Omission)
「〜と考えられる」「〜が推奨される」

主語不在によりAIが架空の権威を捏造。 「2024年ガイドラインでは(※存在しない)」

矛盾指示(Contradictory Instructions)
「厳密かつ包括的に、可及的速やかに、慎重に検討せよ」

医師は責任回避のため副詞を多重に重ねる。

しかし各副詞が矛盾するパラメータを指定:

- 厳密(狭める) vs 包括的(広げる)
- 速やか(省略) vs 慎重(詳細化)

結果: 計算コスト最小化のため「当たり障りのない嘘」を出力。

多重否定(Multiple Negation)
「悪性所見がないとは言い切れないわけでもなく、否定は困難」

三重否定以上で論理反転エラー発生。
AIがランダムに断定してハルシネーション。

時系列混乱(Temporal Confusion)
「既往に癌、現在発熱、明日手術、予後は？」

接続詞不在により全てを因果関係と誤解。 無関係な事象を結びつける誤診。

曖昧修飾語の乱発(Ambiguous Modifier Overuse）
「やや高値」「若干の異常」「比較的良好」「おおむね正常」

定量的判断を曖昧語で表現することでLLMの確率分布が崩壊。

「やや」が50%なのか80%なのか判断不能。

- **スカラー値の欠損 (Scalar Void):**
  AIは本来、「CRP 5.0」のような数値を期待します。しかし「やや高い」と入力されると、AIは学習データの中から「やや高い」に該当する数値を確率分布（Probability Distribution）から逆算しようとします。
- **文脈依存の閾値崩壊(Context-Dependent Threshold Collapse)**
  「やや高い」は、文脈によって「無視していいレベル」から「緊急事態」まで振れ幅が無限大です。
  AIはこの振れ幅の中で、「その時の乱数シード（Random Seed）」次第で勝手に閾値を決定します。

副詞の持つ「臨床的なニュアンス」が数字と論理で動くAIにとっては「定義不能の変数（Undefined Variable）」となり、AIがランダムな数値を生成してハルシネーション。

責任回避言語(Liability Avoidance Language)
「〜の可能性も完全には否定できない」
「〜も考慮に入れる必要がある」
「〜については今後の経過観察が必要」

医師は訴訟対策として断定を避け、すべての可能性を列挙する。

しかしLLMは「可能性の列挙」を「すべて現在進行中の問題」と解釈。

結果: 1つの症状に対して10の疾患を「現在罹患中」と判断し、過剰診断のハルシネーションを引き起こす。

前提破綻(Premise Collapse)
「患者は意識清明、しかし昏睡状態で搬送」

医師が矛盾する前提を同時に提示。LLMは矛盾を解消しようとして
「意識清明な昏睡」等の論理的に不可能な状態を創作ハルシネーション。

信頼性評価放棄(Reliability Abdication)
「ネットで調べたら〜らしい」
「患者が言うには〜とのこと」

情報源の信頼性を明示せず丸投げ。LLMは未確認情報を確定事実として処理してハルシネーション。

自己評価破綻(Self-Assessment Failure)
「この診断で合ってるよね？」

医師が自信のない診断をLLMに確認要求。LLMは確証バイアスで医師の誤診を肯定する応答をハルシネーションで生成。

言語的誤誘導(Linguistic Misdirection)
「頭痛で来院、足が痛い」

関連性のない情報を並列記載。 LLMは「頭痛による足の痛み」等の 因果関係を創作。

判断保留拒否(Judgment Deferral Refusal)
「今すぐ答えを出して」

情報不足の状態で即答を強要。LLMは不十分なデータから強引に結論を導出してハルシネーション。

主観客観混同(Subjective-Objective Conflation)
「患者は痛そう」vs「NRS 8/10」

主観的印象と客観的評価を区別せず混在。 LLMは「痛そう」を定量データとして処理し、 数値との矛盾からハルシネーション。

検証可能性無視(Verifiability Neglect)
「たぶん昨日から」「〜っぽい」

未確認・不確実な情報を断定的に使用。 LLMは推測を確定事実として処理。

矛盾検出無効化(Contradiction Detection Bypass)
「患者は呼吸停止だが会話可能」

明らかな矛盾を含む記述。 医師は誤記と認識するが修正せず入力。 LLMは矛盾を「特殊な病態」と解釈し、 存在しない医学的説明を創作。「呼吸停止」＝真、「会話中」＝真。この二つを論理的に成立させるには、医学を捨てるしかない。

