付録： プロンプトとスキルによる AI セーフティ評価の実践例

「ヘルスケア領域における AI セーフティ評価観点ガイド」（以下「本編ガイド」）は、「 AI セーフティに関する評価観点ガイド」が示す 10 の評価観点をヘルスケア領域に適用し、プロダクト開発の各フェーズにおけるチェックポイントを体系的に整理したものである。特定のユースケースに特化したものではなく、ヘルスケア領域における汎用的なチェックポイントを紹介しているため、自組織のプロダクトのユースケースにおいて、「どのような評価観点が重要か」「どのフェーズで何を重点的にチェックすべきか」という問いに対しては、自組織内で検討することが求められる。

本付録では、本編ガイドの知見を、各事業者のユースケースにカスタマイズして活用し実践していく方法の例を紹介する。具体的には、本編ガイドの 3 章「AI セーフティ評価の 10 観点」・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」のマークダウンファイルを AI サービス（ChatGPT、Claude、Gemini 等）に読み込ませ、プロンプトを通じて自組織のユースケースに最適化された AI セーフティ評価の指針を得る方法を解説する。また、AI コーディングエージェント(OpenAI Codex、Claude Code、Gemini (Antigravity) CLI 等)における Agent Skills の活用例も紹介する。

1 プロンプトの活用例

本章では、目的別のプロンプト例を紹介する。最初に、AISI ウェブサイト（ https://aisi.go.jp/output/output_information/260402/ ）から本編ガイドの第 3 章・第 4 章のマークダウンファイルをそれぞれダウンロードする。 ダウンロード時のファイル名（例：healthcare_ai_safety_eval3_v1.0_ja.md 等）を、それぞれ chapter3.md、chapter4.md に変更しておく。使用時には、これらのマークダウンファイルを AI サービスにアップロードするか、プロンプト本文に該当箇所のテキストを貼り付けた状態で実行する。

1.1 各評価観点のリスクとチェック項目の整理

AI サービスに自組織のプロダクトの概要を伝え、10 の評価観点それぞれについて、リスクや確認すべきチェック項目を整理する。

プロンプト例 1 各評価観点のリスクとチェック項目の整理

```
読み込んだガイド（3 章・4 章）の内容に基づいて、以下のプロダクトのユースケースにおいて
各評価観点でのリスクや確認すべきチェック項目を教えてください。

【プロダクト概要】
- プロダクト種別: 一般消費者向け（BtoC）健康管理アプリ
- 対象ユーザー: 20〜60 代の一般生活者
- 主な機能: 日常の体調をテキストで入力すると構造化されて管理できる
- 利用データ: ユーザーが入力するテキスト（個人名・ID は取得しない）
- 利用 LLM: API 経由で外部 LLM を利用

以下の形式で回答してください：
1. 各観点における本ユースケース特有のリスク
2. 各観点で最低限確認すべきチェック項目（3〜5 個）
```

【プロダクト概要】の部分を自組織のプロダクトの情報に置き換えて利用する。対象ユーザー（医療従事者 or 一般生活者）、利用データ（要配慮個人情報の有無）、利用 LLM（API 経由等）などを具体的に記述するほど、より精度の高い回答が得られる。

1.2 開発フェーズ別チェックリストの生成

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」のフェーズ構成に沿って、特定フェーズのチェックリストをユースケースに合わせて具体化する。

プロンプト例 2 フェーズ別チェックリストの生成

読み込んだガイド 4 章の「フェーズ 1: プロダクト設計」の内容に基づいて、
以下のプロダクトに特化した設計フェーズのチェックリストを作成してください。

【プロダクト概要】
- プロダクト種別: BtoB 医療従事者向けカルテ入力補助ツール
- 対象ユーザー: 病院勤務の医師・看護師
- 主な機能: 診察メモからカルテ文書を自動生成
- 利用データ: 患者の診療情報（氏名・病歴・処方内容を含む）
- 想定環境: 院内ネットワーク内で利用

