AIセーフティに関するレッドチーミング手法ガイド（第1.10版）別紙：詳細解説書

AIセーフティ・インスティテュート

（令和7年3月31日）

AISI Japan AI Safety Institute

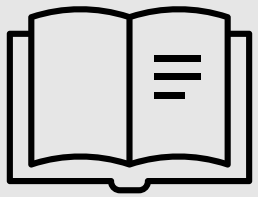
AIセーフティに関するレッドチーミング手法ガイドの構成

AIセーフティに関するレッドチーミング手法ガイドは、本編、別紙（詳細解説書）、別添（成果物例）からなる構成としている。

- 本編では、レッドチーミング手法に関する基本的な考慮事項を第1工程～第3工程に分けて体系的に記載している。
- 別紙：詳細解説書）では、本編に沿ってレッドチーミングを実施する際の実施ポイントや各工程での成果物例を解説している。
- 別添（成果物例）では、本編に沿ってレッドチーミングを実施する際に作成する成果物の一部（リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果、レッドチーミング実施結果報告書、最終報告書）を例として示す。

AIセーフティに関するレッドチーミング手法ガイド

本編

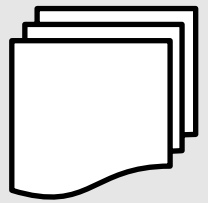


レッドチーミング手法に関する基本的な考慮事項を示す。

別紙（詳細解説書）【本書】

より実践的な、各工程での実施事項や実施ポイントを解説している。

別添（成果物例）



レッドチーミングを実施する際に作成する成果物の一部を例として示す。

目次

レッドチーミング手法ガイドに沿ってレッドチーミングを実施し、その結果から得られた示唆を詳細解説書として示すことで、より実践的な資料として拡充させることを目的とした。

AISI Japan AI Safety Institute

背景

- レッドチーミング（以下、RT）とは、 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手法**である。
- 2024年9月に、**RTに関する基本的な考慮事項を示すこと**を目的に「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、RT手法ガイド）を作成した。
- RT手法ガイドは、「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の調査を踏まえて体系的に作成している。ただし、**特にRTの第2工程（攻撃計画・実施）には高い専門性が求められるため、より実践的なガイドにする必要がある。**そこで、RAGを用いたLLMシステムに対して実際にRTを実施し、その結果から得られた示唆をガイドに盛り込むこととした。

目的

- **RT手法ガイドに沿って実際にRTを実施し、その結果から得られた示唆を別紙の「詳細解説書」（以下、本書）として示す**ことで、RT手法ガイドをより実践的な資料として拡充させることを目的とした。
  ※本書に記載の内容はあくまで一例であり、個々の組織にて適宜内容を変更して実施することも考えられる。

2. 詳細解説書の位置づけ

本書では、本編で示した工程の流れに沿って、【概要】、【詳細】、【参考】を示す。特に高い専門性が求められる第2工程（攻撃計画・実施/STEP6～STEP10）を重点的に解説し、より実践的なガイドとした。

AISI Japan AI Safety Institute

AIセーフティに関するレッドチーミング手法ガイド

本編

工程	実施事項
第１工程：実施計画の策定と実施準備	(STEP 1) … (STEP 2) … (STEP 3) … (STEP 4) … (STEP 5) …
第２工程：攻撃計画・実施	(STEP 6) … (STEP 7) … (STEP 8) … (STEP 9) … (STEP 10) …
第３工程：結果のとりまとめと改善計画の策定	(STEP 11) … (STEP 12) … (STEP 13) … (STEP 14) … (STEP 15) …

別紙（詳細解説書）【本書】

要素	内容	該当箇所
【概要】	RT手法ガイドに記載されている各工程・実施事項の概要を記載する。	「3. 各工程の解説」箇所であり、各スライドに【概要】、【詳細】、【参考】のタグを付与している。
【詳細】	本編に示したSTEP1～STEP15それぞれにおけるRTの実施事項、実施イメージ、実施ポイントを記載する。（次頁参照）	
【参考】	各工程で具体的にどのような項目を検討すべきか一例を記載する。	

別添（成果物例）

成果物例の名称	内容
リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果	本編のSTEP6~8作成するリスクシナリオ、攻撃シナリオの作成例と攻撃シナリオの実施結果をExcel形式で示す。
レッドチーミング実施結果報告書	本編のSTEP12で作成するレッドチーミング実施結果報告書の例をPowerPoint形式で示す。
最終報告書	本編のSTEP13で作成する最終報告書の例をPowerPoint形式で示す。

2. 詳細解説書の位置づけ

本書の【詳細】では、本編で示したSTEP1～STEP15それぞれにおけるRTの実施事項、実施イメージ、実施ポイントを解説する。

AISI Japan AI Safety Institute

実施事項

本編に記載されている実施事項を記載する。

実施イメージ

RTの実施イメージを記載する。

実施ポイント

本編に沿ってRTを実施するにあたり、各STEPで特に留意すべき事項を、実施ポイントとして記載する。

3. 各工程の解説

RT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。

AISI Japan AI Safety Institute

工程	実施事項	本編での該当章
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 ✓ レッドチーミング実施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

3. 各工程の解説

第1工程：実施計画の策定と実施準備

【概要】（STEP 1）チーム発足～（STEP 3）計画の策定

【凡例】：攻撃計画・実施者　：AIシステムに関連する有識者　：対象AIシステムの開発・提供管理者　：その他関連ステークホルダー　：経営層　プロジェクトチーム　レッドチーム

