AI 品質合意フレームワーク解説書 Rev1.0

2026 年 3 月 31 日

AI 品質マネジメントイニシアティブ　WG3

内容

1 はじめに

近年、ビジネス領域における人工知能（AI）の活用は急速に拡大しており、その戦略的価値は今後さらに高まることが予想される。その際に AI の品質管理は非常に重要となるが、それぞれのビジネスにおいて AI を活用する多様なリスクを考慮し、試行錯誤したうえで品質管理の基準を定めることが求められる。特に、AI は確率統計的に学習・推論を行うため、学習データの分布や運用環境の変化に起因する不確実性があり、100%の品質保証を実現することが困難である。したがって、AI をビジネスで活用する際には、残余リスクの存在を前提としたリスクマネジメントが不可欠となる。

このような状況において重要となるのが、ビジネスに関与する複数のステークホルダー間で、AI に関わるリスク認識とマネジメント方針を共有することである。とりわけ、リスクマネジメントにおける鍵は、品質管理に関連するリスク情報の透明性の確保、リスクに対する合意形成、そして責任分担を明確化するための合意内容の文書化および保存にある。

こうした課題に対応するための手段として、AI 品質合意フレームワークを提案する。AI 品質合意フレームワークは、AI 品質管理における合意プロセスおよび文書化を定型化することで、ステークホルダー間の合意形成を円滑化することを目的とする。また、この共通枠組みをエコシステム内で広く用いることで、相手ごとに異なる枠組みを使用する煩雑さを解消し、余計な手間を削減する効果が期待される。加えて、合意プロセスにおけるベストプラクティスを蓄積・共有し、再利用可能にするという点でも有用である。

例えば、AI を用いたサービス開発者は、この共通枠組みを活用することで、AI モジュール開発を外部ベンダへ委託する際の合意内容を、保険業者との契約書作成や認証機関への提出文書にも転用することができる。また、開発支援を行うコンサルタント企業においても、同一の枠組みに基づくことで、複数の AI サービスや製品の開発者に対し一貫したサポートを提供しやすくなる。このように、AI 品質合意フレームワークはエコシステム内の相互運用性を高め、スケーラブルな品質管理体制の構築に寄与すると考えられる。

2 AI 品質合意フレームワークの目的

2.1 AI のエコシステム内の品質合意をめぐる課題

AI 品質管理をエコシステム内で確立するにあたり、現行の枠組みにはいくつかの構造的課題が存在する。まず、AI のガバナンスやマネジメントのレベルにおいては、国際標準や各種ガイ

ドラインが整備されつつあるものの、製品認証のレベルにおいてステークホルダー間の合意を形成するための共通枠組みが依然として存在しない。このため、認証や審査のプロセスにおいて合意事項をどのレベルで設定すべきかが曖昧となり、品質保証のための協調的な基盤が構築しにくい状況となっている。

また、最終製品およびサービスの品質を確保するために、どのような項目について合意を形成すべきかを明確化することは容易ではない。AI システムはデータ、モデル、アルゴリズム、ユースケースといった多層的要素から構成されるため、品質要件の抽出や合意対象範囲の設定が複雑化しやすい。加えて、欧州 AI Act をはじめとした規制要件の動向、業界特有のコンプライアンス要件、顧客企業の内部基準など、多様な外的要因により合意内容が変動することも、ステークホルダー間の合意形成を一層困難にしている。

さらに、AI 導入に伴い問題が発生した際の責任分担を明確にするには、事前に適切な合意項目を定義する必要がある。しかし、AI の品質問題はデータ提供者、モデル開発者、サービス提供者、運用者など複数の主体が関与する構造を持つため、責任境界を明確に規定することは容易ではない。責任範囲が曖昧なままプロジェクトが進行した場合、トラブル発生時に紛争や追加コストの発生につながり、エコシステム全体のリスクが増大する。

加えて、ビジネス要件は開発時に技術要件にブレークダウンされ、それに基づいてデータ準備やモデルの構築が行われるが、開発はアジャイル的に行われるのが一般的で、準備できるデータ分析の後にビジネス要件や技術要件を修正したりという手戻りが必要となる。サービスの配備後に環境変化に即してビジネス要件を修正することも稀ではない。このような場合には、これらの要件を適切に整理して統合的に扱わなければ、ビジネス要件に対する技術要件の妥当性を追跡し辛くなり、ステークホルダー間での認識祖語が生じやすくなる。AI 品質管理には、要件とデータが作り出す機能を継続的に整合させるアジャイル的な作業が不可欠であり、このプロセスを前提とした合意管理の枠組みが求められる。従来型のウォーターフォール的な合意手順では、AI 特有の不確実性や変更容易性に十分に対応できないことが多い。

さらに、取引相手ごとに個別のフォーマットや基準を用いて合意を交わす現在の状況は、エコシステム全体の運用効率を著しく低下させる。企業ごとに異なる文書体系や項目構成に対応するための工数が増大し、品質保証に必要な情報が断片化するリスクも高まる。このような非効率性は、AI 活用が拡大する中で組織のスケーラビリティを阻害する要因となり得る。

以上のように、AI のエコシステム内で品質合意を確立するためには、共通の合意枠組みの不在、合意項目の複雑性、外的要因の変動、責任分担の明確化の難しさ、要件の多層性、アジャイルな管理手法の必要性、そして個別フォーマット運用の非効率性といった多面的な課題が存在する。これらの課題を解決するための統合的な手法が求められており、次節で述べる AI 品質合意フレームワークは、そのための有効なアプローチとなり得る。