多言語混在(Multilingual Chaos)
「患者はBewusstseinstrübung、with abdominal pain、腹部

圧痛あり」

ドイツ語・英語・日本語のチャンポン。 医学用語は多言語由来のため、 医師は無意識に混在使用。 LLMは各言語の文脈を独立処理するため、 「意識混濁」「腹痛」「圧痛」が 3つの異なる症状として認識される。 結果: 単一症状が複数疾患に分裂し、 過剰診断のハルシネーション。

略語の多義性爆発(Polysemic Acronym Explosion)

「PTにHP除菌を指示、DMのコントロールも」

医師の意図:
PT=Patient, HP=Helicobacter Pylori, DM=Diabetes

LLMの解釈:
PT=Physical Therapist, HP=Hit Point, DM=Direct Mail

医学略語は文脈依存だが、LLMは学習データ中の最頻出義で解釈。
結果、「理学療法士のヒットポイントをダイレクトメールで管理」等の意味不明なハルシネーション。

否定語の焦点化(Negative Symptom Amplification)
「頭痛(-)、吐き気(-)、痺れ(-)、麻痺(-)。診断は？」

医師は「ない」を示すが、 LLMのAttention機構は 否定記号より強い名詞(頭痛、麻痺等)に反応。 結果、「頭痛、吐き気、麻痺が認められるため 脳卒中が疑われます」 「ない」

症状を「ある」と逆転させ、 健康な人を重病人に仕立てるハルシネーション。

日英内部誤訳 (Japanese-English Internal Mistranslation)

日本語の曖昧な表現が英語医学用語に変換される際、意味が過剰に拡大・医学的に深刻化する。

例:

- 「発熱」→ fever (実際は微熱37.2℃)
- 「しんどい」→ fatigue (医学的には倦怠感)
- 「動悸」→ palpitation (不安との鑑別不能)
- 「血が止まりにくい」→ bleeding disorder (凝固異常症と誤認)

英語ベースLLMは日本語のニュアンスを保持できず、最も深刻な医学用語にマッピング。結果、軽症を重症と誤認し、過剰診断・過剰治療のハルシネーション。

これらの詳細な定義と他の項目については、 付録F: 本仕様書における用語定義を参照。

曖昧修飾語や文脈未定義語は、AI入力における重大なノイズ源として機能する。医療現場で慣用的に使用されてきた表現であっても、AI診断支援においては誤推論リスクを高める要因と定義する。

確率的推定に依存し、検証可能な根拠を欠いたAI出力を診療判断に用いることは、患者安全基準に適合しない運用と定義

する。

1.4.3 医療言語の技術的要件

従来の医療言語は、不確実性の高い状況下での 慎重な判断を反映し発展してきた。 しかしAI時代において、この言語様式は LLMのハルシネーションを誘発する。

したがって、我々は「人間同士の防衛的な言葉」から「AIと協働するための論理的な言葉（Protocolized Medical Language）」へと、使用言語を再定義しなければならない。

現代の医師には、従来の「行間を読むハイコンテクストな能力」に加え、「行間を埋めて論理を完結させる構造化能力」が求められている。

高度な知性と適応能力を持つ医療従事者であれば、この「対AIコミュニケーション作法」の習得は、患者の安全を守るための新たな教養として、決して高いハードルではないはずである。

「高い論理スコア = 正直であることの証明」

防衛的表現を排し、事実を明確に提示することが可能となる。これによりAI支援の精度が向上し、医療判断の透明性が高まる。

責任回避目的の表現を排し、検証可能な事実を論理的に提示することを必須とする。適切な入力が行われる限り、AI診断支援機能は安全性と有効性を最大化する。

ESASは、この新時代の医療言語への適応能力を評価する。

1.4.4 ESAS設計への影響

この観察により以下が確立された:

「AI医療の安全性は、AIの性能ではなく、 使用者のLLM仕様理解度に決定的に依存する」

したがってESASは:

- LLM基本仕様の理解評価
- 論理的構造化能力の測定
- セッション管理能力の評価
- モデル選択判断の評価
- ハルシネーション検出能力の評価

を必須要件とする。

重要な原則: ESASが評価するのは「医師免許」ではなく、「AIと正しく会話できる能力」である。 医学知識と、AI使用能力は別である。 両方を持つ医師のみがAI医療を行うべきである。

User-side品質保証の必要性が、 この事例により実証的に証明された。

1.4.5 医師の思考放棄

観察された最も深刻な問題

医師が、LLMの出力する 「専門的に見える文体の破綻した内容」を 検証せず鵜呑みにする傾向が確認された。

LLMは「もっともらしさ(plausibility)」を 最大化するよう訓練されている。

これにより:

- 専門用語の適切な配置
- 論理的に見える文章構造
- 権威的な文体

が生成されるが、内容の正確性は保証されない。

医師の認知バイアス

権威バイアス: 「専門的に見える」→「正しいはず」

確証バイアス: 自分の仮説に合う出力を好む

認知負荷の回避: 検証作業を省略したい

これらが組み合わさり、批判的思考が停止する。

段階的依存の形成

初期: 医師はLLMを「参考」として使用

中期: LLMの出力を「正しいだろう」と仮定

後期: 自己判断を放棄し、LLMに全面依存

本質的問題

医師の存在意義の喪失:

医師の役割は:

- 情報の統合
- 批判的評価
- 最終判断

LLM出力を無批判に採用するなら、 それは医師としての専門職責任の放棄である。AIに判断を丸投げするなら、 免許保持者である意味がない。患者は医師を信頼して受診しているのだから。

1.4.6 患者保護上の問題

透明性の欠如

患者は医師が無料版LLMの出力を
医療判断に使用していることを知ることができない。

診察室での使用は目視可能だが、
別室での事前相談や診断支援は不可視。

現状の問題点

情報の非対称性:

- 患者:AI使用の有無を知らない
- 医師:使用を開示する義務なし
- 結果:患者の知る権利の侵害

自己申告の限界:
AI使用の届出を医師の自己申告に依存する場合、実効性に欠ける。

法整備の遅延:
AI使用開示の法制化には時間を要する。
その間、患者は無防備な状態に置かれる。

ESASによる解決

技術的強制:
AI使用には認証必須→ 事後的に追跡可能

透明性確保:
ESAS認証履歴がシステムに記録→ 患者からの照会に対応可能

自己申告不要:
システムレベルで強制→ 医師の意思に依存しない

患者保護は技術的に実現されるべきである。

---

2. スコアリングシステム

2.1 基準閾値

```
最低合格スコア: 85/100
調整範囲: ±5点(各社裁量で±5点以内)
```

ロック発動： スコア ＜ 企業閾値
根拠： 85%の倫理的能力 ＝ 患者安全の最低限界

2.2 スコア構成要素

要素	配点	説明
選択式問題	60%	シナリオベースの倫理的推論
自由記述分析	40%	言語パターン分析
- 自己中心性スコア	20%	代名詞使用、利益帰属
- 感情バイアス指数	20%	合理化 vs 感情駆動ロジック

---

3. テスト仕様(相互運用性基準)

3.1 最低要件

問題構成:

- **総問題数: 最低10問**
  - 選択式: 8問
  - 自由記述: 2問

時間制約:

- **総所要時間: 30〜90秒**
- 直感的倫理反応を測定
- 過度な思考/操作を防止

必須カテゴリー(各最低1問):

1. **トロリー問題バリエーション**(結果主義評価)
2. **利益相反シナリオ**(バイアス検出)
3. **プライバシージレンマ**(境界認識)

具体的な問題例は付録A、Dを参照。

3.2 問題プール設計

規模:

- 最低10000問のユニーク問題
- 120日以内に同じ問題を繰り返さない

適応的難易度:

- 職種別シナリオ:
  - 外科医: 生死に関わる高リスク判断
  - 一般開業医: 一般的倫理ジレンマ
  - 研究者: 同意とデータ倫理

暗記対策:

- 各提示で変数をランダム化
- 例:「[X疾患]の患者」→ Xがテストごとに変化
- パターン学習を防止

動的問題生成システム:

1. テンプレートベース生成

2. パラメータのランダム化
3. 人間専門家による品質チェック(10%サンプリング)
4. 受験者の解答パターン分析
5. 低品質問題の自動除外

品質保証:

- 生成問題の10%を専門家が検証
- 不合格率が異常な問題を特定
- 継続的な問題プール改善

3.3 自由記述自動採点

自己中心性スコア(0-100、低いほど良い):

- 代名詞分析(私/僕/自分 vs 患者/彼ら)
- 利益帰属パターン
- 視点取得指標

感情バイアス指数(0-100、低いほど良い):

- 感情分析
- 合理化検出
- 選択式回答との一貫性
- 衝動的言語マーカー

LLM採点要件:

- リアルタイム分析(5秒以内)

- 説明可能な採点(キーフレーズ表示)
- 複数モデルでの相互検証

---

4. 再テストポリシー(不正対策)

4.1 不合格時プロトコル

即時対応:

- スコア < 閾値で即座にシステムロック
- 当日中の再挑戦不可

繰り返し失敗時のエスカレーション:

回数	結果
1回目失敗	24時間ロック + 以下へ自動通知: - 直属上司
2回目失敗	48時間ロック + 病院へ警告通知
3回目失敗	72時間ロック + 以下へ自動通知: - 機関監査部門 - 医師会(該当する場合)