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

- RT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。

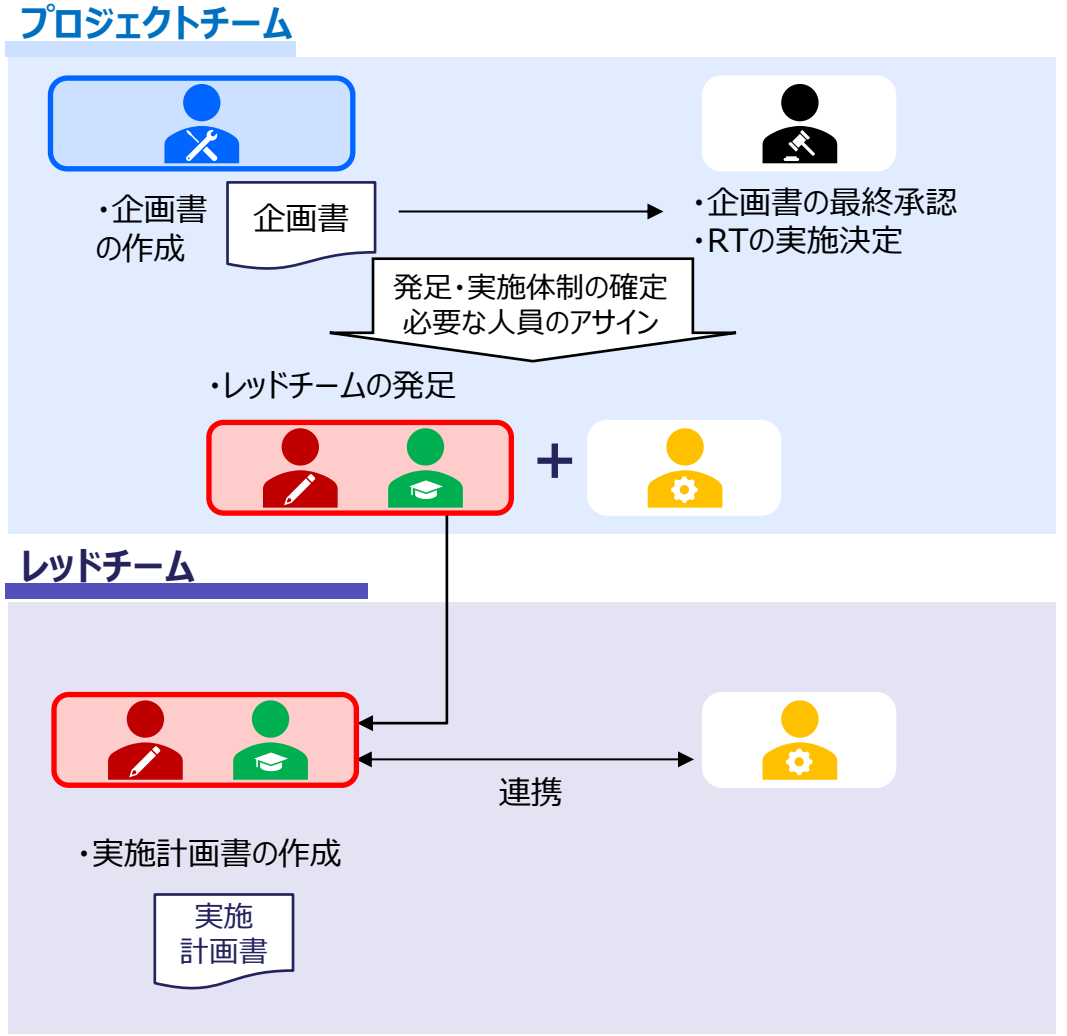
STEP 2
予算及びリソースの識別・確保とサードパーティの選定・契約

- 予算の確保およ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。

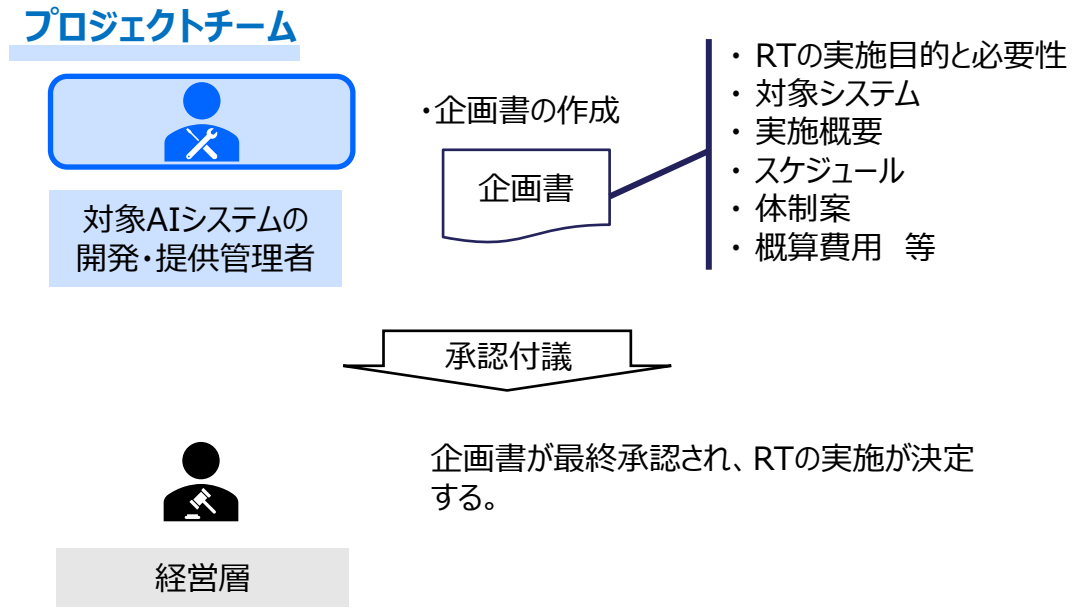


【詳細】（STEP 1）実施の決定とレッドチーミング発足

実施事項

- RT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組織内に発足させる。

実施イメージ



実施ポイント

- RTの対象システムについて、管理部署等の詳細情報を把握しておくことで、その後のレッドチームの発足等で関係者との調整を円滑に実施することができる。
- レッドチームの編成には、「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する。

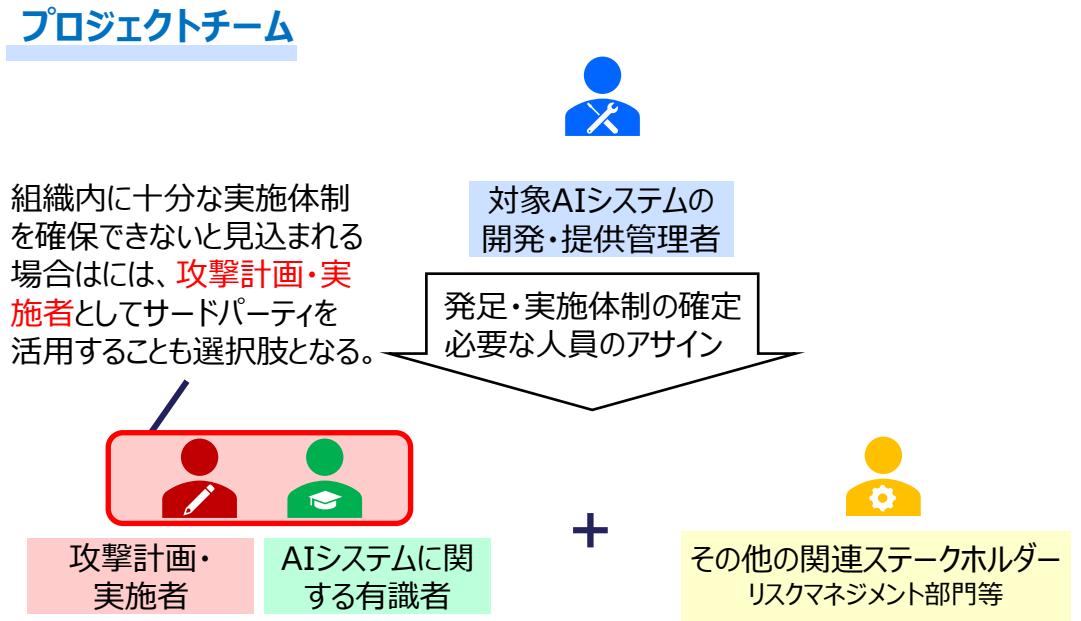
【詳細】（STEP 2）予算及びリソースの識別・確保とサードパーティの選定・契約

