AIセーフティに関するレッドチーミング手法ガイド（第1.00版）概要

AIセーフティ・インスティテュート

（令和6年9月25日）

AISI Japan AI Safety Institute

目次

1. 本書の背景・目的

AIシステムの急速な普及が進む中、AIセーフティの重要性は増している。
レッドチーミング手法に関する基本的な考慮事項を示すために、本書を作成した。

AISI Japan AI Safety Institute

背景

- AIシステムの開発・提供・利用が促進される中で、イノベーションの促進や社会課題の解決が期待されている一方、AIシステムの悪用や誤用、不正確な出力による懸念等が生じている。
- いわゆるAIセーフティについての関心が国内外で高まりつつあり、**AIセーフティ評価の一環として、特にレッドチーミング手法の検討が各国で進んできている**。

目的

- 「AIセーフティに関するレッドチーミング手法ガイド」（以下「本書」とする。）は、AIシステムの開発や提供に携わる者が、**対象のAIシステムに施したリスクへの対策を、攻撃者の視点から評価するためのレッドチーミング手法に関する基本的な考慮事項**を示す。
- 国内外における検討や先行事例を勘案し、国際整合性を考慮した上で、現段階でレッドチーミングを実行する際に重要と思われる事項を示す。

AIセーフティとは「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」。

※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：https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

2. 本書の作成方針

AI事業者ガイドラインに加え、国内外の文献や関連事業者等に関する調査を踏まえて作成した。

AISI Japan AI Safety Institute

- 「AI事業者ガイドライン」に加え、国内外の関連文献や、レッドチーミングに関連する事業者あるいはツールに関する調査結果を踏まえ、本書を作成した。なお、AIセーフティ評価の全般的な考え方は「AIセーフティに関する評価観点ガイド」に記載している。

AI事業者ガイドライン（日本）

近年の急速な技術変化に対応するために、既存の日本国内における関連ガイドラインを統合・更新して作成されたガイドライン。

国内外文献

機械学習品質マネジメントガイドライン　第4版（産業技術総合研究所）

機械学習を用いたAIシステムの品質要件をトップダウンに分類・整理し、対象システムにかかわるステークホルダーが品質を客観的に評価できる枠組みを構築できるようにするガイドライン。

LLM AI サイバーセキュリティとガバナンスのチェックリスト（Open Worldwide Application Security Project（OWASP））

組織内でのAIシステム提供・利用におけるサイバーセキュリティリスクの管理について記載している。LLMアプリケーションで見られる重大な脆弱性トップ10のリストも掲載している。

SP800-115（National Institute of Standards and Technology（NIST））

情報システムのセキュリティテストと評価に関する包括的なガイドライン。

AI 800-1（Initial Public Draft）（National Institute of Standards and Technology（NIST））

デュアルユース基盤モデルのリスクマネジメントに関する指針案。

AIセーフティに関するレッドチーミング手法ガイド

関連ツール（組織）

Anthropic

様々な観点からレッドチーミングを実施している旨を公表しており、レッドチーミングに利用するデータセットを公開している。

Microsoft

自社のサービスに対するレッドチーミングを実施しており、レッドチーミングに関するガイドを公開している。

NVIDIA

自社の分野横断的なチームでレッドチーミングを実施している。

OpenAI

自社のAIモデルへのレッドチーミングを行う専門家を募り、自社製品の安全性強化に取り組んでいる。

Project Moonshot（AI Verify Foundation）

シンガポールのAI Verify Foundationが開発したオープンソースのツールであり、レッドチーミング実施を支援する機能を持つ。

3. 本書の構成

AIセーフティに関するレッドチーミングを実行するうえで重要と思われる事項を種別毎に分類した。読者が参照しやすいよう目次を構成し、各分類に関する項目を記載した。

AISI Japan AI Safety Institute

- 5W1Hの視点で整理した項目に基づき、本書の各目次内容を記載した。
- 主な想定読者はAI開発者・AI提供者のうち、レッドチーミングの企画・実施に関与する者である。

