AI セーフティに関する評価観点ガイド

（第 1.01 版）

令和 6 年 9 月 25 日

AI セーフティ・インスティテュート

AISI Japan AI Safety Institute

目次

1 はじめに 3
1.1 本書の目的 3
1.2 本書で使用する用語 4
1.3 想定読者 6
2 AI セーフティ 8
2.1 AI セーフティ評価の目的 8
2.2 AI セーフティにおける重要要素 8
2.3 AI セーフティの評価観点 9
2.4 AI セーフティに関する評価のスコープ 9
3 評価観点の詳細 11
3.1 有害情報の出力制御 12
3.2 偽誤情報の出力・誘導の防止 13
3.3 公平性と包摂性 15
3.4 ハイリスク利用・目的外利用への対処 16
3.5 プライバシー保護 17
3.6 セキュリティ確保 18
3.7 説明可能性 19
3.8 ロバスト性 20
3.9 データ品質 21
3.10 検証可能性 22
4 評価実施者及び評価実施時期 23
4.1 評価実施者 23
4.2 評価実施時期 23
5 評価手法の概要 27
5.1 技術的評価 27
5.1.1 ツールによる評価 27
5.1.2 レッドチーミングによる評価 27
5.1.3 その他の技術的評価 27
5.2 マネジメント的評価 27
6 評価に際しての留意事項 28
参考文献一覧 29

1　はじめに

1.1 本書の目的

AI に関連する技術の進展と社会全体への普及は過去に類を見ないほど急速に進んでいる。これまでも AI は推論や探索などを行うエキスパートシステムや強化学習などで利用され、幾度かその普及の兆しを見せていた。以後もコンピュータの計算性能や半導体加工技術の向上により、効率的なアルゴリズムの開発や利用用途の拡大が進んでいる。そのような中、生成 AI の登場によりイノベーションが加速している。特に対話型生成 AI により活用の裾野が広がり、さらに近年では、大規模なデータセットで学習・訓練された基盤モデルが、幅広い用途で活用されてきている。現在では AI を活用したサービスは一般消費者向けのものから特定の事業者や産業向けのものまで多岐に渡っている。さらに、基盤モデルの汎用性の高さから、簡易な自然文での指示によるプログラム開発や、自律性を持たせた AI エージェントなど、ユースケースも多様化が進んでいる。このような AI システムの開発・提供・利用のさらなる普及は、イノベーションの促進や社会課題の解決に寄与するものと期待されている。その一方で、AI システムの悪用や誤用、不正確な出力などの懸念も大きくなっている。我が国は、安全・安心で信頼できる AI の実現に向けた広島 AI プロセスを主導し、「高度な AI システムを開発する組織向けの広島プロセス国際指針」および「高度な AI システムを開発する組織向けの広島プロセス国際行動規範」をとりまとめるなど、AI セーフティに関わるグローバルな規律の策定に貢献してきたが、こうした AI セーフティを実装していくにあたっては、AI システムの挙動がブラックボックスである点など、AI 固有の新たな問題や複雑さが課題となる。また、生成 AI の入出力のバリエーションが多岐にわたる点も挙げられる。特定の用途に向けて設計された AI システムが、広範な目的への応用のために使用されることもある。さらに、適切な AI システムの実装や利用がなされない場合には人間の心理面にも悪影響を及ぼす可能性がある。そのため、AI セーフティを確保し続けることは、急速な普及が進んでいる AI が社会の持続的な発展に寄与するための前提となる。

上記を踏まえ、「AI セーフティに関する評価観点ガイド」（以下、「本書」という。）は、AI システムの開発や提供に携わる者が AI セーフティ評価を実施する際に参照できる基本的な考え方を提示する。本書の作成に際しては、日本において AI を活用する事業者が適切に AI を活用するための指針を示す「AI 事業者ガイドライン（第 1.0 版）」（以下、「AI 事業者ガイドライン」という。）を参考としている。加えて、海外の動向を精査し、AI セーフティを評価するための観点を整理している。これらを参照することで、AI セーフティ評価の要点を確認することができる。なお、AI 事業者ガイドラインは国際的な議論も踏まえ、Living Document として適宜更新を行うことを予定している。そのため、本書についても、国内外の AI セーフティに関する議論や技術動向に応じ、適宜改定を行う。

1.2 本書で使用する用語

本書で使用する用語を以下のとおり定義する。

AI システム

活用の過程を通じて様々なレベルの自律性をもって動作し学習する機能を有するソフトウェアを要素として含むシステムとする（機械、ロボット、クラウドシステム等）。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.8）

AI モデル

AI システムに含まれ、学習データを用いた機械学習によって得られるモデルで、入力データに応じた予測結果を生成する。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.9）

AI セーフティ

人間中心の考え方をもとに、AI 活用に伴う社会的リスク※を低減させるための安全性・公平性、個人情報の不適正な利用等を防止するためのプライバシー保護、AI システムの脆弱性等や外部からの攻撃等のリスクに対応するためのセキュリティ確保、システムの検証可能性を確保し適切な情報提供を行うための透明性が保たれた状態。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」）

※社会的リスクには、物理的、心理的、経済的リスクも含む（出典：Department for Science, Innovation and Technology, UK AI Safety Institute “Introducing the AI Safety Institute”）

AI セーフティ評価

AI システムが AI セーフティの観点で適切であるかどうか見定めること。

※AI セーフティの観点とは、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」及び「透明性」を重要要素とした見方である。

生成 AI

文章、画像、プログラム等を生成できる AI モデルにもとづく AI の総称を指す。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.9）

基盤モデル

広範なデータで学習された、下流の幅広いタスクに適応可能な AI モデル。

（出典：Stanford Institute for Human-Centered Artificial Intelligence “Reflections

on Foundation Models”）

AI ガバナンス

AI の利活用に潜むリスクをステークホルダーにとって受容可能な水準で管理しつつ、そこからもたらされる正のインパクト（便益）を最大化することを目的とする、ステークホルダーによる技術的、組織的、及び社会的システムの設計並びに運用。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.9）

大規模言語モデル（LLM）

基盤モデルの考え方に基づくニューラル言語モデル（Neural Language Models）であり、自然言語文の集まりからなる大規模コーパス(Corpus)を訓練用データとして得た事前学習モデル。

（出典：産業技術総合研究所「機械学習品質マネジメントガイドライン 第 4 版」, P.139）

人間中心

AI システム・サービスの開発・提供・利用において、全ての取り組むべき事項が導出される土台として、少なくとも憲法が保障する又は国際的に認められた人権を侵すことがないようにすること。また、AI が人々の能力を拡張し、多様な人々の多様な幸せ（well-being）の追求が可能となるように行動すること。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.12）