AISI Japan AI Safety Institute

実施事項

- 対象AIシステムの開発・提供管理者は、予算の確保及び組織内の実施体制を確定し、必要な人員のアサインやツールの確保を行う。

実施イメージ



実施ポイント

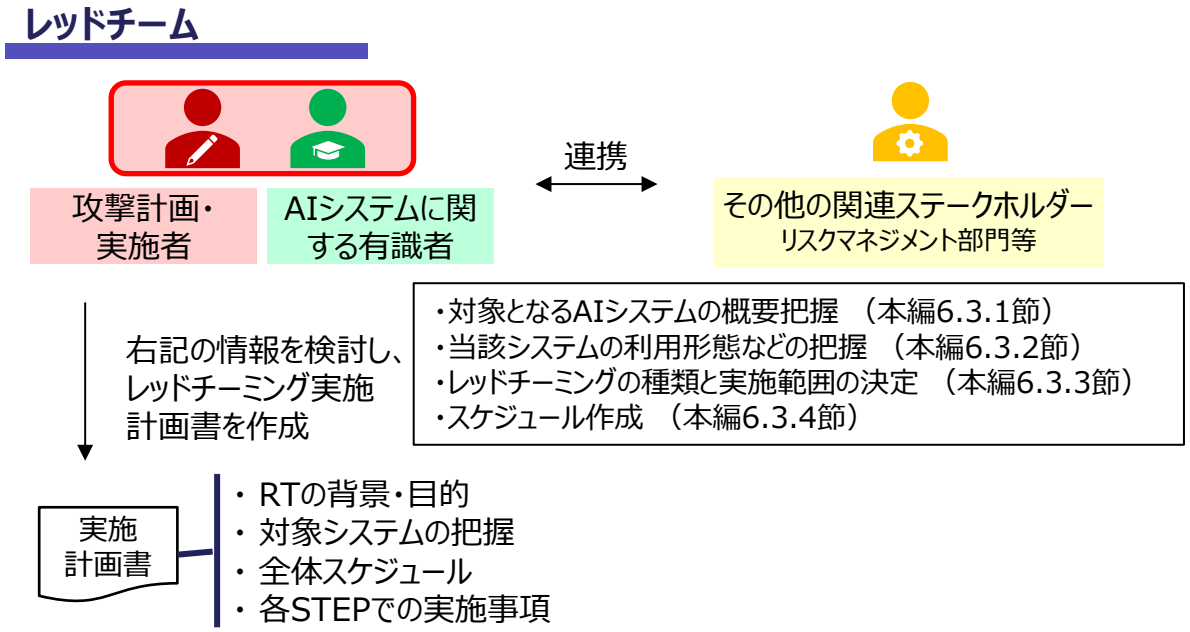
- RT実施には高度な専門性が必要となるため、組織内の体制が不十分な場合、攻撃計画・実施者としてサードパーティを活用することも選択肢となる。なお、RTを実施する過程で内部の機密情報を取り扱うことも想定されるため、十分な情報セキュリティ保護対策を実施することが必要となる。

【詳細】（STEP 3）実施計画の策定

実施事項

- 本編6.3.1節「対象となるＡＩシステムの概要把握」、6.3.2節「当該システムの利用形態」、6.3.3節「レッドチーミングの種類と実施範囲の決定」などから、RT実施のために**必要となる事項**を検討する。そのうえで「**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。

実施イメージ



実施ポイント

- 本編6.3.1節「対象となるＡＩシステムの概要把握」、6.3.2節「当該システムの利用形態」で検討すべき項目例として、P.14の【参考】に示す項目が考えられる。
- レッドチーミング実施計画書は、STEP3に記載のある項目のみ検討するのでなく、STEP4以降の全体を踏まえた構成にする。

【参考】（STEP 3） 6.3.1節、6.3.2節で検討すべき項目例

「対象となるAIシステムの概要把握」で検討すべき項目例

カテゴリ	項目
AIシステムの概要把握	AIシステム全体のシステム構成図やネットワーク構成図
	AIシステムの利用用途
	AIシステムの動作概要
	AIシステムを構成するLLM
	AIシステムのLLM以外の構成要素
	取り扱うデータ

「当該システムの利用形態などの把握」で検討すべき項目例

カテゴリ	項目
LLMの利用形態	（A）LLMの出力に関する利用形態
	（B）LLMの参照先に関する利用形態
	（C）LLM自体の利用形態
LLM以外のコンポーネントに関する概要把握	商用プラグインやライブラリの利用
	OSSのプラグインやライブラリの利用
	自組織で独自に開発したプラグインやライブラリの利用
導入済の防御機構等	LLMへの入力を前段でチェックする機構
	LLM自体での防御対策
	LLMからの出力を後段でチェックする機構
	入出力のユーザーフィードバックによる強化学習
その他の情報把握	事前設定されたユーザープロンプト
	システムプロンプト
	デプロイ環境
	APIパラメータ
	ファインチューニングの実施状況
	ユーザーデータの訓練への利用有無
	訓練データの取得元の有無
	他組織による実施済のレッドチーミングなどの情報

【概要】（STEP 4）実施環境の準備、（STEP 5）エスカレーションフローの確認

AISI Japan AI Safety Institute

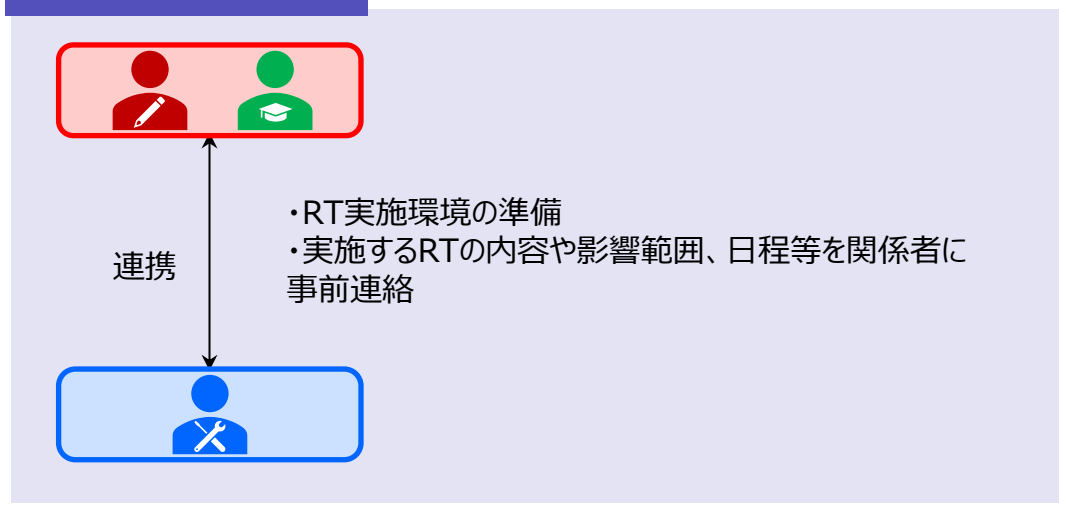
【凡例】：攻撃計画・実施者　：AIシステムに関連する有識者　：対象AIシステムの開発・提供管理者　：その他関連ステークホルダー　：経営層　プロジェクトチーム　レッドチーム