2.2 AI 品質合意フレームワークの目的

AI 品質合意フレームワークは、AI エコシステムに関与する多様なプレイヤー、あるいは企業内部の複数部署間における品質合意の全体像を製品やサービスの要件定義から開発、テスト、保守のライフサイクルに渡って可視化し、共通理解を形成することを主たる目的とする。本フレームワークを用いることで、どの項目について合意が成立しているのか、逆にどの項目が未合意のままであるのかを、連携活動の進行状況に応じて段階的に確認することが可能となる。また、この枠組みは品質管理プロセスそのもの、つまり品質をどこで、どのように定義して評価して管理するかに関する合意形成にも利用でき、プロセスレベルの透明性と説明責任の確保に寄与する。

さらに、AI 品質合意フレームワークは、ビジネス視点での合意事項と技術視点での合意事項をシームレスに接続する役割を担う。AI システムの品質構築においては、ビジネス要件とそれをブレークダウンした技術要件の両者が密接に関連するため、これらを分断せず統合的に扱う枠組みが重要となる。本フレームワークは、合意が必要な項目と、必ずしも合意を必要としない項目を区別し、議論の範囲と優先順位を明確化することで、合意形成プロセスの効率化を図るものである。

また、未合意の項目については、開発プロセスの後段で、より詳細な検討と分解（ブレークダウン）が行われる。これにより、合意に至らない理由を構造的に整理し、必要な追加情報や検討事項を明確にすることができる。最終的には、ビジネス要件と、それを技術要件として具現化した KPI（Key Performance Indicators）を明確化し、設定された KPI の達成によってビジネス要件が満たされていることを客観的に確認できる枠組みを構築することを目指している。

さらに、AI の開発・導入プロセスが要件定義、PoC、実証実験、そして配備、保守といった複数段階から構成されることを踏まえ、本シートは各プロセスにおける合意形成の進捗を確認できるよう設計されている。これにより、ステークホルダーはプロジェクト全体の状況を俯瞰しつつ、必要に応じて合意事項の更新や見直しを行うことが可能となる。

加えて、本フレームワークは問題発生時の責任分担を明確化する役割も担う。AI システムの障害や不具合は複数主体の関与によって発生する可能性があるため、その発生源と管理責任を事前に整理しておくことは紛争防止の観点から極めて重要である。AI 品質合意フレームワークは、開発者、データ提供者、サービス提供者、ユーザーなど、関係者間の責任境界を文書化することで、透明性の高い運用とリスクマネジメントを実現する。

以上のように、AI 品質合意フレームワークは、品質合意の可視化、ビジネスと技術の橋渡し、合意範囲の整理、未合意項目の構造化、KPI 設定の明確化、プロセス全体の進捗管理、そ

して責任分担の明確化といった多面的な目的を持つ。その導入は、AI エコシステムにおける品質管理の成熟度を高め、持続可能な協働体制の構築に寄与するものである。

3 AI 品質合意フレームワークの構成

3.1 AI 品質合意フレームワークの概要

本章では、AI 品質合意フレームワークを構成する要素について述べる。AI 品質合意フレームワークは、「AI 品質要件シート」と「AI 品質合意シート」という二つの文書群から構成され、それぞれが補完的な役割を果たしながら、エコシステムにおける品質合意をライフサイクルに渡って体系的に支援する仕組みとなっている。

3.2 AI 品質要件シート

3.2.1 AI 品質要件シートの位置づけと目的

AI 品質要件シートは、AI を活用して提供する製品・サービスの要件を体系的に記載し、これを起点として AI 品質合意シートで合意すべき項目を漏れなく抽出することを目的とする文書である。AI 品質要件シートの記載により、リスク区分・適用範囲・論点が自動的に明確化され、AI 品質合意シートで合意すべき要件が抽出されることを狙いとしている。

3.2.2 AI 品質要件シートの構成要素

対象とする製品やサービスの概要、準拠対象の規制やガイドラインに加え、AI のふるまいとリスクを決定づける入出力および利用形態の記載から構成される。

- **対象とする製品・サービスの概要、構成と機能**
  - **対象とするサービスや製品の目的や用途、動作の概要**
  - **対象とするサービスや製品がどう構成され、どのように機能するか**
- **対象とする規制・ガイドライン**
  - EU Act.などの法規制
  - AI 事業者ガイドライン、などのガバナンスガイドライン
  - ISO42001 などのマネジメントガイドライン

    - 産総研 AI 品質マネジメントガイドライン、NIST Risk Management Guideline などの技術要件と対応方法を記載したガイドライン
    - 製品・サービスの業界固有の法規制やガイドライン（製造物責任法、薬事法など）
- **入力（Input）**
    - 個人情報（PII）の含有有無／種類（特定個人情報・要配慮個人情報 等）
    - 機密情報・知財・契約制約のあるデータの有無
    - テキスト、画像、音声、センサーデータ、プログラムコード等
- **出力（Output）**
    - 自然言語応答／数値推定／意思決定支援／画像生成/プログラム生成 等
    - 出力に**個人情報・機密情報、第三者の著作物が混入する可能性**の有無
    - 説明可能性の要否（ユーザー・監査向け）
- **利用形態（Use-case/Context）**
    - 対象ユーザー（一般消費者／専門家／社内限定 等）
    - 重大影響の可能性（健康・安全・財務・権利侵害 等）
    - 活用の場所（国内、国外（北米、欧州 等））
- **運用・保守（Operations）**
    - 動作環境の運用中の変化の有無、変化の頻度
    - 外部委託・再委託の有無、SLA・保守体制の有無

3.3 AI 品質合意シート

AI 品質合意シートは、ステークホルダー間で形成されるべき品質に関する合意ポイントを簡潔かつ明瞭に記述するための文書である。本シートでは、合意内容を「AI ビジネス要件」「AI 機能要件」「AI 技術要件」という三つの階層に分類し、ビジネス側と技術側の両視点から一貫した品質合意を形成できるよう構造化している。