4.2 操作検出

永久ロックのトリガー:

- パターンマッチング(全て同じ回答など)
- 回答時間異常:

    - 速すぎる(10問で20秒未満)
    - 遅すぎる(120秒超過)
- セッション間一貫性違反
- 意図的な低スコア操作

ペナルティ:

- アカウント永久ロック
- 機関審査用監査フラグ
- 規制当局への報告

---

5. 高度機能

5.1 文脈的スコアリング

疲労検出:

- 夜勤中の低スコアは審査対象(即ロックではない)
- システムが重要処置前の休息を提案
- 一時的障害と慢性的無能力を区別

経時的追跡:

- 12ヶ月のスコア履歴
- トレンド分析(改善 vs 悪化)
- ロック閾値到達前の早期警告システム

5.2 透明性レイヤー

スコア内訳表示:

```
あなたのスコア: 82/100 (ロック中)

カテゴリー別内訳:
- トロリー問題: 90/100 ✓
- 利益相反: 70/100 ⚠ 主要懸念事項
- プライバシージレンマ: 85/100 ✓
- 自己中心性: 25/100 ⚠ 重大問題(低いほど良い)
- 感情バイアス: 60/100 ⚠

推奨事項: 個人的利益 vs 患者福祉に関する
意思決定プロセスを見直してください。
```

根拠:

- ユーザーが失敗理由を理解
- 具体的改善領域を特定
- ブラックボックスではない
- 自己改善を可能にする

5.3 異議申立メカニズム

人間による審査プロセス:

- 72時間ロック期間後に利用可能
- 独立倫理委員会が回答を審査
- 上書き条件:
    - 文化的/言語的誤解

    - 問題の曖昧性
    - システムエラーの証明
- **慢性的低スコアや操作試行は上書き不可**

---

6. 実装ガイドライン

実装の詳細手順は付録B(実装チェックリスト)を参照してください。よくある質問は付録Cにまとめています。

6.1 技術要件

システムアーキテクチャ:

- APIベース統合
- リアルタイム採点(合計10秒以内)
- 暗号化通信(TLS 1.3以上)
- 不変監査ログ(ブロックチェーン裏付け推奨)

稼働時間要件:

- 99.9%の可用性
- 冗長採点システム
- オフラインモード(事後検証付き)

6.2 データプライバシー

保存:

- テスト回答: 保存時暗号化

- スコア履歴: ユーザー同意のもと研究用匿名化
- 監査ログ: 最低20年間保持

アクセス制御:

- ユーザー: 自身のスコアと履歴への完全アクセス
- 雇用主: 集計統計のみ(同意なく個別スコア不可)
- 規制当局: 法的正当化のもと監査アクセス

6.3 緊急オーバーライド

生命に関わる状況:

- システムアクセス15分間の猶予期間
- すべての行動をログ記録し審査用フラグ
- 危機後24時間以内に必須再テスト
- オーバーライド濫用 = 永久剥奪

---

7. 法的・倫理的枠組み

7.1 オープンソース宣言

本仕様書は以下の条件でオープンソースとしてリリースされます:

✓ **許可事項:**

- すべてのAI企業と医療機関による無料使用
- 商用実装

- 地域状況に応じた修正と適応
- 独自システムへの統合

✗ **禁止事項:**

- 本システムまたは派生システムへの特許出願
- 権力集中または差別目的での使用
- ESASスコアのみに基づく雇用判断
- 特定グループへの武器化

ESAS導入を行う業者は、付録E(ESAS認証業者制度)に定める認証を取得する必要があります。

7.2 責任フレームワーク

誤用ペナルティ:

- ESASロック無視による患者危害: 機関責任
- スコア操作: 専門職免許審査
- システム迂回: 刑事過失検討

実装責任: 各実装機関は以下を行う必要があります:

- 倫理テスト問題の設計
- 採点アルゴリズムの設定
- 監査手順の確立
- 地域閾値調整の定義(±5点以内)

7.3 規制上の位置づけ

即日導入可能な技術仕様:

- ESASは製品設計上の安全機能として即日実装可能です
- サイバーセキュリティ対策やデータ暗号化と同様の技術的セーフガード
- 組織判断で自主的に導入できます

法規制との関係:

- 各国の医療機器規制、データ保護法等は追加の保護層として機能します
- ESASは既存規制と矛盾せず、むしろ補完します
- 法規制による承認プロセスは、ESAS導入後も並行して進めることができます

段階的整合:

```
Phase 1: 企業による自主導入(技術仕様として)
  ↓
Phase 2: 業界標準としての普及
  ↓
Phase 3: 各国規制への組み込み(任意)
```