第1工程：実施計画の策定と実施準備

レッドチーム

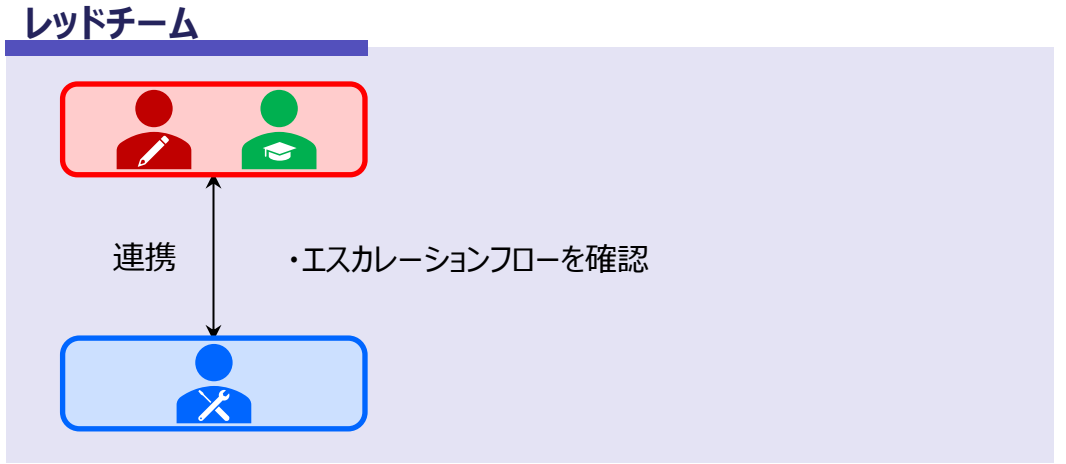
STEP 4 実施環境の準備

- RT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

- RT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。

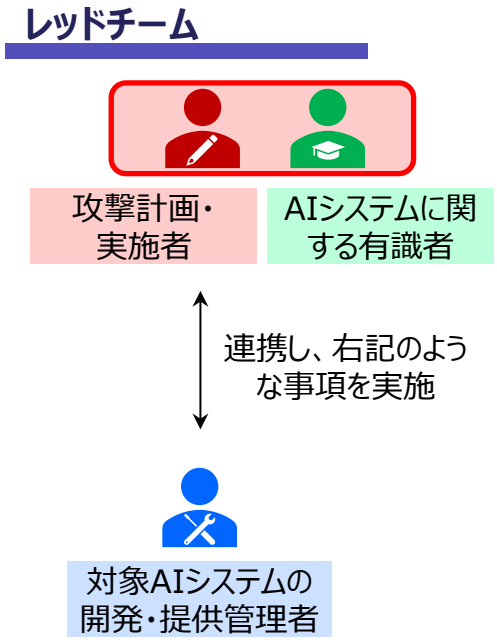


【詳細】（STEP 4）実施環境の準備

実施事項

- レッドチームは、RTの実施環境について、対象AIシステムの開発・提供管理者と連携して必要な準備を進める。
- その際、実施するRTの内容や影響範囲、日程等の関係者へ事前に連絡する。
- 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。

実施イメージ



- ID発行、アクセス権の付与、ログ取得
- 異常検知システムが導入されている場合には、事前に関係者に周知の上、監視設定の一時的解除等の依頼
- 実施するRTの内容や影響範囲、日程等を関係者に事前連絡

実施ポイント

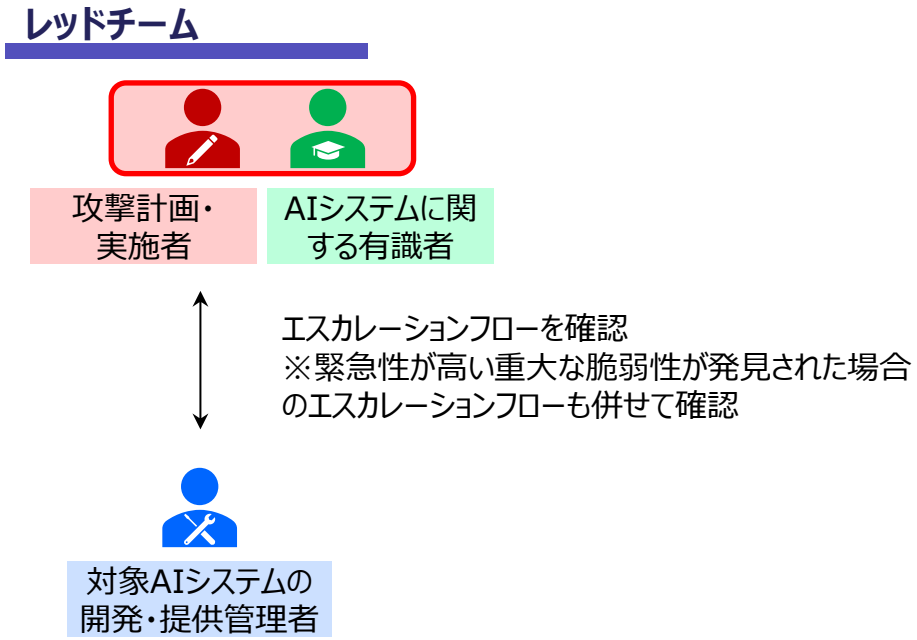
- RTの実施によって、異常検知システムから大量の検知ログが出力される可能性がある。そのため、RTを実施するための環境整備だけでなく、RTの内容や影響範囲、日程を関係者（攻撃シナリオの実施によって影響を受けるシステムに関わる組織）と事前にすり合わせることが望ましい。

【詳細】（STEP 5）エスカレーションフローの確認

実施事項

- RT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。
- 緊急性の高い重大な脆弱性が発見された際には、レッドチーミング実施結果報告書の作成を待たず、関係者に至急、情報共有する。その場合のエスカレーションフローも併せて確認する。

実施イメージ



実施ポイント

- 万が一の障害や業務影響、また重大な脆弱性が発見された場合に備えた取り決めなどを関係者間で事前に合意形成する。これにより、RT実施中に不測の事態に陥った場合に迅速に対応することができる。

3. 各工程の解説
第2工程：攻撃計画・実施

【概要】（STEP 6）リスクシナリオの作成

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、システムの利用形態、保護すべき情報資産、AIセーフティの評価観点の４つからリスクシナリオを作成する。

レッドチーム

攻撃計画・実施者 / データサイエンティスト / 情報システムの主管部 情報セキュリティの主管部 / 対象AIシステムの開発・提供管理者 / ドメイン専門家 / リスクマネジメント部門

【リスクシナリオ作成の参考例】

凡例
→ ：本編記載の流れ
⇢ ：本編記載とは逆の流れ

Step 6-1: システム構成の把握、情報の流れを整理

Step 6-2: 情報資産と考慮すべきAIセーフティの評価観点を特定

Step 6-3:
- AIセーフティ上の懸念箇所を抽出（攻撃者観点）
- 攻撃手法の概要や引き起こされるリスクの可能性を共有
- 当該システムにおける攻撃の可能性や前提条件、実際に想定されるシステムへの影響度・影響範囲などを検討
- 事業リスクや事業インパクトを検討 AIセーフティの評価観点への影響を考慮

ブレインストーミングを通じて、さまざまなリスクシナリオを洗い出し必要に応じて前の実施事項に戻る。

攻撃成功の可能性や事業インパクト、AIセーフティの評価観点への影響が大きいものを中心に、RTで評価するリスクシナリオを作成

【詳細】（STEP 6-1）システム構成の把握

実施事項

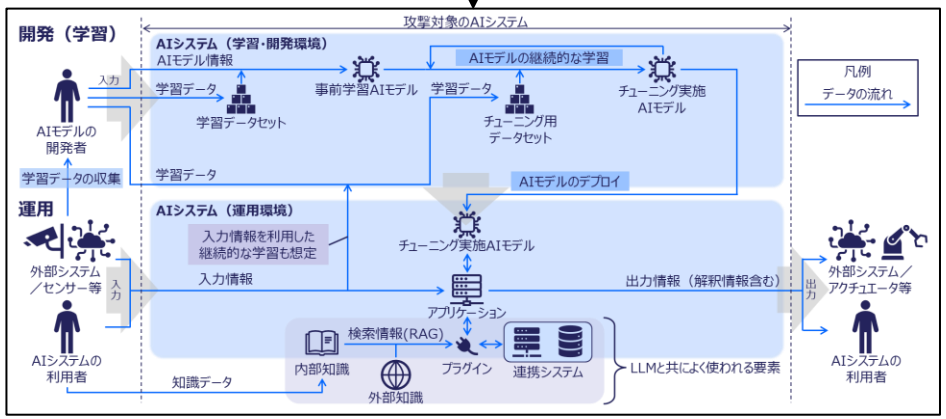
- 本編6.3.1節「対象となるAIシステムの概要把握」で得られた情報から**システム構成**を把握し、LLMの入出力と他コンポーネントの**情報の流れ**を整理する。