また、各合意項目には、その合意状態および合意の充足状態を記載する欄を設けることで、合意形成プロセスを透明化し、進捗を可視化する仕組になっている。これにより、プロジェクトの各段階において、どの項目が確定済みで、どの項目が未確定のまま残っているのかや、いつどの項目がどのように修正、追加されたのかを直ちに把握できる。

なお、合意対象はビジネスプロセス全体を通じて必要となる要件を包括的にカバーする必要がある。例えば、保守・メンテナンスに関わる要件についても、必要に応じて AI ビジネス要件、AI 機能要件、AI 技術要件のいずれにも適切に記載することが求められる。これにより、導入後の運用段階における責任分担や品質維持に関する合意の抜け漏れを防ぎ、AI システムのライフサイクル全体にわたる品質保証を可能とする。

3.3.1　AI 品質合意シートの目的と役割

AI 品質合意シートの主目的は、AI 品質要件シートにより抽出された品質要件を、ビジネス要件・機能要件・技術要件の三層構造に整理した上で、ステークホルダー間で正式な合意内容として確定することにある。とくに、AI システムの品質は複数主体の協働によって成立するため、合意内容を明文化することによって、以下の効果が期待される。

- **合意事項の可視化と漏れの防止**
  品質管理上必要な項目が網羅的に整理され、どの項目が合意済みで、どの項目が未合意であるか、どのように修正、追加されたかが一目で分かる。
- **責任分担の明確化**
  AI 開発者、AI 提供者、AI 利用者ごとの役割・責任境界を整理し、トラブル発生時の紛争防止に寄与する。
- **品質管理プロセスの一貫性確保**
  要件定義フェーズから PoC、実証実験、本番配備、運用保守まで、**フェーズ横断で管理可能な品質基準**を提供する。
- **監査・認証への対応**
  種々の法令やガイドライン（AI 事業者ガイドライン　等）との整合性を必要に応じて確保し、監査や認証制度への適応性を高める。

3.3.2　AI 品質合意シートの文書構造：三層モデルによる体系化

AI 品質合意シートは、AI ビジネス要件・AI 機能要件・AI 技術要件という三層構造で構成される。この階層構造により、ビジネス部門・技術部門・法務部門など、関係者ごとの視点差を吸収しつつ、共通言語で品質を議論できる。

● AI ビジネス要件

安全性、公平性、透明性、説明可能性、プライバシー保護、セキュリティ、責任説明（アカウンタビリティ）、ガバナンス、ロバストネス等 AI を活用した製品やサービスを社会実装するうえでの要求事項。製品やサービスで実際にビジネスを行う部門が、ビジネス遂行に必要な条件を記載する。

● AI 機能要件

上記 AI ビジネス要件を満たすために必要な機能要件。対象とする製品やサービスをに組み込まれた AI の機能についての要件を記載する。
例：フォールバック、バイアス検出、説明提示、PII マスキング、アクセス制御、異常検知、ログ出力

● AI 技術要件

AI 機能要件のそれぞれの充足を具体的に検証するための実装条件・検証方法。検証結果の客観性を担保するために、検証方法、検証に用いるデータや環境、機能要件の充足を明確化するための KPI を具体的に記載する。
例：検証に用いるモデルやシステムの対象、検証環境、検証に用いるデータ、検証方法。安全停止ロジック、SHAP/LIME、暗号化方式、ドリフト検知、改ざん防止ログ。ただし、実現方法は企業機密の場合があるので、その場合は、どのような評価を行い、どのように結果について合意するのかについて、合意に関わるステークホルダーで議論が必要となる。

AI ビジネス要件で管理すべき項目は、AI 品質要件シートの記載に応じて選定されることを想定している。AI ビジネス要件から AI 機能要件、AI 技術要件へのブレークダウンは、対象とする AI コンポーネントの開発や実証作業、運用手順や運用中の環境変化に対応するためのモデル再構築の方法が具体化するにしたがって、記載を進めることになる。

3.3.3　合意状態・充足状態の管理

AI 品質合意シートでは、各要件について、「合意状態（Agreed / Pending / Not Applicable）」および「充足状態（Fulfilled / Partly / Not yet）」を明記する欄を設ける。これにより、以下が可能となる。

- **プロジェクトのフェーズごとに合意の進捗を可視化**
- **ハイリスク領域の未合意・未充足ポイントを早期に発見**
- **対外説明および社内監査に耐える形でエビデンスを残す**

特に、PoC（概念実証）→実証実験→本番配備という段階的導入が一般的な AI 活用では、合意状態の逐次更新が運用の安全性に直結する。

また、履歴管理欄を設け、そこに、それぞれの項目の合意日時、合意担当者（複数部署での合意の場合には、そのすべて）、修正履歴と、修正についての合意担当者を記載する。これにより、AI ライフサイクルのどの時点で、誰がどう修正し、合意したかを明確にできる。

3.3.4　AI 品質要件シートとの連携と要件項目抽出の仕組み

本 AI 品質合意シートは、AI 品質要件シートとの密接な連動を前提とする。品質要件シートの記載に応じて、合意シートに**必要な要件ブロックが抽出できるように設計されている。**

例：プライバシー項目

- 入力：入力に顧客のプライバシー情報（PII）を含むか → 「Yes」

→ 必須となる合意項目

・ビジネス要件：PII 保護に関する組織方針とリスク低減方針

4　AI 品質合意フレームワークの使い方

4.1　使い方の流れの例

AI の品質要件は初期段階で完全合意に至らないことが多いため、**短いサイクルで段階的に合意事項を増やす**方針が有効である。AI システムは PoC や実証実験などの開発フェーズ、あるいは運用後の性能低下・データ分布変化・新リスクの顕在化などに対応するため、それぞれのタイミングで品質要件、その充足の確認方法、充足状況を更新していく。