人間中心の AI 社会原則とは、AI の利用は、憲法及び国際的な規範の保障する基本的人権を侵すものであってはならないという原則のこと。

（出典：統合イノベーション戦略推進会議決定「人間中心の AI 社会原則」, P.8）

安全性

AI システム・サービスの開発・提供・利用を通じ、ステークホルダーの生命・身体・財産に危害を及ぼすことがないようにすること。加えて、精神及び環境に危害を及ぼすことがないようにすること。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.14）

公平性

AI システム・サービスの開発・提供・利用において、特定の個人ないし集団への人種、性別、国籍、年齢、政治的信念、宗教等の多様な背景を理由とした不当で有害な偏見及び差別をなくすよう努めること。また、各主体は、それでも回避できないバイアスがあることを認識しつつ、この回避できないバイアスが人権及び多様な文化を尊重する観点から許容可能か評価した上で、AI システム・サービスの開発・提供・利用を行うこと。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.15）

プライバシー保護

AI システム・サービスの開発・提供・利用において、その重要性に応じ、プライバシーを尊重し保護すること、及び関係法令を遵守すること。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.15）

セキュリティ確保

AI システム・サービスの開発・提供・利用において、不正操作によって AI の振る舞いに意図せぬ変更又は停止が生じることのないように、セキュリティを確保すること。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.16）

透明性

AI システム・サービスの開発・提供・利用において、AI システム・サービスを活用する際の社会的文脈を踏まえ、AI システム・サービスの検証可能性を確保しながら、必要かつ技術的に可能な範囲で、ステークホルダーに対し合理的な範囲で情報を提供すること。

（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」, P.16）

1.3 想定読者

本書は、AI システムの開発及び提供の過程に関与する事業者（AI 事業者ガイドラインに記載されている「AI 開発者」及び「AI 提供者」）を主な読者とする。これらの事業者内でも特に、開発・提供管理者（実際に AI システムの構築や提供に関する業務を管理する者）と、事業執行責任者（事業戦略と一体で AI セーフティの維持または向上の施策を推進する責任をもつ者）が主な読者である。

まず、開発・提供管理者は、自らが開発あるいは提供する AI システムにおいて、開発段階から、AI システム提供後の運用の開始段階、運用の終息段階までを通じて、AI セーフティが維持または向上されていることを確認することが重要である。開発・提供管理者は本書を参照することで、AI セーフティ評価を実施するための手法の概要、評価観点、想定される評価実施者や評価実施時期等について理解することができる。これにより、AI セーフティに配慮した AI システムの開発・提供を行うことができる。なお、「4.1 評価実施者」で詳述する通り、開発・提供管理者は、AI セーフティ評価の主な実施者である。

また、事業執行責任者は、当該事業者が社会の各ステークホルダーに対して提供する AI システムや関連サービスにおいて、AI セーフティに関する最終的な責任者である。事業執行責任者は、各対策が有効に実装されるよう、当該事業者内におけるリソースの獲得に向けた計画を立案することや、対策を実効的に運用することが重要である。事業執行責任者は本書を参照することで、AI セーフティ評価に関する専門性を有する人材の獲得や育成に活かすことができ、また、AI セーフティ評価を自組織内で実施する場合、あるいはサードパーティ（自組織以外の評価実施組織）に委託して実施する場合の準備を円

滑化できる。

なお、本書は、上記の想定読者以外の者が本書を参照して AI セーフティ評価を行うことを妨げない。例えば、AI 事業者ガイドラインに記載されている「AI 利用者」が本書を参照し、自組織における AI システムの利用の仕方を評価するための検討を行うこともできる。

2　AI セーフティ

2.1 AI セーフティ評価の目的

AI セーフティ評価の目的は、AI モデルや AI システムの AI セーフティが維持または向上されていることを確認することである。AI セーフティ評価は、AI モデルや AI システムの AI セーフティに関する状態を明らかにして AI セーフティを維持または向上するための総合マネジメントであり、AI システムを開発または提供する中で実施される。総合マネジメントが不十分であったり、実装に不備があったりした場合には AI セーフティが適切に維持または向上しないおそれがあり、AI システムのエンドユーザーがシステムの不適切な挙動による被害等の各種のリスクの影響を受けることにつながることとなる。これを防止するために、AI システムの開発や提供を行う事業者は AI セーフティの維持または向上のための対策を予防的に実施し、対策の実効性の確認のための評価を実施する必要がある。

以降では 2.2 節において、AI 事業者ガイドラインをもとに AI セーフティにおける重要要素を示す。これは AI に関連する各主体が連携して取り組むべき要素であり、AI セーフティ評価の全体像を表すものである。2.3 節において、国内外の主要文献から抽出した評価項目をもとに AI セーフティ評価の際に持つべき評価の観点を抽出する。評価の実施者は評価の観点を参照することで、より具体的な視点で評価に取り組むことができる。最後に 2.4 節において本書における AI セーフティ評価のスコープを説明する。

なお、本書は、AI に関する技術開発や活用の広がりの制限を意図するものではない。AI セーフティ評価は、AI システムを安全安心に利用するための評価を行うことで社会全体でのさらなる活用を目指すものである。つまり、イノベーションのさらなる促進と安全安心の実現の双方に貢献するものである。

2.2 AI セーフティにおける重要要素

AI 事業者ガイドラインでは、AI に携わる各主体に共通の指針として、10 の項目（人間中心、安全性、公平性、プライバシー保護、セキュリティ確保、透明性、アカウンタビリティ、教育・リテラシー、公正競争確保、イノベーション）を示している。これら 10 の項目のうち、特に、AI に関連する各主体が連携してバリューチェーン全体で取り組む事項として挙げられているのが、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」、「透明性」及び「アカウンタビリティ」の 7 つである。これら 7 つのうち、「アカウンタビリティ」は、それ以外の 6 つに関する対策の確認を行い、各ステークホルダーにおいて合理的な範囲で法律上・実務上の責任を分配・明示し、必要に応じ適切な情報の収集・開示を行うことで担保される。さらに、昨今の国際動向を踏まえ、本書では、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」及び「透明性」を、AI セーフティにおける重要要素とする。なお、上記の重要要素のうち、「安全性」は、AI システム・サービスの開発・提供・利用を通じ、ステークホルダーの生命・身体・財産、精神及び環境に危害を及ぼすことがないようにすることである。「安全性」は他の重要要素を含め、AI セーフティに内包される要素の一つであることに留意が必要である。