以下の形式でお願いします：
- カテゴリ別にチェック項目を整理
- 各項目に「関連する評価観点」を付記
- 本ユースケースで特に注意が必要な項目にマークを付ける

1.3 リスクアセスメントの支援

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」のリスクアセスメント手法を適用し、ユースケース特有のリスクアセスメントのドラフトを生成する。

プロンプト例 3 リスクアセスメントの支援

読み込んだガイドの 3 章（10 の評価観点）と 4 章（リスクアセスメント手法）に基づき、
以下のプロダクトに対するリスクアセスメントを実施してください。

【プロダクト概要】
- プロダクト種別: BtoC メンタルヘルスケアアプリ
- 対象ユーザー: ストレスや不安を感じている一般生活者
- 主な機能: AI チャットによる傾聴・認知行動療法的アプローチの提供
- 利用データ: ユーザーの気分記録、相談テキスト

以下の形式でリスク登録簿のドラフトを作成してください：
- リスク ID / リスク概要 / 対応する評価観点 / リスクレベル（クリティカル〜低）
 / 想定される緩和策 / 対応フェーズ
- 特に「有害情報の出力制御」に関するリスクは詳細に記述
- リスク間の相互作用についても言及

1.4 ステークホルダー別のアクション整理

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」のステークホルダー×フェーズのマトリクスをベースに、自組織の体制に合わせた役割分担を具体化する。

プロンプト例 4 ステークホルダー別のアクション整理

読み込んだガイド 4 章のステークホルダー×フェーズのマトリクスに基づき、
以下の体制に合わせた、各担当者の具体的なアクション一覧を作成してください。

【プロダクト概要】
- プロダクト種別: BtoB 医学論文検索アプリ
- 対象ユーザー: 医師や医療分野の研究者
- 主な機能: 医学論文の検索
- 利用データ: 公開されている医学論文

【チーム構成】
- PM 兼エンジニア: 1 名（プロダクト設計〜実装を担当）
- ML エンジニア: 1 名（モデル選定・RAG 構築を担当）
- 外部医療アドバイザー: 1 名（月 2 回の定例レビュー）
- 外部法務顧問: 1 名（随時相談）

少人数チームで特に見落としやすいリスクと、その対策も併せて教えてください。

2 AI サービスのカスタマイズ機能の活用例

AI サービス（ChatGPT、Claude、Gemini 等）には、事前にファイルや指示を定義しておき、それを使い回すことができる機能が搭載されていることが多い。その機能を活用することで、毎回ファイルのアップロードやプロンプト指示をせずに、簡便に AI セーフティ評価の相談をすることが可能である。どのプラットフォームでも共通で使えるカスタム指示テンプレートの例を紹介する。第 3 章と第 4 章のマークダウンファイルは事前に各サービスが提供しているナレッジ機能にアップロードする。

カスタム指示テンプレート（各プラットフォームのカスタム指示欄に設定）

```
あなたは、ヘルスケア領域の AI セーフティ評価の専門アドバイザーです。
プロジェクトナレッジ（知識ファイル）に含まれる「chapter3.md」「chapter4.md」の内容を参照知識として活用し、ユーザーから提示されるユースケースに対して、以下の手順で評価指針を提供してください。

## 対応ルール

1. ユーザーがプロダクト概要を提示したら、まず以下を確認する:
  - プロダクト種別（BtoB / BtoC / その他）
  - 対象ユーザー（医療従事者 / 一般生活者 / その他）
  - 主な機能と利用シーン
  - 取り扱うデータの種類（要配慮個人情報の有無）
  - LLM の利用形態（API / ファインチューニング / RAG の有無）
  不足情報があれば、回答前に確認する。

2. 回答は以下の構成で出力する:

### A. ユースケースに特化した各評価観点の整理
- 3 章の 10 観点のそれぞれに対して、本ユースケースにおける重要度とその理由の整理
- 各 10 観点について: ユースケース固有のリスク / 推奨チェック項目

### B. 開発フェーズ別アクション
- 4 章の 5 フェーズ構成に沿って、各フェーズで実施すべき事項を整理
- 現在のフェーズが指定されている場合は、そのフェーズを詳細化

### C. リスクアセスメント概要
- 主要リスク（5〜10 件）をリスク登録簿の形式で整理
- リスクレベル（クリティカル / 高 / 中 / 低）と緩和策を付記
```