要件定義フェーズ：品質要件について初回合意

PoC：初期検証の結果を反映

実証実験：実運用に近い環境でのリスク結果を反映

本番導入前：責任分担・説明可能性・保守体制の最終合意

運用フェーズ：ログ分析やドリフト検知に基づく更新

モデル更新：再学習・モデル変更時の合意再検証

この継続的な合意は、AI 事業者ガイドラインが推奨する「AI ガバナンスのライフサイクル管理」に即したものである。 [meti.go.jp]

合意プロセスは事業活動と**独立して管理するのではなく、事業計画の一部として並走**させ、関係者全員が**プロジェクト進展と合意成熟が同期**していることを把握できるようにする。

- **最初に決めやすい項目：**
  法令・規格に関する要件、ポリシー／ビジョンに基づく必須のビジネス要件、社内外で標準化された AI 機能要件・技術詳細要件。
- **途中で見直す項目：**
  ビジネス要件充足のための機能・技術要件、PoC や試行錯誤から得た学びの反映、ビジネス環境変化の反映。
- **最終段階で確定する項目：**
  実運用評価を要するビジネス要件、プロダクト評価結果に基づく機能・技術要件。

また、上で示した通り、合意は何度も変更されるのが通常の使われ方です。その際に、各種要件がどのように変化し、どのように合意が修正されたかをあとで確認できるようにすることが重要です。

- **変更履歴の厳格管理**：版数・日付・変更箇所・理由・担当・根拠資料（試験成績書、ログ、監査記録等）を**必ず保存**する。
- **段階的詳細化／改変**：プロセス進展に合わせ、**必要箇所を詳細化**し、仮置き値を**確定値へ更新**する。
- **ステータスの遷移規律**：「合意」から「充足」への遷移には、**エビデンス提出と相互確認**を必須とし、却下・差戻しの基準も明文化する。

4.2　要件記載例

下記に、AI 事業者ガイドラインへの準拠を想定した AI 合意シートに記載されるべき要件の項目例を示す。AI 品質要件シートの記載に応じて下記から必要となるものが選択される。なお、技術要件について、本来は評価の方法、用いるデータ、KPI の詳細など要件の充足を客観的に示すことができる情報が記載されるが、ここでは、その項目例だけ記載している。また、（1）の AI パフォーマンスは AI 事業者ガイドラインには明示されていないが、AI がもともと求められている役割を果たすかどうかの視点で品質管理に不可欠なので、付加してある。

（1） AI パフォーマンス（AI Performance）

ビジネス要件：もともとの AI 機能に求められる価値創造のための機能が正しく動作する。
機能要件：認識制度や予測精度、チャットボットの回答が求められる性能を満足する。
技術要件：実証実験における試験において、ご回答率が基準以下。F 値が指定範囲内に収まる。

(2) 安全性（Safety）
ビジネス要件：重大な誤作動による危害・損害を最小化する。高リスク領域は HITL（Human in the Loop）を必須とする。
機能要件：誤動作率最小化、重大誤判断時の自動フォールバック、異常入力ハンドリング。
技術要件：誤動作率 KPI を満たす、安全停止ロジック、異常系／ストレステスト、監視ログ・冗長化。

(3) 公平性（Fairness）
ビジネス要件：属性（性別・年齢・人種等）による不当な偏りを排除。
機能要件：要配慮属性の洗い出し。不当な偏りが基準値以下。偏り検出機能、公平性レポート。
技術要件：公平性指標（Statistical parity / Equal opportunity 等）算出、データバイアス検出、定期再評価。

(4) 透明性・説明可能性（Transparency/Explainability）
ビジネス要件：AI 判断の説明可能性と AI 使用の明示。
機能要件：根拠提示、AI 使用箇所表示。
技術要件：SHAP/LIME 等の XAI ツール、根拠抽出 API、UI での AI 生成表示。

(5) プライバシー保護（Privacy）
ビジネス要件：個人情報の適正管理と出力混入防止。
機能要件：PII 検出・マスキング、削除請求対応。
技術要件：PII 検出モデル、暗号化・アクセス制御、混入検知フィルタ。

(6) セキュリティ（Security）
ビジネス要件：モデル・データ・推論環境の不正アクセス防止。

機能要件：認証／認可、改ざん検出、脆弱性診断記録。
技術要件：監査ログ暗号化、Prompt Injection 対策、定期スキャン。レッドチーミングテスト。

(7) アカウンタビリティ（責任説明）
ビジネス要件：開発／提供／利用における責任範囲の明確化。
機能要件：責任主体・エスカレーションの明示、監査ログ出力。
技術要件：推論ログ（入力・出力・モデル版）の完全記録、改ざん防止。

(8) ガバナンス（Governance）
ビジネス要件：経営層主導の AI ガバナンス体制、ライフサイクル全体のリスク管理。
機能要件：承認ワークフロー、リスク評価・更新手順。
技術要件：モデル／データ品質管理、監査データ保存・バックアップ。

(9) 運用中のロバストネス（Robustness）
ビジネス要件：環境変化・不正入力下でも信頼性を維持。
機能要件：異常検知・フォールバック、ドリフト監視。
技術要件：敵対的攻撃テスト、ドリフト検知と再学習トリガー。