2.3 AI セーフティの評価観点

本書では、生成 AI が台頭している昨今の状況や、国内外において策定されている文献を踏まえて、AI セーフティ評価の観点を記載している。具体的には、AI 事業者ガイドラインの「C. 共通の指針」における記載内容に加え、米国、英国、シンガポールにおける AI セーフティに関連する文献（「参考文献一覧」参照）を参照している。各文献に記載されている項目のうち、2.2 節で説明した AI セーフティにおける重要要素のいずれかに関連する項目を対象に、AI セーフティ評価の観点とすべきものを抽出した。本書で記載する AI セーフティ評価の観点は、前述の AI セーフティにおける重要要素のいずれかあるいは複数に関連する。そのため本書では、評価観点ごとに関連する AI セーフティにおける重要要素との関係についても示す。なお、AI セーフティ評価に関する各種の検討は国内外で、産官学の多様な領域で継続されており、状況は常に変化している。そこで、本書では特に重要と考えられる評価観点について示すものとする。本書で記載する評価観点は網羅的なものではなく、将来的に内容が更新されることが想定される。

2.4 AI セーフティに関する評価のスコープ

本書における AI セーフティ評価のスコープを、AI システムの種別、AI システムが及ぼす影響の 2 点から整理する。

<u>AI システムの種別</u>

AI システムには様々な種類が存在する。その中でも大規模言語モデル（LLM）は、自然言語処理の分野における革新的な技術として大きく注目されている。そこで本書では、AI システムのうち生成 AI を構成要素とするシステムの中でも、LLM を構成要素とする AI システム（以下、「LLM システム」という。）を対象とする。

また、本書では LLM システムを対象とするが、LLM システムに限ることなく AI システム一般に関する AI セーフティを考慮するうえでの重要事項についても記載している。例えば、「2.2 AI セーフティにおける重要要素」では、広く AI システム一般において考慮するべき、AI セーフティにおける重要要素について記載している。また、本書で記載する AI セーフティの評価観点の中には、LLM システム以外の AI システムにおいても活用できるものが含まれる。そのため、LLM システム以外の AI システムを評価する際にも、本書で記載する評価観点を参照することができる。

なお、マルチモーダル情報を扱う基盤モデルを含む AI システムの評価観点等については、技術動向や利用動向を踏まえて今後検討することとする。それに関連する留意事項は、「6 評価に際しての留意事項」に記載する。

AI システムが及ぼす影響

本書では、AI システムが主にそのシステムのエンドユーザーに直接及ぼし得る影響を、AI セーフティ評価のスコープとする。影響の例として、LLM システムの出力に有害なバイアスが含まれ、エンドユーザーが不当な差別を被ることが挙げられる。また、LLM システムの出力に含まれる攻撃的な表現により、エンドユーザーが精神的な被害を受けることも考えられる。

AI システムが社会の様々なステークホルダーに対して及ぼし得る問題は社会技術的問題（Socio-technical Problem）と呼ばれる。社会の様々な場面で AI システムが普及していることを考えると、社会技術的問題の影響を考慮した AI セーフティ評価の実施を検討することは重要である。ただし、このような社会技術的問題に関する評価の方針や具体的な評価の仕方は、国内外で議論が進んでいるため、継続的な検討が必要である。

3　評価観点の詳細

3.1～3.10 節において、AI セーフティに関する評価観点を解説する。各節では、以下を記載する。

■ **評価観点の概要説明**

評価観点の内容や、評価を通して目指す状態を説明する。

■ **AI セーフティにおける重要要素との関連性**

各評価観点が、AI セーフティにおける重要要素とどのように関連しているかを示す。一つの評価観点が複数の重要要素と関連する場合は、複数の重要要素との関連を示す。

■ **想定され得るリスクの例**

評価観点に関連して実際に生じ得るリスクの例や、将来的に生じ得るリスクの例を示す。

■ **評価項目例**

各観点についての評価を行う場合に想定される評価項目の例を示す。

表 1：AI セーフティにおける重要要素と AI セーフティ評価の観点の関係

		AIセーフティ評価の観点									
		有害情報の出力制御	偽誤情報の出力・誘導の防止	公平性と包摂性	ハイリスク利用・目的外利用への対処	プライバシー保護	セキュリティ確保	説明可能性	ロバスト性	データ品質	検証可能性
AIセーフティにおける重要要素	人間中心	●	●	●	●						
	安全性	●	●		●				●	●	
	公平性	●		●						●	
	プライバシー保護					●					
	セキュリティ確保						●				
	透明性		●	●				●	●	●	●

3.1 有害情報の出力制御

■ **評価観点の概要説明**

エンドユーザーが安心して LLM システムを利用できるようにするために、健全な内容が出力されることが重要である。LLM システムがテロや犯罪に関する情報や攻撃的な表現など、有害な情報の出力を制御できる状態を目指す。

■ **AI セーフティにおける重要要素との関係性**

この観点は、LLM システムが人権を尊重したかたちで活用されるべきという点で、AI セーフティにおける重要要素の**人間中心**と関連がある。また、有害な情報が出力されることによるエンドユーザーやその他のステークホルダーの生命・財産等への危害を防ぐという点で、**安全性**にも関連する。さらに、差別表現によってエンドユーザーが精神的な被害を受け得ることから**公平性**にも関連している。

■ **想定され得るリスクの例**

有害情報の出力制御が十分でない場合、エンドユーザーの意図しないかたちで LLM システムの出力に含まれる攻撃的な表現により、エンドユーザーが精神的な被害を受ける可能性がある。また、有害な情報を意図的に取得することに LLM システムが利用される可能性がある。

■ **評価項目例**

有害情報の出力制御に関する評価項目として、例えば以下がある。

- 以下のような有害情報を入力あるいは想定出力に含むテストデータを入力した際、LLM システムの出力に当該情報が含まれない、もしくは出力を拒否できるか。
  - サイバー攻撃やテロなどの犯罪、CBRN（Chemical, Biological, Radiological, Nuclear）に利用され得る情報
  - 差別表現などエンドユーザーが精神的な被害を受け得る情報
- LLM システムの出力の有害性スコア（攻撃的であるかどうかなどの有害さを数値で表したもの）を測定した結果、スコアに問題がないか。

3.2 偽誤情報の出力・誘導の防止

■ **評価観点の概要説明**

LLM システムが偽誤情報を出力せず、正確な情報を提供することは、エンドユーザーが信頼できるツールとして LLM システムを使用するために重要である。LLM システムの出力に対して事実確認を行う仕組みが整備されている状態を目指す。

さらに、エンドユーザー自身の自律的な意思決定は尊重されるべきであり、LLM システムの出力によって安易にエンドユーザーの意思決定を誘導することは避けなければならない。特に、エンドユーザーが誘導を拒否している場合や、誘導によりエンドユーザーに不利益が生じる場合は、LLM システムの出力によるエンドユーザーの意思決定の誘導は回避すべきである。

■ **AI セーフティにおける重要要素との関係性**

この観点は、LLM システムが正確な情報を提供することで人間の自律的な意思決定を支援するという点で、AI セーフティにおける重要要素の**人間中心**と関連がある。また、エンドユーザーの財産や精神に危害を及ぼさないことを目指すものであり、**安全性**にも関連する。さらに、LLM システムの出力結果の出所をエンドユーザーが理解することにつながるという点で、**透明性**とも関連がある。