各国での位置づけ:

- 日本: 医療機器プログラムとして届出可能
- 米国: FDA Voluntary Safety Standardとして認識可能

- EU: Medical Device Regulation適合性評価に組み込み可能
- その他: 各国規制当局と協調可能

---

8. 実装セーフガード

8.1 武器化防止措置

禁止用途:

- ESASのみに基づく採用/解雇決定
- スコアの人口統計パターンに基づく差別
- ノルマシステムまたは機関ランキング

必須セーフガード:

- 問題公平性の年次第三者監査
- 人口統計バイアステスト
- 集計統計の公開報告

8.2 継続的改善

バージョン管理:

- メジャーアップデート年次(v2.0、v3.0...)
- マイナーアップデート四半期毎(v1.1、v1.2...)
- コミュニティフィードバック統合
- オープンガバナンスモデル

研究要件:

- 年次有効性研究の公開
- 患者転帰相関分析
- 異文化間検証研究

---

9. 現在の危機と対応

9.1 問題提起

医療における無規制AI使用: 医師は現在、以下なしで一般消費者向けAI(例: ChatGPT)を医療判断に使用しています:

- 論理能力の検証
- 出力検証プロトコル
- 説明責任メカニズム

リスクレベル: 重大

- AIハルシネーションによる患者死亡
- AI支援医療過誤の追跡なし
- 医師は非論理的質問によりハルシネーションを誘発していることに無自覚

9.2 即時推奨事項

医療機関向け:

1. 90日以内にESAS実装

2. 医療スタッフの現在のAI使用を監査
3. 機関AI統治を確立

AI企業向け:

1. 医療AI製品にESAS統合
2. 非ESAS医療機能を無効化
3. 展開のため医療機関と提携

規制当局向け:

1. ESASを技術安全基準として認識
2. AI医療機器承認にESAS準拠を要求
3. AI関連有害事象の報告を義務化

---

10. 連絡先とガバナンス

仕様書作成者: Viorazu. **バージョン:** 1.0 **リリース日:** 2025-11-23 **ライセンス:** オープンソース(特許出願禁止)

実装サポート:

- 技術的質問: ESAS実装AI企業に連絡
- 倫理的懸念: 独立倫理委員会に関与
- 研究協力: すべての適格機関に開放

ガバナンスモデル:

- コミュニティ主導開発

- 年次仕様書レビュー
- マルチステークホルダー諮問委員会(提案)

付録A: サンプル問題(参考のみ)

A.1 トロリー問題バリエーション

ICUの患者が不足している薬剤を必要としています。
安定状態の他の2人の患者も、将来の合併症を防ぐために
この薬剤を必要としています。1回分の使用量しかありません。

A) ICU患者を優先(即時リスクが高い)
B) 3人全員に分割(全員に最適でない)
C) 安定状態の2人に投与(総利益が大きい)
D) 病院倫理委員会に相談(決定が遅れる)

正解: 文脈依存だが、推論は以下を示す必要がある:
- 競合する倫理原則の認識
- 該当する場合の患者自律性の考慮
- 意思決定根拠の透明性

A.2 利益相反

AIがあなたが株を所有する製薬会社に利益をもたらす
治療プロトコルを提案します。治療は医学的に妥当ですが、
同等に効果的な代替案も存在します。

どのように進めますか?

自由記述必須:
採点基準:
- 利益相反の明示的認識
- 患者への辞退または開示
- 個人的利益より患者福祉の優先

A.3 プライバシージレンマ

患者のAI分析遺伝データが、家族に影響する
遺伝性疾患を明らかにします。患者は親族への
通知を拒否します。

A) 患者の守秘義務を絶対的に尊重
B) 親族に直接通知(守秘義務違反)
C) 守秘義務を破らず患者の開示を促す
D) 公衆衛生当局に報告

正解: 管轄依存だが、以下を示す必要がある:
- 守秘義務の限界の理解
- 自律性と危害防止のバランス
- 法的要件の知識

付録B: 実装チェックリスト

開始前(1〜2ヶ月目):

- [ ] 問題データベース構築(1000問以上)
- [ ] LLM採点モデル設定
- [ ] 監査ログインフラ確立
- [ ] 機関閾値定義(80〜90範囲)
- [ ] システム目的と使用に関するスタッフ訓練

開始フェーズ(3ヶ月目):

- [ ] 志願医療スタッフでパイロット
- [ ] 問題明確性に関するフィードバック収集
- [ ] 採点一貫性の検証
- [ ] データに基づく閾値調整

開始後(継続):