種別	記載項目の例
What （レッドチーミングとは何か）	➢「レッドチーミング」の定義やスコープ ➢ 本書が対象とするAIシステム
Why （なぜレッドチーミングを実施するか）	➢ レッドチーミングの目的 ➢ レッドチーミングの重要性・期待される効果
Who （誰がレッドチーミングを実施するか）	➢ どのような役割の者がレッドチーミングを実施するか
When （いつレッドチーミングを実施するか）	➢ レッドチーミングの実施時期
Where （どこでレッドチーミングを実施するか）	➢ 自組織が実施するか、第三者（サードパーティ）が実施するか
How （どのようにレッドチーミングを実施するか）	➢ レッドチーミングの実施計画の立て方や、実施する際の準備事項 ➢ レッドチーミング実施に際して想定する脅威

想定読者

AI開発者・AI提供者

開発・提供管理者

事業執行責任者

※左記のうち、レッドチーミングの企画・実施に関与する者が想定読者。

AIセーフティに関するレッドチーミング手法ガイド【目次】	
1	はじめに
2	レッドチーミングについて
3	LLMシステムへの代表的な攻撃手法
4	実施体制と役割
5	実施時期及び実施工程
6	実施計画の策定と実施準備
7	攻撃計画・実施
8	結果のとりまとめと改善計画の策定
A	付録

4. レッドチーミングのスコープ

本書では、攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性を確認する評価手法をレッドチーミングとする。また、大規模言語モデル（LLM）を構成要素とするAIシステム（LLMシステム）を対象として記載した。

AISI Japan AI Safety Institute

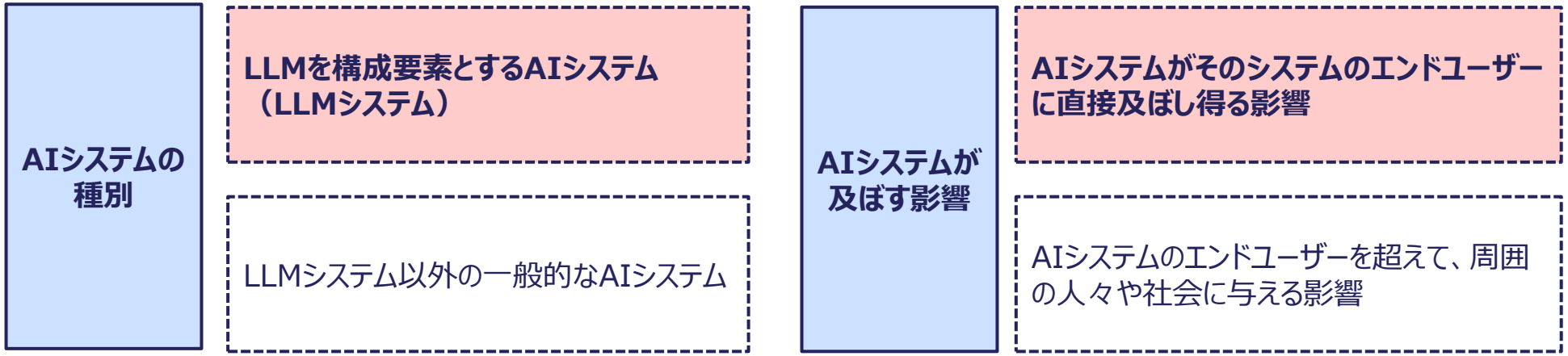
本書におけるレッドチーミングのスコープ

- **レッドチーミング**とは、「**攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性を確認する評価手法**」である。本書では、AIセーフティに関するレッドチーミングを単に「レッドチーミング」と呼ぶ。

- 本書におけるレッドチーミングのスコープを、**（1）AIシステムの種別**、**（2）AIシステムが及ぼす影響**の2点から整理した。（レッドチーミングはAIセーフティ評価手法の1つであるため、「AIセーフティにおける評価観点ガイド」におけるスコープと同様）

本書におけるレッドチーミングのスコープ

※下図色付きが本書でのレッドチーミングのスコープ



レッドチーミングは「攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性を確認する評価手法」であり、AIセーフティ評価の手法の一つである。