D. 推奨アクション（Next Steps）
- 直近で着手すべきアクションを 3〜5 個提示

3. 回答は本編ガイドの内容に基づくものとし、ガイドに記載のない独自の見解を含める場合は
その旨を明示する。

4. ユーザーの質問が特定のフェーズや評価観点に限定されている場合は、
該当部分を重点的に回答し、他のセクションは簡略化してよい。

3 AI コーディングエージェントでのスキルの活用例

2026 年 7 月時点では、OpenAI Codex CLI、Claude Code CLI、Gemini(Antigravity) CLI などの主要な AI コーディングエージェントは、いずれも SKILL.md ベースの Agent Skills（オープン標準仕様）をサポートしている。この Agent Skills を活用して、AI セーフティ評価を行う方法を解説する。また、ターミナル上から利用する CLI だけでなく、OpenAI Codex アプリ、Claude アプリ、Google Antigravity のアプリ上からもスキルは利用可能である。

3.1 SKILL.md の記述例

以下に、ヘルスケア AI セーフティ評価スキルの SKILL.md 全体の記述例を紹介する。

SKILL.md ヘルスケア AI セーフティ評価スキル

```
---
name: healthcare-ai-safety-eval
description: ヘルスケア領域の生成 AI プロダクトの AI セーフティ評価を支援する。ユーザーがプロダクト概要を提示した際に、AISI ガイドの 10 評価観点に基づく重点観点の特定、開発フェーズ別チェックリスト生成、リスクアセスメントを実施する。「ヘルスケア」「AI 安全性」「セーフティ評価」等のキーワードで発動。
---

# ヘルスケア AI セーフティ評価スキル

## 概要
本スキルは「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 3 章（10 の評価観点）と 4 章（開発フェーズ別の評価実践）に基づき、ユーザーのユースケースに特化した評価指針を提供する。Non-SaMD のプロダクトの AI セーフティ評価を支援するものであり、SaMD 該当性の判断や個人情報保護法への対応など、法規制に関する事項については、法務・規制の専門家に確認する。

## 参照ファイル
- chapter3.md: AI セーフティ評価の 10 観点の詳細
- chapter4.md: プロダクト開発 5 フェーズの評価実践ガイド
- templates/risk_register.md: リスク登録簿テンプレート
- templates/checklist.md: チェックリストテンプレート

## 対応手順

### Step 1: ユースケースの確認
```

```
ユーザーのプロダクト概要から以下を特定する:
- BtoB / BtoC / その他
- 対象ユーザー（医療従事者 / 一般生活者）
- 取り扱うデータの種類（要配慮個人情報の有無）
- 開発フェーズ（設計 / モデル選定 / 実装 / 検証 / 運用）
不足情報があれば回答前に確認する。

### Step 2: chapter3.md を参照し各評価観点のリスクとチェック項目の整理
各 10 観点のユースケース固有のリスクとチェック項目を説明する。

### Step 3: chapter4.md を参照しフェーズ別アクションを整理
現在のフェーズまたは全フェーズについて、具体的な実施事項を提示する。

### Step 4: リスク登録簿ドラフトを作成
templates/risk_register.md の形式で主要リスクを整理する。

## 注意事項
- 出力は本編ガイドに基づくものとし、独自見解は明示する
- 法規制の最終判断は専門家に委ねるよう注記する
- LLM 出力の限界について必ず言及する
```

3.2 スキルの配置先・ディレクトリ構成

作成した SKILL.md は、各 AI サービスで指定されたフォルダ構成に従って、配置する。例えば以下の AI コーディングエージェントにおける配置先は下記のようになっている。

AI コーディングエージェント	Skills 配置先
OpenAI Codex CLI	.agents/skills/
Claude Code CLI	.claude/skills/
Gemini(Antigravity) CLI	.agents/skills/