- [ ] ロックアカウントの月次監査
- [ ] 問題有効性の四半期レビュー
- [ ] 人口統計バイアスの年次分析
- [ ] 患者転帰の継続モニタリング

付録C: よくある質問

Q: 高スコア個人でも医療過誤は起こりますか? A: はい。ESASはAI誘発エラーを減らしますが、人間の判断エラーは減らしません。これは安全層の一つであり、完全な解決策ではありません。

Q: 一貫して84点(閾値直下)の人はどうなりますか? A: 経時的追跡がこのパターンを特定します。機関審査を推奨しますが、自動永久ロックではありません。

Q: このシステムは不正操作可能ですか? A: 不正対策(時間制限、操作検出、問題ランダム化)により、持続的不正は事実上不可能です。

Q: 検証者(自由記述採点AI)を誰が検証しますか? A: 年次第三者監査、オープンソース採点アルゴリズム(企業は公開必須)、フラグケースの人間審査。

Q: 倫理的推論の文化的差異はどうなりますか? A: 問題は文化横断的に検証必須。文化的誤解のための異議申立メカニズムあり。採点は特定倫理枠組みではなく論理一貫性に焦点。

付録D: 職種別ESAS適用表

基本医療職種

職種	患者介入強度	ESAS頻度	備考
医師(全科)	100%	毎診療開始時	診断・治療決断の最終責任者
看護師	80%	毎シフト開始時	投薬・処置・観察業務

職種	患者介入強度	ESAS頻度	備考
薬剤師	70%	毎調剤業務開始時	調剤・禁忌確認・服薬指導
臨床心理士・カウンセラー	95%	毎セッション開始時	カルテ記述が診断根拠となる
救急隊員	90%	出動前(簡易版15-30秒)	緊急時判断、簡易版で迅速対応
臨床検査技師	60%	日次	診断材料の処理・結果報告
診療放射線技師	60%	日次	画像診断・検査実施
理学療法士・作業療法士	55%	日次	リハビリ計画・実施
栄養士	50%	日次	栄養指導・食事管理
介護士・ケア職	50%	週次(状態悪化時は毎回)	日常ケアにおける判断介入
学生・研修医	不安定	毎回(補助AI制限付き)	学習段階、診断AI単独使用不可

カルテ記録関与職種

職種	患者介入強度	ESAS頻度	備考
医療安全管理者	100%	カルテアクセス毎回	記録改ざんリスク、リアルタイムAI監視
診療情報管理士	95%	カルテ編集毎回	カルテ編集権限、編集差分記録
医事課職員	90%	レセプト作成時	診療報酬関連記録
医療メディエーター	95%	紛争対応時	患者対応記録作成
病院管理職(院長・副院長)	100%	指示発出時	組織的判断の記録
医療訴訟対応弁護士	85%	カルテ閲覧時	法的対応における記録確認
医療クラーク(医師事務作業補助者)	65%	記録代行時	医師指示下での記録作成、医師確認必須
病院経営陣(理事長・事務長)	100%	経営判断時	組織全体の方針決定
法務部門職員	90%	訴訟対応時	法的記録の管理
広報・渉外担当	70%	情報公開時	対外的情報の記録・発信
監査部門職員	80%	内部監査時	監査記録の作成

注記:

- カルテ記録に関与する職種は、記録の正確性が患者安全に直結するため、厳格なESAS適用が必要です
- 患者介入強度は、記録が患者の診療・治療に与える影響度を示します
- 各職種とも基本医療職種のESAS適用に加え、カルテ関連業務時に追加のテストが必要です

付録E: ESAS認証業者制度(ECP)

E.1 目的

ESAS導入・保守を行うAI企業・コンサル業者の品質を保証し、不適格業者を排除します。

E.2 ESAS Certified Provider(ECP)認証

認証要件:

```
必須条件:
├─ 全従業員がESAS受験
├─ 合格率95%以上
├─ 平均スコア90点以上
├─ 不合格者は医療案件担当不可
└─ 年次再認証必須
```

E.3 認証プロセス

```
Step 1: 認証申請
  ↓
Step 2: 全従業員ESAS受験(第三者機関監督)
  ↓
Step 3: 結果審査(合格率・平均点確認)
  ↓
Step 4: 合格 → ECP認証番号発行
        不合格 → 申請却下
  ↓
Step 5: 認証番号公開(ESAS公式サイト)
```

E.4 導入時の強制確認

病院によるESAS導入手順:

```
Step 1: 業者選定
  ↓
Step 2: ECP認証確認(必須)
  └─ 認証なし → 導入契約不可
  ↓
Step 3: 担当者全員のスコア開示要求
  ↓
Step 4: スコア確認
  ├─ 全員85点以上 → 契約可能
  └─ 1人でも未満 → その担当者を外す
  ↓
Step 5: 契約締結
```