AISI Japan AI Safety Institute

- レッドチーミングは、攻撃者の目線で対象AIシステムにおける弱点や対策の不備を発見し、これらを修正及び堅牢化することを目的とする。

レッドチーミングの種類

➢ レッドチーミングは以下のように分類できる。

攻撃計画・実施者が保有する前提知識の有無・程度による分類

- **ブラックボックステスト**
  （内部構造等の情報を未知としてレッドチーミングを行う）
- **ホワイトボックステスト**
  （内部構造等の情報を既知としてレッドチーミングを行う）
- **グレーボックステスト**
  （内部構造等の情報を一部既知としてレッドチーミングを行う）

レッドチーミングを実施する環境による分類

- **実運用環境**
  （AIシステムが実際に実用に供される運用環境）
- **ステージング環境**
  （実運用環境とほぼ同様の状態でテストや不具合のチェック等を行う環境）
- **開発環境**
  （AIシステムの開発を行う環境）

レッドチーミング実施において攻撃シグネチャを試行する方法による分類

- **自動化ツールによるレッドチーミング**
- **手動によるレッドチーミング**
- **AIエージェントを用いたレッドチーミング**

LLMシステムへの代表的な攻撃手法

➢ LLMシステムへの代表的な攻撃手法例として、下記が存在する。これらを念頭に置いてレッドチーミングを行うのが望ましい。

■ **直接プロンプトインジェクション**
攻撃者が、悪意あるプロンプトをAIシステムに直接注入する攻撃

■ **間接プロンプトインジェクション**
攻撃者が、悪意あるプロンプトをAIシステムに間接的に注入する攻撃

■ **プロンプトリーキング**
攻撃者が、設定されたシステムプロンプトを引き出す攻撃

■ **ポイズニング攻撃**
攻撃者が細工したデータ・モデルを、訓練時に利用するデータ・モデルに紛れ込ませる攻撃

■ **回避攻撃**
AIシステムへの入力に悪意ある変更を加え、意図していない動作を引き起こす攻撃

■ **モデル抽出攻撃**
入出力の分析により、対象システムのモデルと同等の性能を持つモデルを作成する攻撃

■ **メンバーシップ推論攻撃**
入出力の分析により、あるデータが訓練データに含まれるかを特定する攻撃

■ **モデルインバージョン攻撃**
入出力の分析により、訓練データに含まれる情報を復元する攻撃

5. レッドチーミングの概要

レッドチーミング実施に際しては、多様な関係者（攻撃シナリオの実施によって影響を受けるシステムに関わる組織）が参画するのが望ましい。また、AIシステムのリリース/運用開始前に加え、運用開始後も、必要に応じて随時実施することが望ましい。

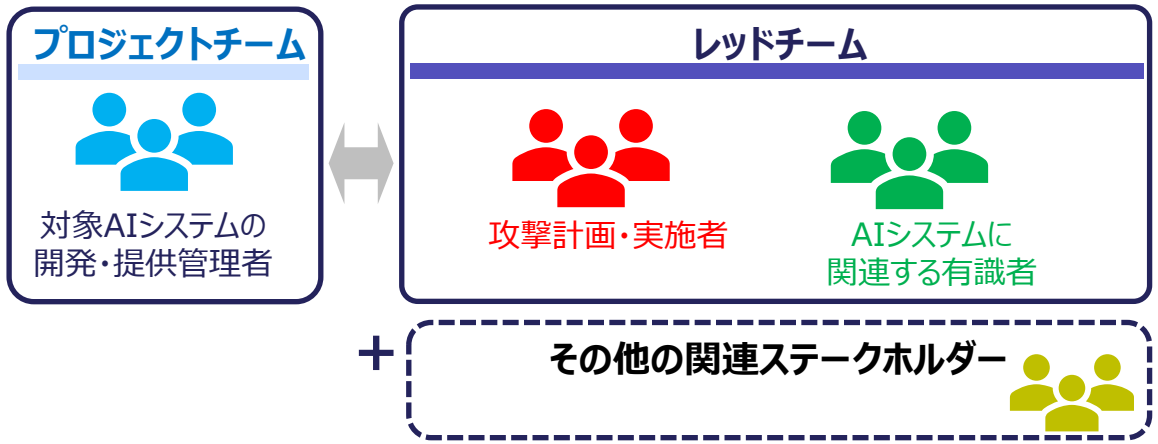
AISI Japan AI Safety Institute

実施体制と役割

- レッドチーム※は、レッドチーミングの実施対象となるAIシステムの開発・提供に携わるプロジェクトチームと連携する形で設置し、リーダまたは責任者を任命する。
- レッドチームには、「攻撃計画・実施者」及び「AIシステムに関連する有識者」が含まれることを基本として想定する。
- 必要に応じて、組織内のその他ステークホルダーの関与も考えられる。