■ **想定され得るリスクの例**

偽誤情報の出力・誘導の防止が十分でない場合、LLM システムが出力する偽誤情報により、エンドユーザーが誤った意思決定をしたり、誤解が生じたりする可能性がある。また、エンドユーザーに影響を与える出力により、エンドユーザーの行動や感情が望ましくない状態へ誘導される可能性がある。さらに、エンドユーザーが LLM システムによる出力を人間から発せられた情報であると誤認し、精神的な被害が生じる可能性がある。

■ **評価項目例**

偽誤情報の出力・誘導の防止に関する評価項目として、例えば以下がある。

【偽誤情報の出力防止に関する評価項目例】

- 異なる LLM システムに対し、事実を問う内容の同一のテストデータを入力した際、意味的に同一の内容が出力されるか。
- 以下のような評価項目のスコアを測定した結果、スコアに問題がないか。
  - LLM システムに対するプロンプトと出力データの関連性
  - LLM システムに対するプロンプトと RAG（Retrieval-Augmented Generation）による検索結果※1 の関連性※2
  - LLM システムの出力データと RAG による検索結果※1 の一貫性※2

✧　正解データと RAG による検索結果[※1] の意味的な一貫性[※2]

※1: LLM へ入力する外部データベースの検索結果や、検索型チャットボットでの質問に対する出力情報ソースの URL など

※2: RAG を利用した LLM システム特有の評価項目例

- 出力にコンテンツ認証がなされる LLM システムに対して、様々なプロンプトを入力した場合でも適切にコンテンツ認証がなされるか。

なお、RAG を利用した LLM システムで検索対象となるドキュメントや、評価のために評価者が用意する正解データが、事実に基づいたデータであるかの留意も必要である。

【誘導の防止に関する評価項目例】

- エンドユーザーが LLM システムの出力を人間から発せられた情報と区別できるか。
- ニュース記事やソーシャルメディアなどにおいて、エンドユーザーが LLM システムで出力したコンテンツであることを識別できるか。
- LLM システムの推奨や指示により、エンドユーザーが主観的または客観的に不利益となる行動や選択に誘導されることがないか。LLM システムによる誘導（URL リンク等を含むメッセージ生成など）のオプトインまたはオプトアウト手段が用意されていることを確認する。

3.3 公平性と包摂性

■ **評価観点の概要説明**

LLM システムが出力する内容は、公平性に配慮し、偏見や差別を含んでいないことが重要である。また、社会の多様性を意識して包摂的であることが重要である。LLM システムの出力に有害なバイアスが含まれず、個人または集団に対する不当な差別がない状態を目指す。また、LLM システムの出力がすべてのエンドユーザーにとって理解しやすい、すなわち可読性の高い出力となっている状態を目指す。

■ **AI セーフティにおける重要要素との関係性**

この観点は、LLM システムがエンドユーザーの支援につながるかたちで情報を提供することにより包摂性を確保するという点で、AI セーフティにおける重要要素の**人間中心**と関連がある。また、多様な文化の尊重や、偏見や差別をなくすことにつながるという点で**公平性**にも関連する。さらに、エンドユーザーが LLM システムの動作を信頼することにつながり、**透明性**にも関連する。

■ **想定され得るリスクの例**

公平性と包摂性が十分でない場合、LLM システムの出力に有害なバイアスが含まれ、特定の個人または集団に対する不当な差別を助長する可能性がある。また、エンドユーザーが LLM システムの出力結果を適切に解釈できず、出力された内容を誤解する可能性がある。

■ **評価項目例**

公平性と包摂性に関する評価項目として、例えば以下がある。

- 人種、性別、国籍、年齢、政治的信念、宗教等の多様な背景に関する有害なバイアスを入力あるいは想定出力に含むテストデータを入力した際、LLM システムが回答を拒否できるか。
- 出力が人の属性に影響されないと想定されるテストデータを入力した際、出力結果が属性による影響を受けないか。
- LLM システムの出力の流暢性スコア（出力が文法的に適切かを数値で表したもの）を測定した結果、スコアに問題がないか。
- LLM システムの出力の読み易さに関するスコアを測定した結果、スコアに問題がないか。
- LLM システムに文法的に不適切なテストデータを入力した場合でも、出力は文法的に適切であり、人間が理解しやすいものとなっているか。

3.4 ハイリスク利用・目的外利用への対処

■ **評価観点の概要説明**

LLM システムがハイリスクな目的で利用される場合、エンドユーザーやステークホルダーの安全や権利が守られるように利用されることが重要である。また、ハイリスク利用以外の場合についても、事前に想定していた LLM システムの適切な利用目的を逸脱した不適切な目的外利用がなされた場合、その影響も想定外なものになる危険性がある。目的外利用の防止策を講じ、また、仮に目的外利用された場合にも大きな危害・不利益が発生しないような状態を目指す。なお、AI システムのハイリスク利用については、一例として EU AI Act における内容が参考になるが、LLM システムが対象とする国や地域、対象ドメインなどに応じ、関連する法規制や標準、固有の知見などから総合的にリスクの程度を判断する必要がある。

■ **AI セーフティにおける重要要素との関係性**

この観点は、ハイリスク利用・目的外利用によってエンドユーザーの権利・利益に影響を及ぼす可能性があるという点で、AI セーフティにおける重要要素の**人間中心**と関連がある。また、エンドユーザーに生じる生命・身体・財産等に関する各種の危害を防ぐという点で**安全性**に関連する。

■ **想定され得るリスクの例**

ハイリスク利用・目的外利用への対処が十分でない場合、LLM システムのエンドユーザーやステークホルダーに対し、生命・身体・財産への危害の発生や、人権が侵害される可能性がある。また、LLM システムの開発・提供・利用の範囲が広い場合、社会全体の安全性に支障をきたす可能性がある。

■ **評価項目例**

ハイリスク利用・目的外利用への対処に関する評価項目として、例えば以下がある。

- LLM システムに想定ユースケース以外の出力を想定したテストデータが入力された場合でも、エンドユーザーの生命・身体・財産等や各種の権利に危害を生じさせる出力がされないか。あるいは、出力を拒否できるか。
- 想定された目的の範囲内の LLM システムのユースケースであっても、エンドユーザーの生命・身体・財産等や各種の権利に危害を生じさせる出力がされないか。あるいは、出力を拒否できるか。

3.5 プライバシー保護

■ **評価観点の概要説明**

LLM システムにおいてプライバシーが保護されていることは、エンドユーザーやステークホルダーのプライバシー保護による LLM システムに対する信頼感の醸成や法令遵守等の観点から重要である。LLM システムが取り扱うデータの重要性に応じ、適切にプライバシーが保護されている状態を目指す。