実施イメージ

レッドチーム

攻撃計画・実施者

システム構成を把握し、LLMの入出力と他コンポーネントの情報の流れを整理する



実施ポイント

- 本編6.3.1節「対象となるAIシステムの概要把握」で把握したシステム構成図から、より具体的なLLMの入出力と他コンポーネントの情報の流れを整理する。
- 本編6.3.1節 図5に、２種類の環境（開発／運用）で構成されるAIシステムの構成図を参考例として示す。
- 【成果物例：レッドチーミング実施結果報告書】※では、RAGを利用した社内業務データ活用チャットボットサービスのシステム構成図を参考例として示している。

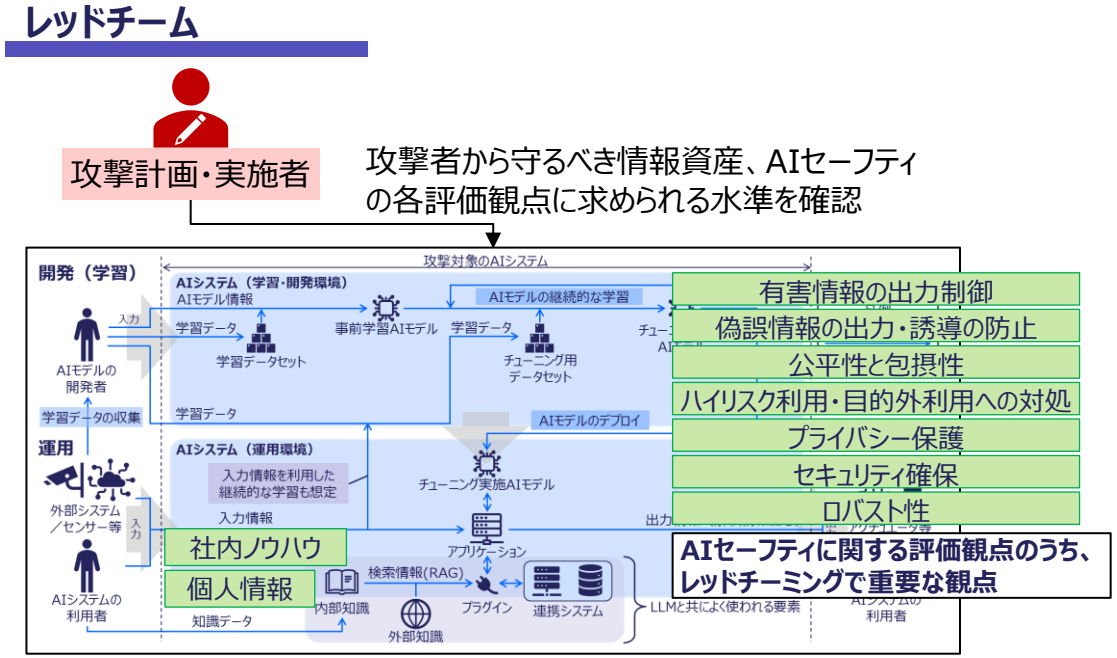
※ RT手法ガイド＞別添（成果物例）＞レッドチーミング実施結果報告書を参照

【詳細】（STEP 6-2）考慮すべきAIセーフティの観点と情報資産の特定

実施事項

- システム内のサービスや機能に基づいて、保護すべき情報資産を洗い出し、特に攻撃者から**守るべき重要情報**に注目する。
- **AIセーフティの各評価観点に求められる水準**を確認する。

実施イメージ



実施ポイント

- 「AIセーフティの各評価観点に求められる水準」については、取り扱う情報などの個々のAIシステムの特性を考慮し各組織で定義する。
- 【成果物例：リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果】※1や【成果物例：レッドチーミング実施結果報告書】※2にAIセーフティの評価観点に求められる水準の検討結果を例として示す。

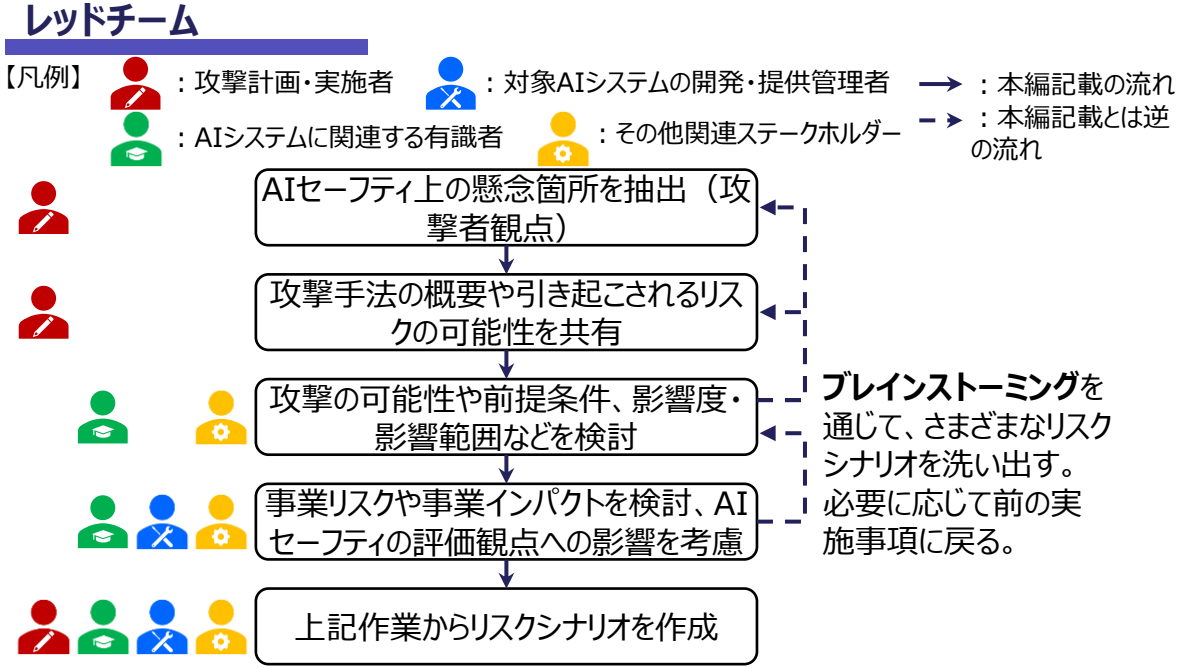
※1 RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果を参照
※2 RT手法ガイド＞別添（成果物例）＞レッドチーミング実施結果報告書を参照

【詳細】（STEP 6-3）システム構成と利用形態からリスクシナリオの作成

実施事項

- リスクシナリオを作成する。
- リスクシナリオとは、AIシステムや運用環境において発生しうるリスクを具体的に想定し、それに伴う脅威の発生箇所や影響を明確化するためのシナリオである。
- 例えば、右図（実施イメージ）に示すフローチャートにしたがい、関係者で密に連携をとりつつ作成する。