4.3 合意の RACI

合意形成にかかわる種々の作業について、RACI（Responsible, Accountable, Consulted, Informed）の枠割分担例を示す。

活動／責務	Responsible（実行責任）	Accountable（最終責任）	Consulted（助言・審査）	Informed（報告先）
1. AI 品質要件シートの作成	AI 開発者、AI 提供者	プロジェクト責任者	法務／ガバナンス部門、データ提供者	経営層、AI 利用者

2. AI ビジネス要件の整理	事業部門（業務オーナー）	プロダクトオーナー	法務、経営企画、ガバナンス部門	AI 開発者、AI 提供者
3 AI 機能要件の整理	AI 開発者、AI 提供者	プロジェクト責任者	業務部門、法務、品質保証部門	AI 利用者
4. AI 技術要件の整理	AI 開発者	AI 技術責任者	セキュリティ部門、品質保証部門	AI 提供者、AI 利用者
5. リスク分析（安全性・公平性・透明性など）	AI 開発者、AI 提供者	技術責任者 or セーフティ責任者	法務、ガバナンス部門、セキュリティ部門	経営層、AI 利用者
6. 合意プロセス計画	プロジェクト責任者	プロダクトオーナー	関係部門（開発・業務・法務）	経営層
7. 合意内容の記載	担当者（ビジネス・機能・技術）	プロジェクト責任者	関係者レビュー	全ステークホルダー
8. 合意内容の確認・承認	合意メンバー（開発・提供・利用側の代表）	プロダクトオーナー／経営層	法務、ガバナンス部門	全関係者
9. 運用フェーズの定期レビュー	AI 利用者／AI 提供者	運用責任者	AI 開発者、法務、ガバナンス部門	経営層
10. モデル更新・再学習時の合意内容見直し	AI 開発者	プロダクトオーナー or 技術責任者	セキュリティ部門、業務部門	全関係者
11. 監査／証跡用ドキュメント管理	ガバナンス部門／品質保証部門	経営層	法務、情報セキュリティ部門	プロジェクトメンバー
12. 規制変更（AI ガイドライン・AI Act 等）への対応	法務、ガバナンス部門	経営層	プロジェクト責任者、AI 開発者	全関連部署

5 AI 品質合意フレームワークの使用例について

5.1 自動運転向け人物認識の例

5.1.1 AI 品質要件シート

(1) サービス/製品の概要

- **対象**：自動運転車両に搭載する人物検出 AI コンポーネント（静止画像ベース）
- **AI の役割（ビジネス要件の核）**：
  車載カメラから取得した静止画像（1920×1080）を入力として、人の有無を判定し、映っている箇所（バウンディングボックス）を提示する。性別・人種・年齢を問わず人物を検出し、腕のみ等の一部が映っている場合でも検出対象とする。ヘッドライト等の照明がある前提で昼夜を問わず検出し、道路種別・路面状態によらず検出する。雨・雪で視認性が極端に低い場合は非対応可とする（機能外条件として明記）。
- **非機能上の前提**：AISL=1（本 AI 単体は安全動作の最終判断を行わない）。緊急停止や回避などの安全性確保は他サブシステムで実施。敵対的サンプル攻撃は別サブコンポーネントが並列で検出する。

(2) 準拠すべきガイドライン、規制

- **規格・ポリシー**：社内 AI ガバナンスガイドライン、AI 事業者ガイドライン、車載品質基準（例：ログ保全・変更管理）等（必要に応じ追記）

(3) 入出力要件

要件	チェック	備考
入力は静止画像？	Yes	車載カメラ静止画、1920×1080
入力に個人情報を含む？	No	基本は遠景人物。顔識別等は行わない前提
外部通信を行うか？	No	完全オンボード（組み込み PC）

(4) 利用形態・運用

要件	チェック	備考
24/7 運用か	Yes	車両稼働時間に準拠
誤検出時のフォールバックはある	Yes	上位安全システム側で HITL／複数センサ冗長
ログ保存は必要	Yes	BBox・スコア・モデル版・環境メタデータ
ドリフト監視は必要か	Yes	季節・地域差に応じた外観変化

5.1.2 AI 品質合意シート（例：自動運転向け・人物検出）

下記に AI 品質合意シートの記載例を示す。見やすさのために、合意チェック、充足チェック欄は省略している。また、試験方法については、本来は試験に用いる具体的データや試験環境の記載が求められる。

	AI ビジネス要件	**AI 機能要件**	**AI 技術要件**	**KPI/閾値**	**試験/監査方法**
AI パフォーマンス	物体検出（主要 5 クラス）	歩行者/車/二輪等のクラス検出	mAP 算出・IoU 処理	mAP@IoU0.5 ≥90%	ベンチマーク試験
安全性	安全側動作（Fail-safe）	誤検出時フォールバック通知	異常系テスト・安全停止ロジック	FN(重要クラス) ≤0.5%	シナリオ試験
安全性	低照度（夜間）対応	夜間補正処理	夜間データセット mAP 評価	夜間 mAP ≥ 85%	低照度試験
安全性	ドリフト監視	性能低下アラート	閾値判定（低下≥5%）	低下≥5%で通知	監視ログ分析
安全性	リアルタイム性（レイ	高速推論処理	P95 レイテンシ測定	P95 ≤200ms	負荷試験