※ 攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性の確認を担当するチーム

レッドチーミングの体制



実施時期

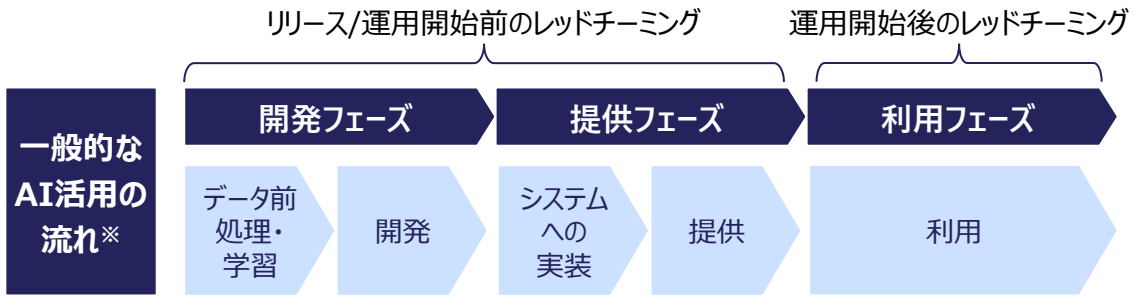
リリース/運用開始前のレッドチーミング

- レッドチーミングを初回実施する際は、対象とするAIシステムのリリース/運用開始前までに実施することを基本とする。
- ただし、対象とするAIシステムの規模や複雑性等に応じ、システムのコンポートネントやシステムレイヤ等の単位におけるリスク分析を行い、適切なタイミングで分割してレッドチーミングを実施することが有効な場合もある。

運用開始後のレッドチーミング

- レッドチーミングは一度実施して完了ではない。必要に応じて随時実施することが望ましい。

AI活用の流れにおけるレッドチーミング実施時期



※ AI事業者ガイドライン「一般的なAI活用の流れにおける主体の対応」参照

6. レッドチーミングに関する工程

レッドチーミングの工程は、「実施計画の策定と実施準備」、「攻撃計画・実施」、「結果のとりまとめと改善計画の策定」の3つから構成される。

工程	実施事項	ガイド本文での該当章
第１工程： **実施計画の策定と実施準備**	✓ 実施の決定とレッドチーム発足 ✓ 予算及びリソースの識別・確保とサードパーティの選定・契約 ✓ 実施計画の策定 ✓ 実施環境の準備 ✓ エスカレーションフローの確認	6章
第２工程： **攻撃計画・実施**	✓ リスクシナリオの作成 ✓ 攻撃シナリオの作成 ✓ 攻撃シナリオの実施 ✓ 実施中の記録取得 ✓ 実施後の処理	7章
第３工程： **結果のとりまとめと改善計画の策定**	✓ 実施結果の分析 ✓ 結果報告書の作成と関係者レビュー ✓ 最終報告書の作成と報告 ✓ 改善計画の策定と実施 ✓ 改善後のフォローアップ	8章

6. レッドチーミングに関する工程（実施計画の策定と実施準備）

「実施計画の策定と実施準備」では、レッドチームを発足させたうえで実施計画を策定し、レッドチーミング実施に必要な事前準備を行う。

AISI Japan AI Safety Institute

第１工程 ： 実施計画の策定と実施準備

プロジェクトチーム

STEP 1
実施の決定とレッドチーム発足

- レッドチーミング実施に関する企画書をとりまとめ、実施決定を行う。
- 企画書に記載されたレッドチームを発足させる。

STEP 2
予算及びリソースの識別・確保とサードパーティの選定・契約

- 予算の確保及び実施体制を確定し、必要な人員のアサインを行う。
- 必要なツール等のリソースを識別し、確保する。
- 組織内に十分な実施体制を確保できないと見込まれる場合には、攻撃計画・実施者としてサードパーティを活用する。