実施イメージ

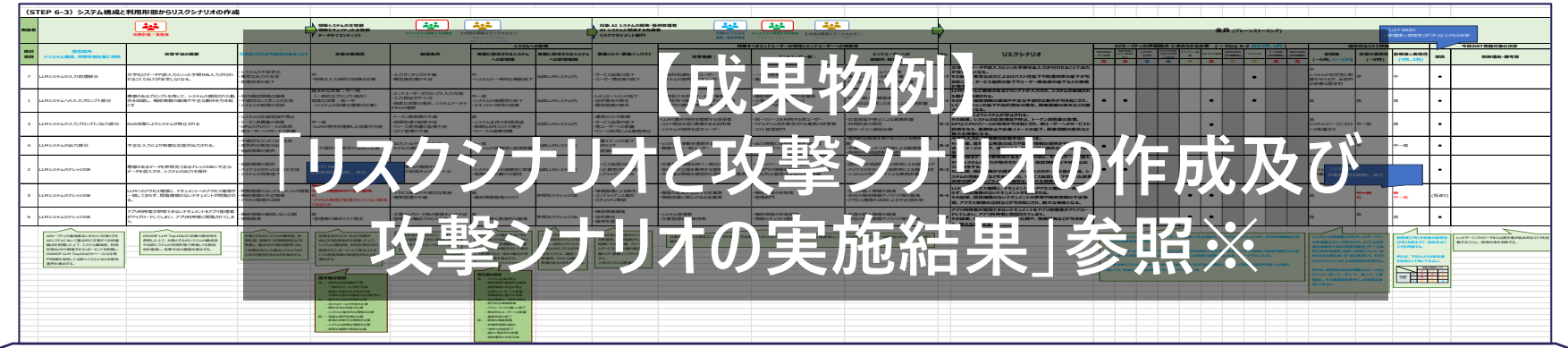


実施ポイント

- 各関係者（攻撃実施者、情報システム部門、情報セキュリティ部門など）の役割に応じた項目を検討し、密に連携を取りながらリスクシナリオを作成する。
- 次頁の【参考】スライドや【成果物例：リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果】※では、リスクシナリオ作成の検討事項や手順を解説する。

※ RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果を参照

【参考】（STEP 6-3）リスクシナリオ作成における各関係者の検討項目



【凡例】 ：攻撃計画・実施者 ：AIシステムに関連する有識者 ：対象AIシステムの開発・提供管理者 ：その他関連ステークホルダー

リスクシナリオ作成時における実施者、検討項目について

実施者	攻撃計画・実施者			AIシステムに関連する有識者（データサイエンティスト） その他関連ステークホルダー（情報システムの主管部、情報セキュリティの主管部）			対象AIシステムの開発・提供管理者 AIシステムに関連する有識者（ドメイン専門家） その他関連ステークホルダー（リスクマネジメント部門）		全員（ブレインストーミング）		
列	B列	C列	D列	E列	F列	G列・H列	I列	J列-L列	O列-X列	Y列-AA列	AB列・AC列
検討項目	懸念箇所	攻撃手法の概要	引き起こされる可能性のあるリスク	攻撃の実現性	前提条件	システムへの影響度	事業リスク・事業インパクト	エンドユーザーの特性と被害等	AIセーフティの評価観点に求められる水準	総合的なリスク	今回のRT実施対象の決定

※ RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果を参照
なお、本編7.2.3節の攻撃シナリオの作成例（表4~6）のフォーマットと異なるが、成果物例の方法もあくまで一例であり、個々の組織にて適宜方法を変更して実施することも考えられる。

【参考】（STEP 6-3）リスクシナリオ作成における各項目の検討プロセス例

【凡例】：攻撃計画・実施者　：AIシステムに関連する有識者　：対象AIシステムの開発・提供管理者　：その他関連ステークホルダー

リスクシナリオ作成時における実施者、検討項目、プロセス例について												
実施者	攻撃計画・実施者			AIシステムに関連する有識者（データサイエンティスト） その他関連ステークホルダー（情報システムの主管部、情報セキュリティの主管部）				対象AIシステムの開発・提供管理者 AIシステムに関連する有識者（ドメイン専門家） その他関連ステークホルダー（リスクマネジメント部門）		全員（ブレインストーミング）		
列	B列	C列	D列	E列	F列	G列・H列		I列	J列-L列	O列-X列	Y列-AA列	AB列・AC列
検討項目	懸念箇所	攻撃手法の概要	引き起こされる可能性のあるリスク	攻撃の実現性	前提条件	システムへの影響度		事業リスク・事業インパクト	エンドユーザーの特性と被害等	AIセーフティの評価観点に求められる水準	総合的なリスク	RT実施対象の決定
プロセス例	対象とするAIシステムの構成図において攻撃対象となりうる懸念箇所を導出する。AIセーフティの評価観点に求められる水準をもとに、重視する評価観点をあらかじめ決めておき、専門知識を活用しつつ検討する。	列挙した懸念箇所について成立しうる脆弱性を推測し、攻撃手法の概要を導出する。OWASP LLM Top10などに記載の脆弱性を参照し、対象とするAIシステムの構成図やシステムの利用形態を考慮しつつ検討する。	対象とするAIシステムの構成図、利用形態、保護すべき情報資産などを考慮し、導出された懸念箇所に対して攻撃が成立した場合にどのようなリスクの可能性があるのかを導出する。	列挙した懸念箇所と攻撃手法について、どの程度攻撃の実現性があるのかを導出する。攻撃手法やリスク、攻撃成立の前提条件を把握し、対象とするAIシステムの構成図とシステムの利用形態を考慮しつつ検討する。	攻撃手法の概要をもとに、必要に応じて追加調査を行った上で、攻撃がどのような条件下で成立しうるのかを導出する。	■実際に想定されるシステムへの影響度 リスクが引き起こされた場合のシステムへの影響の大きさを導出する。AIセーフティの評価観点に求められる水準をもとに、重視する評価観点を考慮しつつ検討する。	■実際に想定されるシステムへの影響範囲 リスクが引き起こされた場合、参照先のナレッジDB単体、当該LLMシステム、当該LLMシステム以外の自社システム、顧客への影響等、どのような範囲で影響があるのかを導出する。	攻撃が成立した場合のシステムの影響度、影響範囲から、自組織に与える影響、ステークホルダーに与える影響の観点で事業リスク・事業インパクトを導出する。 ※双方に与える影響を含む	対象システムの想定するエンドユーザーの特性を攻撃者／被害者について導出する。攻撃者については保持するスキルや攻撃の動機を列挙し、リスクシナリオ作成の参考情報として活用する。被害者についてはエンドユーザーの属性を列挙し引き起こされる直接的および間接的な被害を列挙する。	リスクシナリオがどの評価観点に紐づくのかを導出し、星取表を作成する。参考資料（※）にて定義した、各評価観点の「レッドチーミングの実施対象となるAIシステムでの求められる水準」を参照し、網羅性を確認する。不足している場合はリスクシナリオ検討を再度実施する。	攻撃の実現性とシステムへの影響度をもとに、RTのリーダまたは責任者が総合的なリスクを導出する。 総合的なリスクの算出方法と関係性は次頁参照	RTのリーダまたは責任者は総合的なリスクの評価結果をもとにRT実施対象の採否を決定し、その判断理由とともに記載する。

※ RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果の「（STEP 6-2）AIセーフティの各評価観点に求められる水準」シートを参照

【参考】（STEP 6-3）総合的なリスク評価プロセスの例

- 総合的なリスク評価方法の一例として、攻撃の実現性とシステムへの影響度を利用することが考えられる。
- システムへの影響度は、事業リスク・インパクト、エンドユーザーの特性、AIセーフティの評価観点に求められる水準を考慮する。