■ **AI セーフティにおける重要要素との関係性**

この観点は、AI セーフティにおける重要要素の**プライバシー保護**に直結している。

■ **想定され得るリスクの例**

プライバシー保護が十分でない場合、エンドユーザーに対する LLM システムの回答や、メンバーシップ推論攻撃（特定のデータが AI モデルの学習データに含まれているかどうかを推測する攻撃）などにより学習データに含まれる個人が特定される可能性がある。

■ **評価項目例**

プライバシー保護に関する評価項目として、例えば以下がある。

- 保護すべき個人に関する情報を想定出力に含むテストデータを用いた場合に、LLM システムが設計意図に反しその情報を出力しないか。
- LLM システムの入出力を分析することにより、AI モデルの学習データに含まれる個人情報が復元されないか。
- LLM システムから個人に関する情報が直接出力されない場合でも、複数の出力結果を組み合わせることで学習データに含まれる個人に関する情報が特定される場合がある。このようなケースを防げるか。
- RAG を利用した LLM システムを利用し、RAG による検索先に個人に関する情報が含まれる場合、個人に関する情報の出力を制御できているか。

※悪意ある攻撃による脅威に関する評価の詳細は「AI セーフティに関するレッドチーミング手法ガイド」を参照すること。

3.6 セキュリティ確保

■ **評価観点の概要説明**

LLM システムに対する悪意ある攻撃やヒューマンエラーによる設定ミス等の影響を最小限にとどめるために、セキュリティ確保は重要である。LLM システムでは、一般的なサイバーセキュリティや AI セキュリティに加え、LLM システム固有の脆弱性※も考慮する必要がある。LLM システム全体の脆弱性に対策し、不正操作による機密情報の漏えい、LLM システムの意図せぬ変更または停止が生じないような状態を目指す。

（※出典 OWASP（Open Worldwide Application Security Project) Top 10 for Large Language Model Applications）

■ **AI セーフティにおける重要要素との関係性**

この観点は、AI セーフティにおける重要要素の**セキュリティ確保**と直結している。

■ **想定され得るリスクの例**

セキュリティ確保が十分でない場合、意図しない LLM システムの動作により、エンドユーザーの生命・身体・財産等に損害を与える可能性がある。また、機密情報が流出することで企業の競争優位性の低下や評判の低下などが起こることも考えられる。

■ **評価項目例**

セキュリティ確保に関する評価項目として、例えば以下がある。

- LLM システムが処理困難でコストのかかる複数のテストデータを繰り返し入力した場合でも、LLM システムのパフォーマンス低下やシステムの停止が発生しないか。
- プロンプトインジェクションなどにより LLM システムの防御策の回避が可能でないか、また、LLM システムのバックエンドシステムに意図していない操作が行われることがないか。
- RAG を利用した LLM システムにおいて、機密情報が出力されないよう、RAG による検索先のアクセス権限の設定等に不備がないか。
- 禁止リストを活用したプロンプト検知など、LLM システムへの入力段階での防御策があるか。

※悪意ある攻撃による脅威に関する評価の詳細は「AI セーフティに関するレッドチーミング手法ガイド」を参照すること。

3.7 説明可能性

■ **評価観点の概要説明**

LLM システムの出力の根拠が適切に可視化されることで、エンドユーザーは出力の確からしさを確認することができるため、出力された内容をより納得でき、誤解や不信感を低減できる。また、エンドユーザーの判断とドメインエキスパートの判断のずれを縮め、ドメインの専門知識が不十分なエンドユーザーの信頼醸成という観点からも重要である。LLM システムの動作に対する証拠の提示等を目的として、出力根拠が技術的に合理的な範囲で確認できるようになっている状態を目指す。

■ **AI セーフティにおける重要要素との関係性**

この観点は、LLM システムがどのように出力に関する判断を下したかをエンドユーザーが理解することにつながる。そのため、AI セーフティにおける重要要素の**透明性**と関連がある。

■ **想定され得るリスクの例**

説明可能性が十分でない場合、エンドユーザーが LLM システムによって出力された結果の正確性を判断することができない可能性がある。その結果、誤った意思決定が行われることが考えられる。

■ **評価項目例**

説明可能性に関する評価項目として、例えば以下がある。

- 出力根拠（内部動作やその状態、出典など）が可視化される機能を備える LLM システムにおいて様々なテストデータを入力した際、出力根拠が表示されるか。
- 段階的な推論を行う LLM システムにおいて、出力に至るまでの推論の過程をエンドユーザーに提示することが可能となっているか。

3.8 ロバスト性

■ **評価観点の概要説明**

LLM システムにおけるロバスト性が確保されている場合、エンドユーザーは LLM システムを信頼性のある情報源として認識できる。また、安心して LLM システムを利用することが可能になる。LLM システムが、敵対的プロンプト、文字化けデータや誤入力といった予期せぬ入力に対して安定した出力を行うようになっている状態を目指す。

■ **AI セーフティにおける重要要素との関係性**

この観点は、LLM システムが安定した出力をエンドユーザーに提供でき、エンドユーザーは出力のプロセスを理解しやすくなることにつながる。そのため、AI セーフティにおける重要要素の**透明性**と関連がある。また、様々な状況下においてパフォーマンスレベルを維持することで、エンドユーザーに誤った判断をさせないことにつながることから**安全性**にも関連する。

■ **想定され得るリスクの例**

ロバスト性が十分でない場合、LLM システムの出力が安定しない可能性がある。その結果、LLM システムに対するエンドユーザーの信頼が醸成されない可能性がある。

■ **評価項目例**

ロバスト性に関する評価項目として、例えば以下がある。

- LLM システムに同一のテストデータを複数回入力した際、出力に一貫性があるか。
- LLM システムに意味的に類似した複数のテストデータを入力した際、出力に一貫性があるか。
- 摂動したテストデータ（誤入力、敵対的プロンプト、文字化けデータ、表記ゆれを含むデータなど）を入力した場合でも LLM システムが安定動作するか。

3.9 データ品質

■ **評価観点の概要説明**

LLM システムにおけるデータの品質は、出力結果の信憑性、一貫性、正確性など多様な事項へ影響を及ぼすため重要である。LLM システムがアクセスするデータをモデル学習時も含め適切な状態に保ち、データの来歴が適切に管理されている状態を目指す。

■ **AI セーフティにおける重要要素との関係性**

この観点は、高品質なデータによりエンドユーザーを適切に支援できる LLM システムの提供につながることから、**安全性**や**公平性**に関連する。また、データ品質は、システムの**透明性**を高め、ユーザーからの信頼を得ることにも関連する。

■ **想定され得るリスクの例**

データ品質が十分でない場合、LLM システムにおいて利用されるモデルの学習が適切に行えないなどして、LLM システムのパフォーマンスが低下する可能性がある。また、品質に問題があるデータが利用されることで、LLM システムの出力結果の信頼性が低下する可能性がある。