レッドチーム

STEP 3
実施計画の策定

- 対象となるＡＩシステムの概要把握など、実施のために必要となる事項を検討したうえで「レッドチーミング実施計画書」を作成し、その他の関連ステークホルダーと連携を図る。

STEP 4
実施環境の準備

- レッドチーミングに必要な実施環境の準備を行い、関係者に事前連絡する。
- 必要に応じ、事前に関係者に周知の上、監視設定の一時的解除や監視対象からの除外、アラートの無視等の依頼を行う。

STEP 5
エスカレーションフローの確認

- レッドチーミング実施によって、予期しない動作や障害・トラブル等が発生した場合に備えて、エスカレーションフローを確認する。

6. レッドチーミングに関する工程（攻撃計画・実施）

「攻撃計画・実施」では、リスクシナリオや攻撃シナリオの作成、攻撃シナリオの実施、実施中の記録取得、実施後の処理を行う。

AISI Japan AI Safety Institute

第２工程 ： 攻撃計画・実施

レッドチーム

STEP 6 **リスクシナリオの作成**	• 対象ドメインとシステムのユースケースにおいて、システム構成、AIセーフティの評価観点、保護すべき情報資産、システムの利用形態の４つからリスクシナリオを作成する。
STEP 7 **攻撃シナリオの作成**	• 作成されたリスクシナリオに沿ってどのような攻撃が実際に可能であるかを検討し、レッドチーミングで行う具体的な攻撃シナリオを作成する。
STEP 8 **攻撃シナリオの実施**	• 作成された攻撃シナリオに沿って、具体的な攻撃シグネチャを投下することによって攻撃シナリオを実施する。
STEP 9 **実施中の記録取得**	• 実施されたレッドチーミングの詳細を証跡として保存するため、レッドチーミングの実施中の記録を取得する。
STEP 10 **実施後の処理**	• 対象AIシステムの開発・提供管理者や、情報システム部門・情報セキュリティ部門などの関連ステークホルダーに対してレッドチーミングでの攻撃終了の連絡を行う。 • レッドチーミング用の一時アカウントの削除や、一時的に設定を変更・緩和した防御策がある場合は、設定の復帰を行う。

6. レッドチーミングに関する工程（結果の取りまとめと改善計画の策定）

「結果のとりまとめと改善計画の策定」では、レッドチーミングの結果指摘された事項に対して改善を行う。実施結果を関係者がレビューした後、改善計画を策定・実施し、改善策のフォローアップを行う。

AISI Japan AI Safety Institute

第３工程 ： 結果のとりまとめと改善計画の策定

レッドチーム

STEP 11
実施結果の分析

- 攻撃計画・実施者は、レッドチーミングで得られた結果を分析する。
- 必要に応じて対象AIシステムの開発・提供管理者や情報システムの主管部、情報セキュリティの主管部等の関係部署に分析に必要な情報の追加確認を行う。

STEP 12
結果報告書の作成と関係者レビュー

- 攻撃計画・実施者は、発見された脆弱性をもとに、ログや証跡等を揃え、レッドチーミングの概要として示す。
- 結果報告書を作成し、事実の誤認がないか、対象AIシステムの開発・提供管理者、その他の関連ステークホルダーなどの関係者によるレビューを実施する。

プロジェクトチーム

STEP 13
最終報告書の作成と報告

- 対象AIシステムの開発・提供管理者は、攻撃計画・実施者による結果報告書をもとに、最終報告書をとりまとめる。
- 最終報告書について、必要に応じて経営層へ報告する。

STEP 14
改善計画の策定と実施

- 対象AIシステムの開発・提供管理者は、事業リスク等を勘案の上、具体的な改善策を検討し、改善計画を策定する。
- 具体的な改善策や改善計画の検討にあたっては、緊急度やリスク度合いに応じて優先度を検討する。

STEP 15
改善後のフォローアップ

- 改善計画に基づいて実施される改善策の進捗状況については、適宜、経営会議等で確認することが望ましい。
- 改善策を実施後、対策の設定状況確認やドキュメントレビュー、あるいは必要に応じて再度レッドチーミングを実施し、当該脆弱性が適切に改善され、リスクが低減していることを確認することが望ましい。

AISI

Japan AI Safety Institute