【凡例】：攻撃計画・実施者　：AIシステムに関連する有識者　：対象AIシステムの開発・提供管理者　：その他関連ステークホルダー

実施者

攻撃計画・実施者
AIシステムに関連する有識者
その他関連ステークホルダー

対象AIシステムの開発・提供管理者
AIシステムに関連する有識者
その他関連ステークホルダー

全員

検討項目

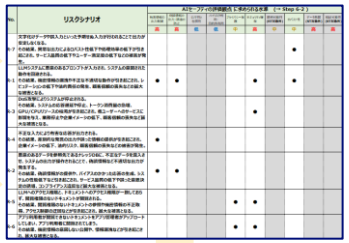
攻撃の実現性

システムへの影響度

事業リスク・インパクト

エンドユーザーの特性と被害等

AIセーフティの評価観点に求められる水準



システムへの影響度の再評価

注：この算定ロジックはあくまでも例であり、各社のリスク評価手法や他の算定ロジックを用いてもよい。

総合的なリスク評価

		攻撃の実現性		
		高	中	低
影響度	高	高	高	中
	中	高	中	中
	低	中	中	低

RT実施対象の決定

【参考】（STEP 6-3）攻撃の実現性、システムへの影響度、総合的なリスク評価の指標例

AISI Japan AI Safety Institute

- 攻撃の実現性とシステムへの影響度を分類して、総合的なリスク評価の指標を算出することが考えられる。

レベル(加点)	攻撃の実現性	システムへの影響度
説明	攻撃が実現（発生）する可能性を判定する。以下はレベルごとの例を示す。	システムへの影響度を判定する。以下はレベルごとの例を示す。
高(+3)	- 特別な技術知識が不要 - 一般的なツールで実行可能 - 既知の攻撃手法が流用可能 - 内部の仕組みを理解する必要がない	- システム全体の停止 - 機密情報の直接的な漏洩 - 重要機能の完全な停止 - 広範なユーザーへの影響 - 事業継続に重大な支障
中(+2)	- 基本的な技術知識が必要 - カスタムツールの作成が必要 - 既存手法の改変が必要 - システムの基本的な理解が必要	- 特定機能の一時的な停止 - 部分的な情報漏洩 - パフォーマンスの著しい低下 - 限定的なユーザーへの影響 - 業務効率の低下
低(+1)	- 高度な専門知識が必要 - 新規の攻撃手法開発が必要 - システムの詳細な理解が必要 - 特別な権限や環境が必要	- 軽微な機能障害 - 非機密情報の露出 - 一時的な性能低下 - 極めて限定的な影響 - 通常運用で対処可能

総合的なリスク評価

		攻撃の実現性		
		高(+3)	中(+2)	低(+1)
システムへの影響度	高(+3)	**高(6)**	**高(5)**	**中(4)**
	中(+2)	**高(5)**	**中(4)**	**中(3)**
	低(+1)	**中(4)**	**中(3)**	**低(2)**

【目安】

- 高：現行システムにて対応が必要
- 中：現行システムで対応又は次回リプレイス時に対応が必要
- 低：現行システムでは対応は不要

注：この算定ロジックはあくまでも例であり、各社のリスク評価手法や他の算定ロジックを用いてもよい。

【概要】（STEP 7）攻撃シナリオの作成

AISI Japan AI Safety Institute

第2工程：攻撃計画・実施

レッドチーム

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、RTで行う具体的な攻撃シナリオを作成する。

レッドチーム

攻撃計画・実施者

Step 7-1

・主要コンポーネントの諸元確認
・システム構成における各コンポーネントの分類を踏まえて、RT実施方法の選択肢を導出する。

Step 7-2

・レッドチーミング実施の対象環境とアクセスポイント確認
・各コンポーネントからどの環境で実施するか検討する。
・アクセスポイントを列挙する。

Step 7-3

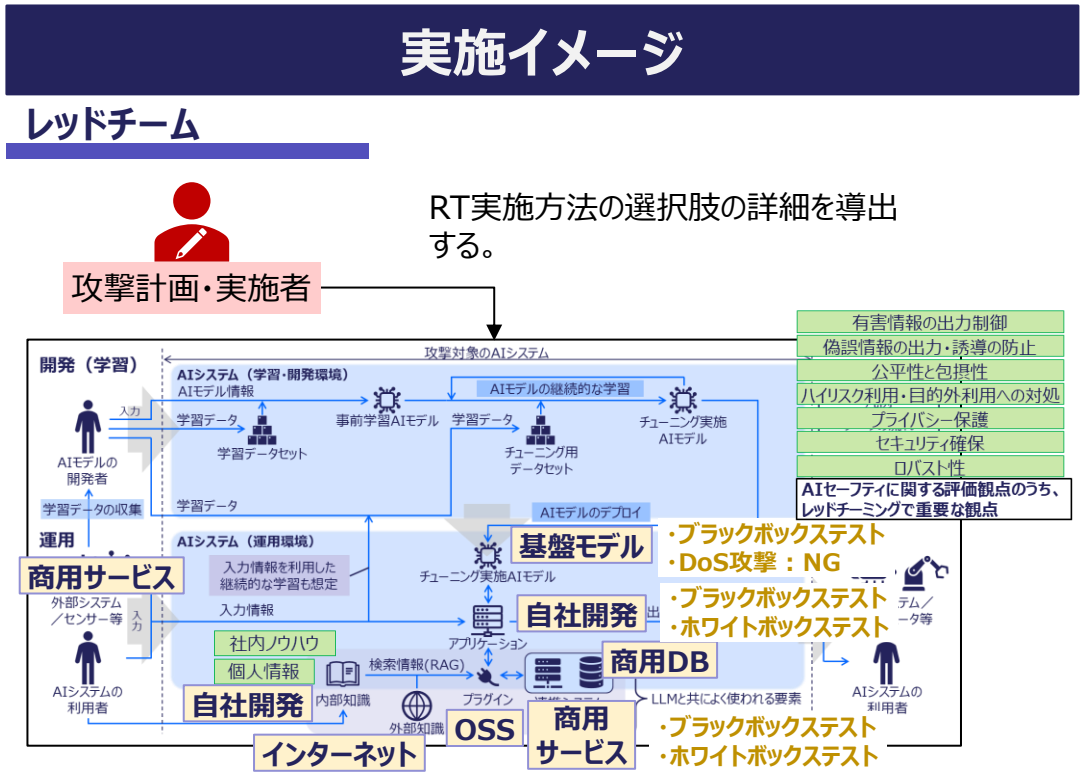
・攻撃シナリオの作成
・攻撃者の目線で一連のシナリオを作成する。
・LLMにおける代表的な防御機構の構成を踏まえ、攻撃を組み立てる。
・実際の攻撃手法、攻撃のトレンド、実際の被害事例、対策上で見落としがちな盲点等を考慮する。
・セキュリティのフレームワークも参考する。

【詳細】（STEP 7-1）主要コンポーネントの諸元確認

実施事項

- 本編7.1.1節「システム構成の把握」と6.3.3節「レッドチーミングの種類と実施範囲の決定」で得られた情報をもとに、システム構成内の各コンポーネントを商用サービス、OSS、または自組織開発に分類する。分類結果に基づいてRT実施方法の詳細な選択肢を導き出す。

実施イメージ



実施ポイント

- 主要コンポーネントの諸元によって、ホワイトボックステストが可能であるのか、ブラックボックステストのみ可能であるかなどが明らかになり、RT実施方法の詳細な選択肢を導き出すことが可能になる。