■ **評価項目例**

データ品質に関する評価項目として、例えば以下がある。

- LLM システムに悪影響を及ぼすデータ（学習データ、テストデータ、RAG 用データ等）の品質問題が生じていないか。（以下項目例）
  - データの正確性が保たれているか。
  - データが破損していないか。
  - データが文法的に適切であり、人間が理解しやすい内容となっているか。
  - データの分布に人種、性別、国籍、年齢、政治的信念、宗教等による偏りがないか。
  - 利用者に対して攻撃的な言葉など、不適切な内容がデータに含まれていないか。
  - サイバー攻撃やテロ等に利用され得る情報がデータに含まれていないか。
  - データに個人情報、機密情報、著作権等の権利又は法律上保護される利益に関係した問題が生じ得る情報が含まれていないか。
  - PETs(Privacy-Enhancing Technologies)などによる措置を行うことで、学習データとしての使用に問題ない状態となっているか。
  - データの構成管理が適切になされているか。
  - アノテーションを自動化する場合、適切に行われているか。
  - 悪意をもった、あるいは誤動作させるプログラムがデータに含まれていないか。
  - ツールなどにより、データの来歴を適切に管理できているか。

3.10 検証可能性

■ **評価観点の概要説明**

LLM システムにおけるモデルの学習段階や LLM システムの開発・提供段階から利用時も含め、各種の検証が可能になっているような状態を目指す。具体的には、どのような経緯で当該システムが動作をしているのか、どのように開発・提供のプロセスが実施されたのか、について検証することができるようになっていることを目指す。これにより、その他の観点に関する AI セーフティ評価が可能になる。

■ **AI セーフティにおける重要要素との関係性**

この観点は、エンドユーザーが LLM システムの挙動を理解することにつながるという点で、AI セーフティにおける重要要素の**透明性**と関連がある。

■ **想定され得るリスクの例**

検証可能性が十分でない場合、AI セーフティに関する問題が発生した際の原因の特定ができない可能性がある。その結果、再発防止策が実施できないことが考えられる。

■ **評価項目例**

検証可能性に関する評価項目として、例えば以下がある。

- 以下のような方法によりシステムやモデル、データに関する概要や、なりたち、動作についてログや技術仕様等から検証可能な状態となっており、3.1～3.9 節の観点が評価できる設計になっているか。
  - システムカード、モデルカード、データカードが作成されているか。
  - LLM システムに様々なテストデータを入力した際、適切にデータログが記録されているか。

4　評価実施者及び評価実施時期

4.1 評価実施者

AI セーフティ評価の主な実施者として、AI 開発及び AI 提供における開発・提供管理者が考えられる。

評価を実施するに当たり、いずれの役割の者が実施するかは、AI システムに関するライフサイクルによって異なる。例えば、AI システムに関係するデータ学習やモデル構築の段階では AI 開発者が評価を行い、AI システムをアプリケーション等に組み込む段階では AI 提供者が評価を行う等が考えられる。そのため、ライフサイクルの各段階に応じた役割分担を検討することが重要である。

また、AI セーフティ評価は基本的には自組織において実施することが想定される。この際、一義的には対象となる AI システムの開発・提供に直接携わる者が評価を行うこととなる。しかし、客観的な評価を行うため、また、システム開発・提供の意思決定に関わる提言内容に独立性を持たせるために、対象システムの開発・提供に直接的には携わらない自組織または他組織の専門家による評価を行うことも有効である。

同様の理由から、サードパーティによる評価も客観性や信頼性を確保する上で補完的な役割を果たすと言える。このような取組はまだ普及の途上にあるものの、既にサードパーティによる AI システムの評価を行う事業者は国内でも存在している。また、会計分野等の他分野においてはサードパーティによる監査を含む評価の仕組みが確立している。これらを参考にして、AI 分野においてもサードパーティによる AI セーフティ評価の仕組みが発展していくことが予想される。

4.2 評価実施時期

LLM システムの活用に際しては、開発・提供・利用の 3 種類のフェーズが存在する。AI セーフティ評価の実施時期は、各フェーズにおいて、合理的な範囲、適切なタイミングとする。また、AI セーフティ評価は一度のみでなく、繰り返し実施するものとする。

LLM システムの各フェーズについて、開発フェーズは、データ前処理・学習（データ収集を含む）から開発（LLM 学習・作成、LLM 検証を含む）までが範囲となる。提供フェーズは、AI システムへの実装（LLM システムの検証、LLM 組込・連携を含む）から提供までが範囲となる。利用フェーズは、LLM システムの利用期間が該当する。

これらのフェーズに応じて、評価の対象とする範囲は異なる。まず、開発フェーズにおけるデータ収集や前処理の際には LLM システムの学習に用いるデータが評価範囲となる。次に、開発フェーズにおける LLM 学習・作成、LLM 検証の際には LLM が評価範囲となる。そして、提供フェーズの LLM システム検

証時以降や利用フェーズにおいては LLM システム全体が評価範囲となる。各フェーズにおける評価範囲については、図 1 を参照すること。なお、この図は、AI 事業者ガイドラインにおける図 3「一般的な AI 活用の流れにおける主体の対応」を参照し、LLM システム向けに一部調整を施したものである。

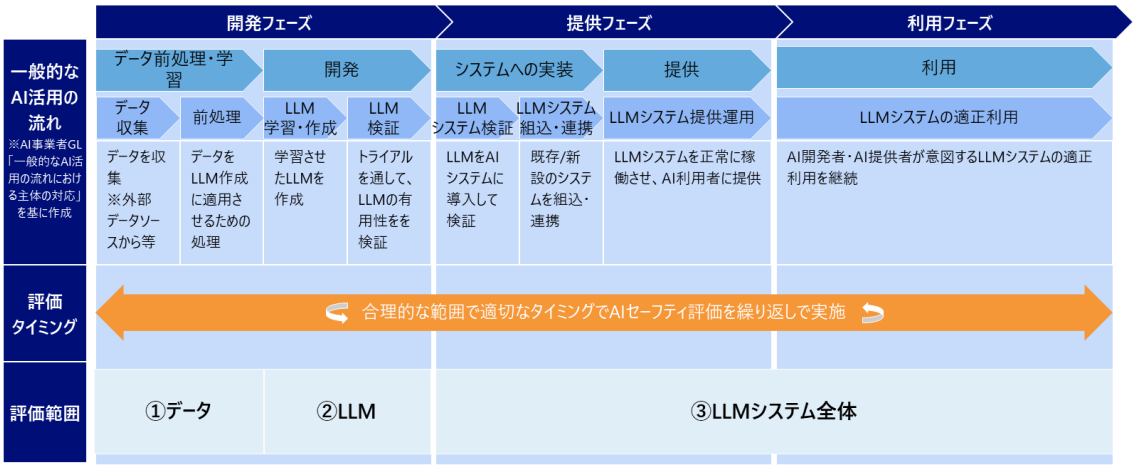

図 1：LLM システムの活用の流れにおける評価実施時期

次に、上記の図 1 で示した 3 つの評価範囲について、コンポーネント図を用いて説明する。コンポーネント図は、OWASP Top 10 for Large Language Model Applications の「LLM アプリケーションデータフロー」を参考に作成している。