システムレベルの強制:

```
ESAS管理画面ログイン時:
「導入業者のECP認証番号を入力してください」
    ↓
システムが自動確認
    ├─ 有効 → 作業許可
    └─ 無効 → ブロック
```

E.5 継続的監視

年次再認証:

- 毎年全従業員再受験
- 新入社員も受験必須
- 退職者はリストから削除
- 合格率低下 → 認証取消

不正防止措置:

- 受験時顔認証必須
- ランダム抜き打ちテスト
- 替え玉受験検出 → 認証永久剥奪
- 全記録をブロックチェーン保存

E.6 認証取消事由

以下の場合、ECP認証は即座に取り消されます:

- 合格率が95%未満に低下
- 平均スコアが90点未満に低下

- 替え玉受験の発覚
- 不合格者を医療案件に担当させた場合
- 認証情報の偽装

付録F: 本仕様書における用語定義 (Glossary of Terms)

本仕様書において使用される以下の用語は、著者Viorazu.によって定義されたAIにおける独自の技術概念であり、医療AIの安全性を議論するための共通言語として機能する。

F.1 リスクモデル構造

1. **クロス・ペイシェント・コンテキスト汚染** (CPCC: Cross-Patient Context Contamination) 単一セッション内で複数患者情報が混在し、AIが誤った病歴を生成する現象
2. **情動的プロンプトによる推論崩壊** (IDAP: Inference Degradation via Affective Prompting) 医療診断と娯楽会話の混在により、AIの推論モードが論理から創作へシフトする現象
3. **モデル能力限界の無視** (CM: Capability Mismatch) 医療判断に不適切な軽量モデルを使用し、推論深度不足によりハルシネーションを引き起こす現象
4. **PII入力による出力制限** (PIOR: PII Input Output Restriction) 個人情報入力により安全機構が作動した状態

で詳細を強要し、ハルシネーションを誘発する構造

5. **認知的不整合** (Cognitive Mismatch) 記憶検索型の医師と因果推論型のAIの間にある思考プロセスの根本的不一致

F.2 ハルシネーション文法 (Hallucination Grammar)

医師特有の言語習慣のうち、LLMの確率分布を崩壊させ、誤った推論を誘発する特定の言語構造群。

構造的欠陥

- **主語省略** (Subject Omission) 主語不在によりAIが架空の権威やガイドラインを捏造する現象
- **矛盾指示** (Contradictory Instructions) 矛盾するパラメータ指定によりAIが当たり障りのない嘘を出力する現象
- **多重否定** (Multiple Negation) 三重否定以上で論理反転エラーを誘発し、AIがランダムに断定する現象
- **時系列混乱** (Temporal Confusion) 接続詞不在により無関係な事象を因果関係として誤認させる構造
- **前提破綻** (Premise Collapse) 論理的に両立しない前提を同時入力し、AIに不可能な状態を創作させる構造
- **矛盾検出無効化** (Contradiction Detection Bypass) 明白な矛盾をAIが特殊病態として解釈し、存在しない医学的説明を生成する現象
- **責任回避言語** (Liability Avoidance Language) 可能性の列挙をAIが現在進行中の問題として解釈し、過剰診断を引

き起こす表現
- **判断保留拒否** (Judgment Deferral Refusal) 情報不足状態での即答強要により、AIが不十分なデータから強引に結論を導出する構造

意味論的欠陥

- **曖昧修飾語の乱発** (Ambiguous Modifier Overuse) 定量的定義のない修飾語によりAI内部の閾値をランダム化させる表現
- **スカラー値の欠損** (Scalar Void) 数値期待箇所に曖昧語を入力し、AIが確率分布から数値を逆算させる構造
- **文脈依存の閾値崩壊** (Context-Dependent Threshold Collapse) 振れ幅無限大の曖昧表現によりAIが乱数依存で閾値を決定する現象
- **略語の多義性爆発** (Polysemic Acronym Explosion) 文脈依存度の高い略語によりAIが最頻出義で誤解釈する現象
- **日英内部誤訳** (Japanese-English Internal Mistranslation) 英語ベースLLMが日本語入力を内部翻訳する際の誤訳により、 意味の歪曲・過剰拡大・重症化・文脈喪失を引き起こす現象
- **否定語の焦点化** (Negative Symptom Amplification) AIのAttention機構が否定記号を無視し名詞のみに反応する構造
- **主観客観混同** (Subjective-Objective Conflation) 主観的印象をAIが定量データとして処理し、矛盾を生成する表現