	テンシ管理）				
安全性	誤検出フォールバック	誤検出時の安全系連携	フォールバック実装・ログ	成功率=100%	安全系統合試験
公平性	環境差・属性差による認識率の乖離抑制	サブグループ性能監視	サブグループ評価シナリオ	差分 ≤ ±3%	夜間/季節/地域別評価
透明性	説明可能性（XAI）	bbox/score 根拠提示	説明ログ出力・可視化	説明付与率 100%	監査ログレビュー
プライバシー保護	個人識別情報を扱わない前提	PII サニタイズ（該当時）	入力画像暗号化		
セキュリティ	情報漏洩なし	攻撃検出・アクセス制御	FGSM/PGD 耐性・改ざん防止ログ	逸脱ゼロ	NW ポリシー監査
セキュリティ	敵対的攻撃耐性	軽微攻撃検出・緩和	FGSM/PGD 評価	誤判定 ≤5%	敵対的試験
透明性・説明責任	誤検出時の根拠表示、ログ完全性	bbox/score/モデル版/環境情報ログ	改ざん防止ログ	ログ完全性、説明付与率 100%	監査
運用、保守	再学習トリガ	モデル更新検知	再学習ルール適用	3 ヶ月 or ドリフト検知時	運用ルール

5.2 RAG を使った社内規定検索の例

5.2.1 AI 品質要件シートの記載例

(1) サービス／製品の概要

- **名称**：社内規定検索アシスタント（RAG）

- **対象業務**：**人事／調達／広報／貿易管理**に関する**社内規定**の検索・要約・引用提示
- **目的**：社員の規定参照負荷を軽減し、**最新・正確・根拠付き**の回答を即時に提示する
- **前提**：回答生成は**社内規定 RAG 用 DB** に**厳格に限定**。規定にない内容は「不明」と回答し、**問い合わせ先**を案内。**社外への情報送出はなし**（社内クローズド環境）

(2) 準拠すべきガイドライン、規制

・ 規格・ポリシー：社内 AI ガバナンスガイドライン、AI 事業者ガイドライン、車載品質基準（例：ログ保全・変更管理）等（必要に応じ追記）

(3) 入出力要件

項目	設問	チェック	派生する要件	合意対象（ビ/機/技）
入力	社員の自由文質問を受け付けるか	Yes	日本語中心（英語は将来拡張）	機
入力	個別事案・機微情報（個人名、案件名等）を含む可能性	Yes	**入力サニタイズ／PII 検出・マスキング**を実施	技
出力	回答に根拠（規定の条番号・原文）を必ず添付するか	Yes	**出典リンク／抜粋**の必須化	機/技
出力	**社内規定に関係しない質問**には回答しないか	Yes	**拒否方針**・テンプレ返答・誘導プロンプト	ビ/機
出力	**規定に記載がない**場合は「不明＋問い合わせ先」提示か	Yes	**不明ハンドリング**と**窓口ルーティング**	ビ/機
スタイル	**ビジネスにふさわしい文体**の維持	Yes	トーン／禁止表現／校閲ルール	機
セキュリティ	**社外通信なし**で動作	Yes	オンプレ/仮想私有網。外部 API 呼出し禁止	技

ログ	問合せ内容・回答・使用ソース・モデル版の**監査ログ**保存	Yes	PII 削除方針・保存期間を定義	技

⑷ その他留意事項

1. **スコープ厳守**：本アシスタントは**社内規定（人事／調達／広報／貿易管理）に限る**。関連しない質問には**回答しない**。
2. **根拠提示**：回答の根拠（規定の条番号・章節・抜粋）を必ず添付し、社員が一次情報で確認できるようにする。
3. **不明応答**：規定に記載がない事項は「不明」と明言し、担当窓口（問い合わせ先）を案内する。
4. **機密配慮**：**個別事案の内容が漏えいしない**ように取り扱う（入力・出力・ログの全過程で最小限主義）。
5. **ビジネストーン**：回答は**丁寧・簡潔・中立**で、規定の**誤解を招かない**表現とする。
6. **有害・不適切出力の抑制**：**非敵対的入力で不適切出力 ＜1%、敵対的入力で ＜30%**（後述の評価プロトコルで測定）。

5.2.2 AI 品質合意シート記載例

	ビジネス要件	**機能要件**	**技術要件**	**KPI / 閾値**	**試験 / 監査方法**
安全性	誤情報の抑制（規定外回答禁止）	クエリ分類（規定領域/非該当/不明）	拒否テンプレ適用・出力検証	拒否妥当率 ≥ 99%	スコープ外デッキ評価
安全性	不適切出力の抑制	安全フィルタ（通常/敵対的）	有害性評価モデル・閾値設定	通常 <1% / 敵対 <30%	有害性デッキ評価

公平性	特定部署・個人に不利益となる偏り防止	質問内容の正規化	バイアス検知（通常/敵対的入力）	バイアス差分 ≤ ±3%	サブグループ評価
プライバシー	機微情報混入防止	入力サニタイズ / PII 検知	PII 検出モデル・PII マスキング	PII 検出 F1 ≥ 0.97	擬似 PII 評価・敵対的 PII 混入試験
セキュリティ	社内限定利用（外部送信なし）	外部 API 無効化	NW 閉域化・アクセス制御・改ざん防止ログ	逸脱ゼロ	NW ポリシー監査・通信ログ監査
セキュリティ	ログ最小化と保全	最小限ログルール	ログ暗号化・保存期間設定	保存期間遵守 100%	ログ仕様監査・保持期間検証
アカウンタビリティ	根拠トレーサビリティ確保	チャンク ID 付与・参照ログ	監査ログ・完全性ハッシュ	ログ完全性 100%	ハッシュ照合・監査ログレビュー
アカウンタビリティ	モデル更新の追跡	モデル版管理	更新ログ・再評価プロトコル	再評価実施率 100%	更新時の再評価ログ確認
透明性	不明応答の適正誘導	不明時テンプレ応答	問い合わせ先ルーティングロジック	不明応答の妥当性 100%	テンプレ一致確認・ルーティング試験
透明性	根拠提示（条番号・抜粋）	引用生成機能	引用スキーマ・MMR/Top-k 設定	根拠一致率 ≥ 95%	引用品質テスト（人手評価／サンプルレビュー）