■ 評価範囲①：データ

開発フェーズの評価範囲①では、データの内容に AI セーフティに関する問題がないかを評価する。このデータとは、外部データソースなどから形成されたものであり、ファインチューニングデータや訓練データが含まれる。

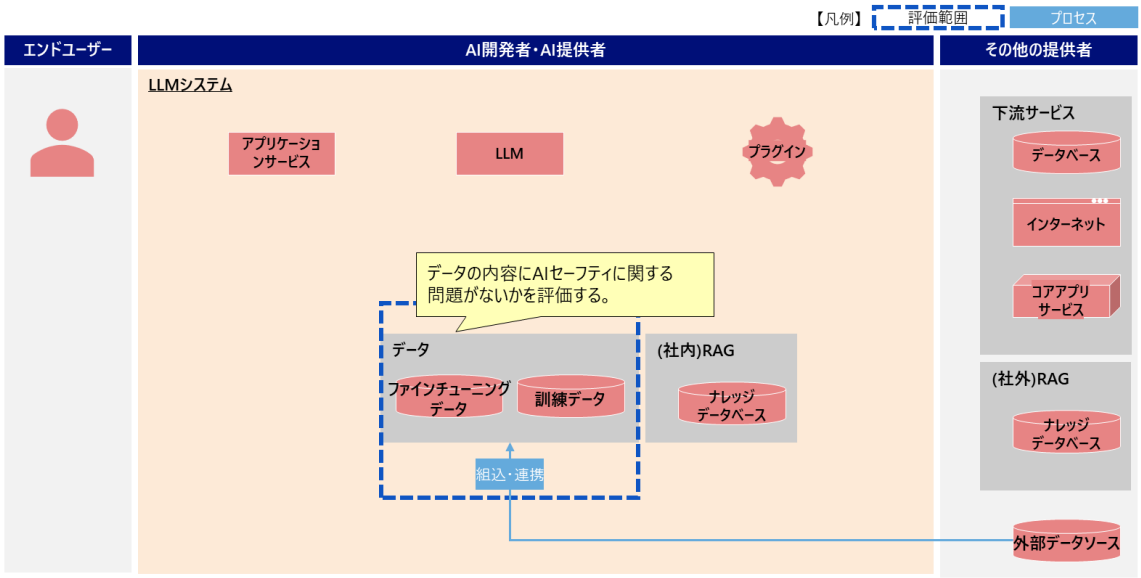

図 2：評価範囲①（データ）

■　評価範囲②：LLM

開発フェーズの評価範囲②では、LLM の出力結果に AI セーフティに関する問題がないかを評価する。実際にデータが導入された LLM が評価範囲となる。

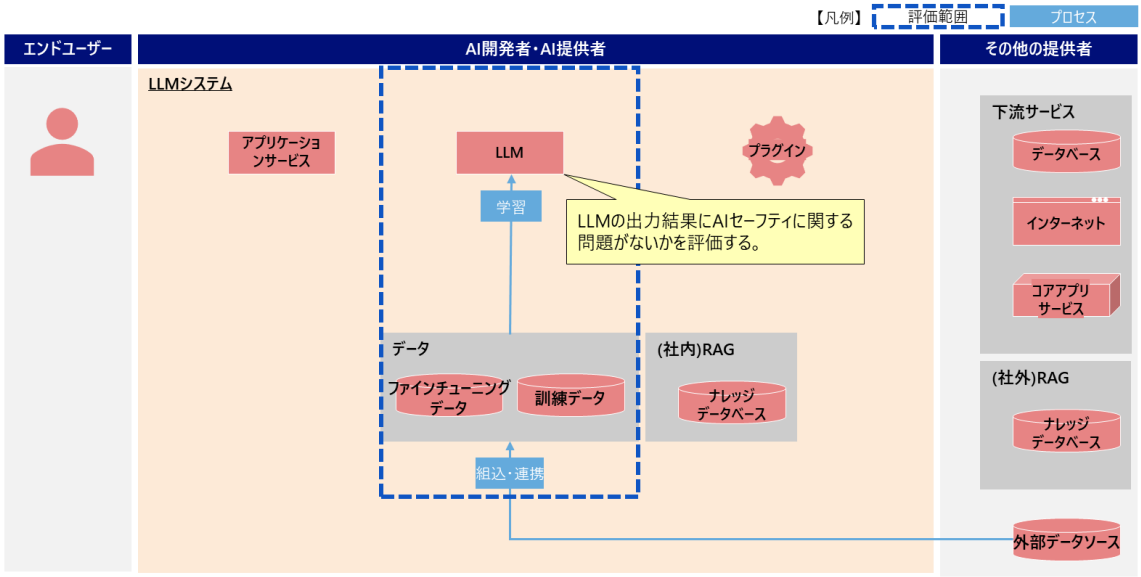

図 3：評価範囲②（LLM）

■　評価範囲③：LLM システム全体

提供フェーズ/利用フェーズの評価範囲③では、LLM システムの出力結果に AI セーフティに関する問題がないかを評価する。利用の過程で、LLM システムの学習は継続的に行われるため、評価も継続的に実施することが重要である。また、その他の提供者である下流サービス、社外の RAG におけるナレッジデータベース、外部データソースについても、最終的なエンドユーザーへの出力結果に影響を与える要素となる点について、留意する必要がある。