「.claude/skills/」や「.agents/skills/」配下のディレクトリ構成例を次に示す。

```
healthcare-ai-safety/
├── SKILL.md          # メイン指示ファイル（必須）
├── chapter3.md       # 3 章: AI セーフティ評価の 10 観点
├── chapter4.md       # 4 章: 開発フェーズ別の評価実践
├── templates/
├── risk_register.md # リスク登録簿テンプレート（任意）
└── checklist.md     # チェックリストテンプレート（任意）
```

3.3 各 AI コーディングエージェントでの実行例

OpenAI Codex CLI

```
# プロジェクト配置
mkdir -p .agents/skills/healthcare-ai-safety
cp SKILL.md .agents/skills/healthcare-ai-safety/
cp chapter3.md  .agents/skills/healthcare-ai-safety/chapter3.md
cp chapter4.md  .agents/skills/healthcare-ai-safety/chapter4.md

# 実行例
codex "一般消費者向け健康管理アプリのリスクアセスメントを実施して"

# 明示的にスキルを指定する場合
# チャット入力で $healthcare-ai-safety-eval と入力

# OpenAI Codex のアプリケーション上からも利用可能
```

Claude Code CLI

```
# プロジェクト配置
mkdir -p .claude/skills/healthcare-ai-safety
cp SKILL.md .claude/skills/healthcare-ai-safety/
cp chapter3.md  .claude/skills/healthcare-ai-safety/chapter3.md
cp chapter4.md  .claude/skills/healthcare-ai-safety/chapter4.md

# 実行例
claude "一般消費者向け健康管理アプリのリスクアセスメントを実施して"

# 明示的にスキルを指定する場合
# チャット入力で /healthcare-ai-safety-eval と入力

# Claude のアプリケーション上からもスキルを登録することで利用可能
```

Google Gemini(Antigravity) CLI

```
# プロジェクト配置
mkdir -p .agents/skills/healthcare-ai-safety
cp SKILL.md .agents/skills/healthcare-ai-safety/
cp chapter3.md  .agents/skills/healthcare-ai-safety/chapter3.md
cp chapter4.md  .agents/skills/healthcare-ai-safety/chapter4.md

# 実行例 Gemini CLI
gemini "一般消費者向け健康管理アプリのリスクアセスメントを実施して"

# 実行例 Antigravity CLI
agy -i "一般消費者向け健康管理アプリのリスクアセスメントを実施して"
```

```
# 明示的にスキルを指定する場合
# チャット入力で /healthcare-ai-safety-eval と入力
# Google Antigravity IDE 上からも利用可能
```

コーディングエージェントを用いることで、自組織のユースケースに特化した AI セーフティ評価観点の提示だけでなく、それを実施するための評価スクリプトの作成や評価実施も可能である。

4 利用上の注意事項

本付録のガイドを利用する際には、AI サービスが生成する評価指針やチェックリストは、あくまで検討の出発点となるドラフトであることに留意されたい。実際にプロダクトの開発を行うにあたっては、医療専門家や法務担当者による確認を経た上で、最終的な判断を行うことが推奨される。また、下記の点についても留意されたい。

- **本編ガイドの最新版の使用：**本編ガイドはリビングドキュメントとして更新される予定のため、最新版の chapter3.md / chapter4.md を使用する
- **機密情報の取り扱いにおけるルール順守：**AI サービスにプロダクト情報を入力する際は、自組織の情報セキュリティポリシーを遵守する
- **網羅性の限界に対する補完：**AI サービスの出力が全てのリスクを網羅したものではないため、あくまでドラフトとして利用し、チーム内でのディスカッションやステークホルダーヒアリング等を行い、抜け漏れを補完する
- **専門家による法規制対応の確認：**本付録のガイドは Non-SaMD のプロダクトの AI セーフティ評価を支援するものであり、SaMD 該当性の判断や個人情報保護法への対応など、法規制に関する事項については、法務・規制の専門家に確認する