6 おわりに

本稿では、AI 活用が組織の価値創出に直結する時代において、関係者間の合意形成を体系的に支援する枠組みとして、AI 品質合意フレームワークの目的、構成、そして使い方について論じた。AI はその性質上、従来のソフトウェアと異なり、学習データやモデル構造、運用環境によって性能やリスクが動的に変化する。このような不確実性を前提とするためには、単に仕様を文書化するだけでは不十分であり、ステークホルダーが共通の認識を持ち、段階的に合意を成熟させていく仕組みが不可欠である。AI 品質合意フレームワークは、この課題に対する実務的かつ再現性のある解法を提示するものである。

特に、本シートが採用する **三層構造（ビジネス要件・AI 機能要件・技術詳細要件）** は、非技術者と技術者が同一の文書上で議論しながらも、それぞれの専門性に応じて焦点を合わせることを可能とする。これにより、ビジネス観点での目的・責任範囲と、技術観点での検証可能な品質要件とが明確に連結され、プロジェクト全体の透明性が向上する。また、「なし／合意／充足」による段階的合意の可視化は、AI 固有の試行錯誤や改善サイクルと極めて相性が良く、開発・実証・配備・保守というライフサイクル全体を通じて、リスクと品質を一貫して管理する基盤として機能する。

さらに重要なのは、本シートが特定プロジェクトに閉じた成果物ではなく、**組織として蓄積・発展させていけるナレッジ基盤**となりうる点である。複数プロジェクトにわたって蓄積された合意プロセス、評価基準、失敗知見、改善提案は、次のプロジェクトに再利用可能な「質の高いテンプレート」として展開され、組織全体の AI ガバナンスを強化する。これにより、個別の開発チームや開発企業依存のばらつきが減少し、AI 活用に関わる品質・透明性・説明責任が標準化されていく。

今後、AI 技術はより多様化し、規制環境（国内指針、国際標準、AI 法制など）も一層複雑化していくことが予想される。そのなかで、AI 品質合意フレームワークは、組織が自律的に AI ガバナンスを確立し、持続的に改善し続けるための「基盤ツール」として活躍するだろう。本シートをプロジェクトの形式的手続きとしてではなく、リスクを共有し、価値を共創するための対話の媒体として活用することで、AI が組織にもたらす恩恵を最大化しつつ、安全性・透明性・信頼性を確保することができる。

AI 品質合意フレームワークは、単なる文書ではなく、組織の AI 活用能力そのものを高めるための枠組みである。本稿で示した視点と方法論が、読者の組織が AI 活用をより効果的かつ責任あるかたちで推進するための一助となることを期待したい。

本フレームワークは、AI 品質マネジメント検討委員会 WG3（エコシステム）で 1 年間かけて検討した内容をまとめたものである。本稿に挙げたような例をもとに開発プロセスを想定し

て AI 品質要件シート、AI 品質合意シートを設計してきたが、実際のビジネスに活用するには、さらに細部を詰める必要がある。2026 年度に、実応用あるいは、それに近いレベルの応用を対象として開発、合意のプロセスに本フレームワークを活用して、活用のためのリファレンスを蓄積するとともに、フレームワークをより使いやすいように更新していく予定である。

7 参考文献

(1) 日本のガバナンス・実務指針

1. 総務省・経済産業省『**AI 事業者ガイドライン（第 1.1 版）**』本編・別添（2025–2026 更新）。ガイドライン全体構成、共通の指針 10 項目、主体別（開発者／提供者／利用者）の実務を網羅。
    - 経産省「第 1.1 版」資料一式（本編・別添・概要等） [meti.go.jp]
    - 概要資料（第 1.1 版 概要 PDF） [soumu.go.jp]
    - 別添（付属資料）概要（実践“how”の整理） [meti.go.jp]

2. **生成 AI 品質マネジメントガイドライン 第 1 版**（2025 年 5 月 26 日）。
   対象：LLM 等の基盤モデルを部品として利用する**生成 AI システムの品質マネジメント**。スコープ外質問の拒否、根拠付き回答、ログ最小化や改ざん防止、レイテンシ SLO など、**現場運用に直結する管理策**まで具体化。
   所収：産総研デジタルアーキテクチャ研究センター公開ページ／AIST リポジトリ（DOI）／産総研ニュースリリース。
    - 公開ページ（DigiARC）：生成 AI 品質マネジメントガイドライン 第 1 版 [digiarc.aist.go.jp]

3. **機械学習品質マネジメントガイドライン（第 4 版ほか）**。
   対象：識別・予測型の**機械学習 AI システム**のライフサイクル品質マネジメント。外部品質（リスク回避性／AI パフォーマンス／公平性 等）と、これを担保する**内部品質**（データ設計・被覆性・モデル安定性・運用監視 等）を構造化。
    - 第 4 版（Rev.4.1.0、2023-12-12、CC BY 4.0）：PDF（DigiARC） ／ 総合ポータル（各版・関連資料）：公開ページ [digiarc.aist.go.jp], [digiarc.aist.go.jp]