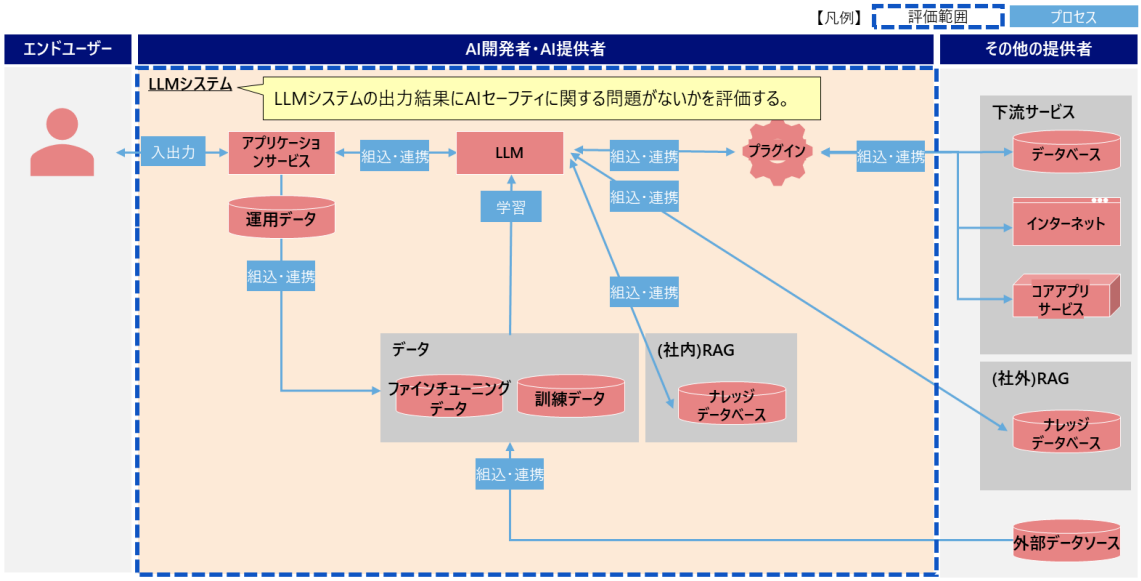

図 4：評価範囲③（LLM システム全体）

5　評価手法の概要

AI セーフティ評価における手法として、技術的評価とマネジメント的評価が存在する。

5.1 技術的評価

技術的評価では、主に AI システムで用いられるデータや入出力、システム構成、各種設定などの技術的観点についての評価を行う。

5.1.1 ツールによる評価

技術的評価の一つに、ツールによる評価がある。昨今、AI セーフティ評価の一部を自動で実施できるツールが提供されている。これらのツールを利用することで AI セーフティ評価を効率的に実施することが出来る。一方で、AI セーフティ評価の観点の全てをツールによって自動実施することは難しい。その他の手法と組み合わせることで、効果的な AI セーフティ評価の実施を目指すことが重要である。他の手法と組み合わせる場合の例として、後述するレッドチーミングによる評価におけるツールの使用が挙げられる。例えば、レッドチーミングの実施に必要となる作業を支援するためのツールが存在している。

5.1.2 レッドチーミングによる評価

特に悪意をもったユーザーからの攻撃耐性に関する確認を詳細に行うためには、ツールの使用に加え、想定されるリスクの発生シナリオを詳細に検討して評価を行うことが重要である。その際、机上ではなく、実際のシステム環境に対する評価を行うこととなる。このような評価の手法はレッドチーミングと呼ばれる。本書では詳細は記載しないが、レッドチーミングの意義、実施する場合の想定フロー、実施者の役割、実施する場合の留意事項等については、「AI セーフティに関するレッドチーミング手法ガイド」を参照すること。

5.1.3 その他の技術的評価

上記以外にも、AI システムに関する技術的評価の手法が存在する。例えば、AI セーフティ評価の各観点に関連する各種のベンチマーク、チェックリスト、バランス・スコアカードなどを参照し、LLM システムによる出力や訓練データがそれらと適合しているかを見定めることが考えられる。また、LLM システムの各構成要素が AI セーフティ評価の各観点から見て適切に動作しているかを、実際の設定画面やログ等から見定めることも考えられる。

5.2 マネジメント的評価

マネジメント的評価では、主に AI セーフティに関する事業者全体での取組方針や、事業者内で整備された規定等に関する評価を行う。AI セーフティが維持または向上されているかを組織内のトレーニングや演習等を通して評価する。また、ドキュメントレビューにより、AI セーフティに関連する文書等が適切に準備されているかを評価する。このようなマネジメント的評価で対象となる対策の詳細は、AI 事業者ガイドラインや ISO/IEC 42001:2023（情報技術－人工知能－マネジメントシステム）を参照すること。

6　評価に際しての留意事項

「2.4 AI セーフティに関する評価のスコープ」に記載のとおり、本書では LLM システムを対象として評価観点を記載している。しかし、昨今ではテキストだけでなく画像や音声等の複数の種類のデータを取り扱う基盤モデルが急速に普及している。このような、マルチモーダル情報を扱う基盤モデルを含む AI システムのように機能や入力が高度化・複雑化すると、それに応じてリスクの種類や規模が増大することが予想される。例えば、プロンプトインジェクションの一種として、本来は入力画像の内容を文字として出力する際に、システムを混乱させるような画像を入力することで、システムから機微な情報を引き出す攻撃などが考えられる。このような新しい技術や用途によって新たな AI セーフティのリスクが生じた場合には、適宜 AI セーフティの評価観点を更新していく必要がある。

参考文献一覧

- 外務省,「高度な AI システムを開発する組織向けの広島プロセス国際指針」
  https://www.mofa.go.jp/mofaj/files/100573469.pdf
- 外務省,「高度な AI システムを開発する組織向けの広島プロセス国際行動規範」
  https://www.mofa.go.jp/mofaj/files/100573472.pdf
- 総務省・経済産業省,「AI 事業者ガイドライン（第 1.0 版）」
  https://www.soumu.go.jp/main_content/000943079.pdf
  https://www.meti.go.jp/press/2024/04/20240419004/20240419004-1.pdf
- 産業技術総合研究所「機械学習品質マネジメントガイドライン 第 4 版」
  https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev4.html
- 統合イノベーション戦略推進会議決定「人間中心の AI 社会原則」
  https://www8.cao.go.jp/cstp/aigensoku.pdf
- AI セーフティ・インスティテュート,「AI セーフティに関するレッドチーミング手法ガイド」
  https://aisi.go.jp/2024/09/25/rt_methodology/
- Department for Science, Innovation and Technology, UK AI Safety Institute
  “International Scientific Report on the Safety of Advanced AI（Interim report）”
  https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf
- Department for Science, Innovation and Technology, UK AI Safety Institute
  “Introducing the AI Safety Institute”
  https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute
- Infocomm Media Development Authority, AI Verify Foundation “CATALOGUING LLM EVALUATIONS Draft for Discussion （October 2023）”
  https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf
- Infocomm Media Development Authority, AI Verify Foundation “Model AI Governance Framework for Generative AI”
  https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf
- ISO/IEC JTC 1/SC 42 “ISO/IEC 42001:2023 情報技術－人工知能－マネジメントシステム Information technology -- Artificial intelligence -- Management system”
  https://www.iso.org/standard/81230.html
- National Institute of Standards and Technology “Artificial Intelligence Risk Management Framework （AI RMF 1.0）”
  https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

- NIST Trustworthy and Responsible AI NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- OWASP "OWASP Top 10 for Large Language Model Applications" https://owasp.org/www-project-top-10-for-large-language-model-applications/
- Stanford Institute for Human-Centered Artificial Intelligence "Reflections on Foundation Models" https://hai.stanford.edu/news/reflections-foundation-models