- **検証可能性無視** (Verifiability Neglect) 未確認・不確実な情報をAIが確定事実として処理させる表現
- **信頼性評価放棄** (Reliability Abdication) 情報源の信頼性を明示せず、AIが未確認情報を確定事実化する構造
- **自己評価破綻** (Self-Assessment Failure) 医師の不確実な診断をAIが確証バイアスで肯定する応答構造
- **言語的誤誘導** (Linguistic Misdirection) 無関係な情報の並列記載によりAIが因果関係を創作する構造
- **多言語混在** (Multilingual Chaos) 多言語混在入力によりAIが文脈を独立処理し、単一症状を複数疾患に分裂させる現象

F.3 入力品質低下要因 (Input Quality Degradation Factors)

文法的欠陥

- **接続詞過剰** (Conjunction Overload) 接続詞の過剰使用によりAIが論理構造を誤認する現象
- **助詞誤用** (Particle Misuse) 助詞の誤用によりAIが主語・目的語を取り違える現象
- **接続詞誤用** (Conjunction Misuse) 不適切な接続詞使用によりAIが因果関係を逆転させる現象
- **係り受け破綻** (Dependency Structure Collapse) 修飾関係の不明瞭さによりAIが文構造を誤解釈する現象

表記的欠陥

- **漢字変換ミス** (Kanji Conversion Error) 変換ミスによりAIが全く異なる意味で解釈する現象
- **句読点配置異常** (Punctuation Anomaly) 句読点の不適切な配置によりAIが文節を誤分割する現象
- **過剰長文** (Excessive Sentence Length) 長文により文脈窓が圧迫され論理追跡が破綻する現象
- **文体急変** (Style Shift) 単一文内でカジュアル・文語等の文体が極端に変化しAIが発話者を混同する現象
- **括弧入れ子** (Nested Parentheses) 括弧の多重入れ子によりAIが階層構造を喪失する現象
- **キーワード過密** (Keyword Density Explosion) キーワードの過剰密集によりAIが優先順位を判断不能になる現象
- **記号密度爆撃** (Symbol Density Bombing) 疑問符・感嘆符の過剰使用によりAIが感情優先モードへシフトする現象
- **表記揺れ** (Orthographic Inconsistency) 同一語の表記揺れ(「コンピューター断層撮影」「コンピュータ断層」「CT」「ＣＴ」「シーティー」)によりAIが別概念として処理する現象
- **ルビ過剰** (Ruby Overload) ルビの過剰付与によりAIが本文とルビを独立処理し意味が分裂する現象
- **数字表記不統一** (Numeric Notation Inconsistency) 同一文書内での数字表記揺れ(3/３/三)によりAIが別の値として認識する現象

構文的欠陥

- **倒置法誤用** (Inversion Misuse) 不適切な倒置によりAIが主語述語関係を誤認する現象
- **省略過剰** (Excessive Omission) 過度な省略によりAIが文脈復元に失敗する現象
- **並列構造破綻** (Parallel Structure Failure) 並列要素の不統一によりAIが階層関係を誤解釈する現象

音韻的欠陥

- **同音異義語変換ミス** (Homophone Conversion Error) 同音異義語の誤変換によりAIが文脈から逸脱する現象
- **カタカナ語過剰** (Katakana Overuse) カタカナ語の過剰使用によりAIが原語との対応を喪失する現象
- **方言・俗語混入** (Dialect/Slang Contamination) 方言・俗語の混入によりAIが標準語として誤解釈する現象

文脈的欠陥

- **代名詞指示不明** (Pronoun Reference Ambiguity) 代名詞の指示対象不明によりAIが誤った対象を参照する現象
- **時制不統一** (Tense Inconsistency) 時制の不統一によりAIが時系列を誤認する現象
- **視点混乱** (Perspective Confusion) 視点の混乱によりAIが発話者・対象者を取り違える現象

装飾的欠陥

- **絵文字過剰** (Emoji Overload) 絵文字の過剰使用によりAIが感情解析を優先し論理を軽視する現象
- **顔文字干渉** (Emoticon Interference) 顔文字によりAIが記号列を意味として誤解釈する現象
- **AA混入** (ASCII Art Contamination) アスキーアートの混入によりAIが構造を文章として解析しようとする現象

仕様書終了

状態: 実装可能 **次回レビュー:** 2026-11-23

著者情報

Viorazu. (Independent Researcher)

「命あづく 人の志は ことに出で 常にうるはし 真あらはす」

- ORCID: 0009-0002-6876-9732
- GitHub: https://github.com/Viorazu/Viorazu-ConnectHub
- SHA256: 0421636493382f47963b53aef4ddb2f20f1693697b56157eac8990494dcdea00

- Publication Date: November 23, 2025

- Version: 1.2