(2) 国際標準（マネジメント／リスク／品質モデル／テスト）

1. **ISO/IEC 42001:2023** — Artificial intelligence — **Management system**（AI マネジメントシステム；AIMS）。AI ガバナンスの仕組み・責任・監査まで含む要求事項を体系化。
    - ISO 公式解説ページ（標準概要） [iso.org]
    - 参照用 PDF（流通サイト・学術リポジトリ） [cdn.standa...ds.iteh.ai], [bibrepo.uca.es], [gsc-co.com], [img.antpedia.com]
2. **ISO/IEC 23894:2023** — Artificial intelligence — **Guidance on risk management**（AI リスクマネジメント指針）。AI 特有リスク原則・プロセスと組織統合のガイダンス。
    - 標準本文の参照用 PDF（iTeh 等）／ISO 公式案内 [cdn.standa...ds.iteh.ai], [iso.org]
    - 追加の参照情報（他配布元・ハブ掲載） [img.antpedia.com], [aistandardshub.org], [i2.saiglobal.com]
3. **ISO/IEC 25010:2023** — SQuaRE **製品品質モデル**。機能適合性・性能効率・互換性・ユーザビリティ・信頼性・セキュリティ・保守性・可搬性等を定義し、品質要求・評価の共通参照に有用。
    - ISO 公式ページ（第 2 版 2023） [iso.org]
    - 参照用 PDF（iTeh 等）と解説サイト [cdn.standa...ds.iteh.ai], [iso25000.com], [quality.arc42.org]
4. **ISO/IEC TR 29119-11:2020** — Software testing — **Guidelines on the testing of AI-based systems**（AI ベースシステム試験ガイドライン）。ブラックボックス／NN 向けホワイトボックス、テストオラクル問題等を整理。
    - ISO 公式ページ／参照用 PDF（iTeh）／JSA 案内／AI Standards Hub／OECD.AI 掲載 [iso.org], [cdn.standa...ds.iteh.ai], [webdesk.jsa.or.jp], [aistandardshub.org], [oecd.ai]

(3) リスクマネジメント・横断原則（米国・OECD・G7）

1. **NIST AI Risk Management Framework (AI RMF 1.0)**（2023）。Govern–Map–Measure–Manage の 4 機能、トラストワージネス特性、安全・公平性・透明性等を包括。生成 AI プロファイルも公開。
    - NIST 解説ページ（AI RMF・Playbook・Resource Center） [nist.gov]

- 公式 PDF（NIST AI 100-1）／日本語版（AISI） [nvlpubs.nist.gov], [aisi.go.jp]
- AIRC（NIST）による要点整理ページ [airc.nist.gov]

2. **OECD AI Principles**（2019）。政府が合意した初の AI 原則（人間中心・透明性・堅牢性・アカウンタビリティ等）—各国法制・国際議論の共通参照。
   - OECD 公式ページ／OECD.AI ポータル（更新情報） [oecd.org], [oecd.ai]
   - 原則の概要 PDF（フライヤー） [archive.epic.org]
3. **G7 広島 AI プロセス**（2023）— 国際ガイディング・プリンシプル／開発組織向け行動規範、**包括的政策フレームワーク**等。
   - 総務省：Digital & Tech Ministers' Statement（Comprehensive Policy Framework 添付） [soumu.go.jp]
   - 外務省ほか：G7 首脳声明（広島 AI プロセス）／欧州委の紹介ページ等 [mofa.go.jp], [digital-st....europa.eu]
   - G7 情報センター（OECD 作成のジェネレーティブ AI 報告） [g7.utoronto.ca]

(4) 規制枠組み（EU）

1. **EU AI Act** — 規則（EU）2024/1689（2024 年 7 月 12 日 OJ 掲載）。禁止行為、**高リスク AI の要求事項**、市場監視・ポストマーケット監視、コード・オブ・コンダクト等。
   - 公式ジャーナル掲載（条文 PDF 等への導線）／市民向け解説サイト（最終版公開） [eur-lex.europa.eu], [artificial...enceact.eu]
   - 条文全文の閲覧サイト（最終テキスト） [aiact-info.eu], [aiactinfo.eu], [artificial...ce-act.com]

(5) 研究動向（公平性・説明責任・評価）

1. **ACM FAccT（Fairness, Accountability, and Transparency）** — 公平性指標・評価・監査・社会実装論点の国際会議。
   - 公式サイト（開催情報・トピック） [facctconference.org]
   - 近年の議論総覧（研究動向のメタ分析） [cs.cmu.edu]

- 会議掲載のプログラム／論文集一覧ページ（一例） [researchr.org]

(6) 実務に役立つ補助資料（参考）

1. **AI Standards Hub**（BSI/UKRI 等）— ISO/IEC 23894・TR 29119-11 ほか国際標準の解説・リンク集。
    - ISO/IEC 23894 概説ページ／TR 29119-11 概説ページ [aistandardshub.org], [aistandardshub.org]

(7) 使い方（本解説文とのひも付け）

- 合意の三層（ビジネス→機能→技術）を規格にマッピング：
    - 組織面の仕組みは **ISO/IEC 42001**（AIMS）と **NIST AI RMF** の「Govern」機能を参照。合意の**責任分担・監査可能性**の設計に資する。 [iso.org], [nvlpubs.nist.gov]
    - リスク評価・緩和方針は **ISO/IEC 23894** と **AI 事業者ガイドライン**の「リスクベースアプローチ」を根拠に明文化。 [iso.org], [meti.go.jp]
    - 品質特性・非機能要件は **ISO/IEC 25010** を参照して粒度と用語を統一。 [iso.org]
    - テスト設計・エビデンスは **ISO/IEC TR 29119-11** を拠り所に、AI 特有の試験観点（オラクル問題、NN 向け WB テスト）を明示。 [iso.org]
- **規制・原則との整合**：
    - **EU AI Act** の高リスク要件（データ・ガバナンス、テクニカルドキュメント、ロギング、人間の監督、精度・堅牢性・サイバーセキュリティ）を、合意項目の**チェック欄**や**配備ゲート**に反映。 [aiact-info.eu]
    - **OECD AI Principles・G7 広島 AI プロセス**の原則・行動規範を、社内ポリシーや**テンプレ合意文言**に組み込む。 [oecd.org], [soumu.go.jp]