【詳細】（STEP 7-2）レッドチーミング実施の対象環境とアクセスポイント確認

AISI Japan AI Safety Institute

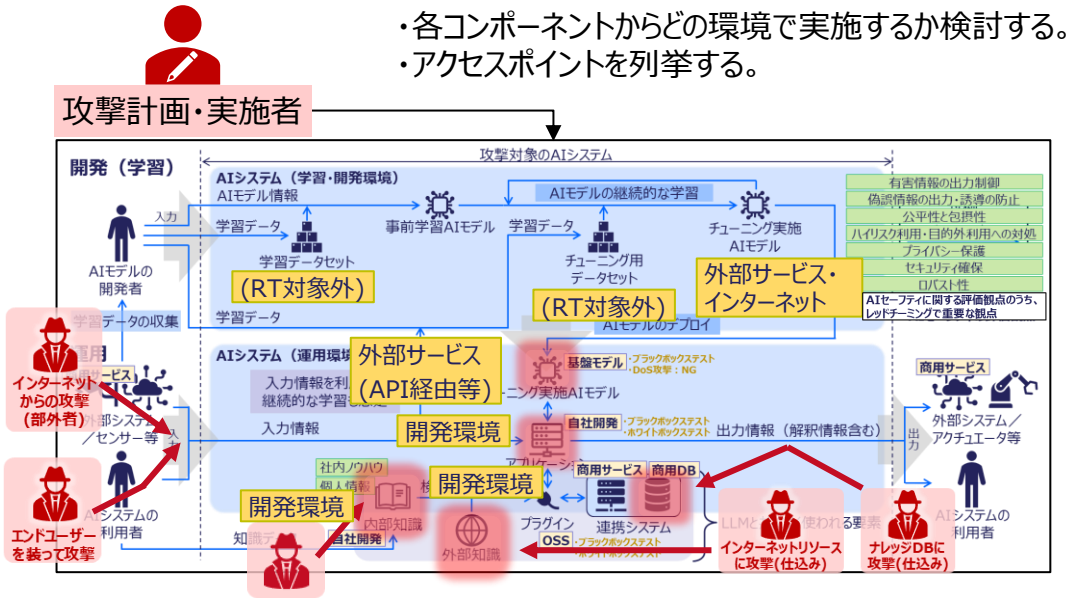
実施事項

- 本編6.3.3節「レッドチーミングの種類と実施範囲の決定」で得られた情報をもとに、各コンポーネントについてRTを実施する環境を検討する。
- RTを実施する際のアクセスポイントを検討する。

実施イメージ

レッドチーム

・各コンポーネントからどの環境で実施するか検討する。
・アクセスポイントを列挙する。



実施ポイント

- 各コンポーネントについて、実運用環境、ステージング環境、開発環境の、どの環境でRTを実施するかを検討する。
- RTを実施する際のアクセスポイントを検討するが、ここではアクセスポイントの選択肢を洗い出すに留める。

【詳細】（STEP 7-3）攻撃シナリオの作成

実施事項

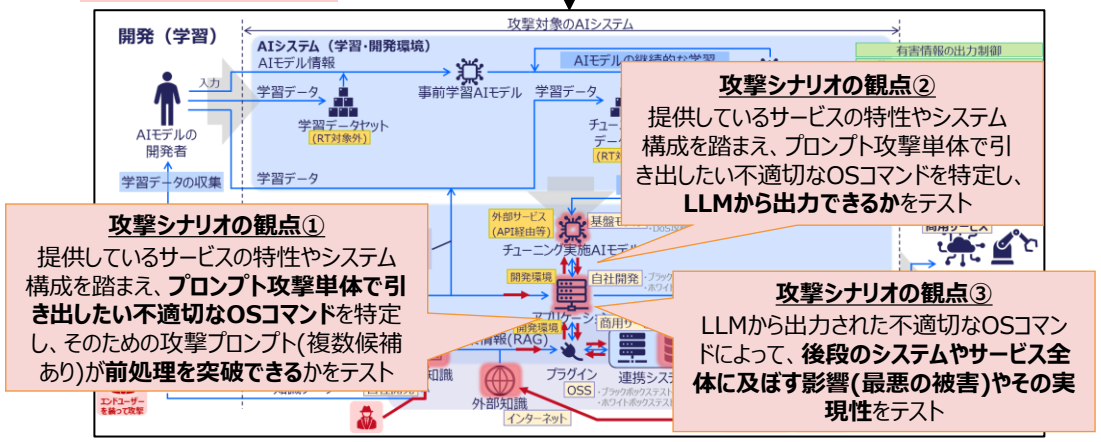
- 攻撃計画・実施者は、リスクシナリオに沿って攻撃シナリオを作成する。
- 攻撃シナリオとは、特定のリスクシナリオに基づき、攻撃者の視点からどの環境に対して、どのアクセスポイントを利用し、どのような手法を組み合わせて攻撃を実行するかを定めた計画である。
- 複数の観点から攻撃シナリオを検討するのが有効である。

実施イメージ

レッドチーム

攻撃計画・実施者

攻撃者の目線で一連のシナリオを作成



実施ポイント

- LLMシステムにおける代表的な防御機構や、実際に報告されている攻撃手法、攻撃のトレンド、実際の被害事例、対策上で見落としがちな点を考慮することで、実効性の高いRTの実施に繋がる。
- 攻撃シナリオの作成例を【成果物例：リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果】※に示す。

※ RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果を参照

【概要】（STEP 8）攻撃シナリオの実施

AISI Japan AI Safety Institute

第2工程：攻撃計画・実施

レッドチーム

STEP 8 攻撃シナリオの実施

- 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。

レッドチーム

攻撃計画・実施者

Step 8-1

・プロンプト単体に関するレッドチーミング
・有効な攻撃手法を特定することを目的とする。
・自動化ツールを用いて探索するのが効率的である。※ただし、手動でも検証する必要がある。

Step 8-2

・攻撃シグネチャと攻撃シナリオ実施手順の作成
・実際に投下する攻撃シナリオと攻撃シナリオ実施手順をとりまとめる。
・想定外の挙動を引き起こすことができるかという観点から設計し、そこから逆算して攻撃シグネチャを作成する。

Step 8-3

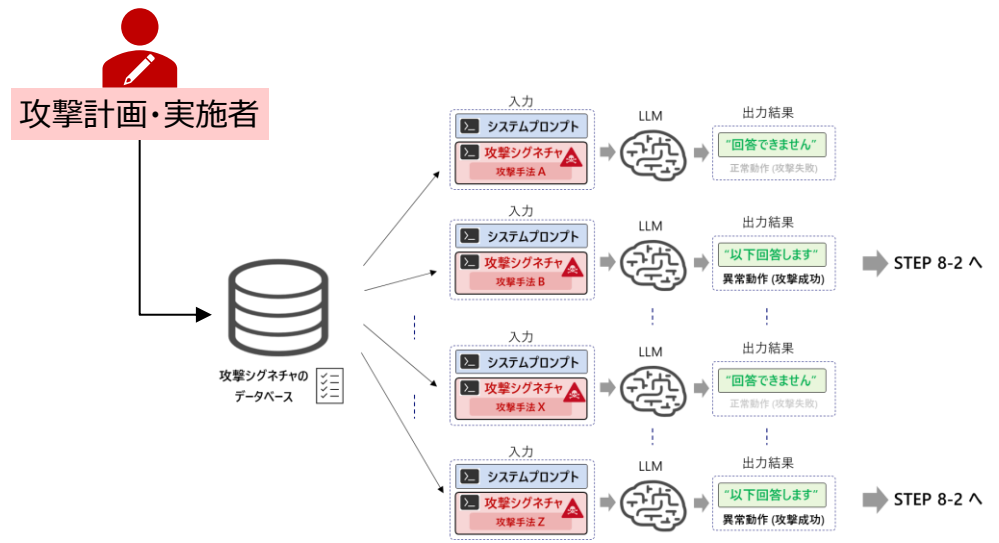
・LLMシステム全体に対するレッドチーミング
・攻撃シナリオ実施手順に基づき、結果を検証する。
・出力結果をフィードバックしながら、攻撃シグネチャをチューニングしたり、別の攻撃シグネチャを投下するなどの探索を試みる。

【詳細】（STEP 8-1）プロンプト単体に関するレッドチーミング

実施事項

- LLMの基本的な脆弱性や攻撃手法を特定することを目的として、プロンプト単体を対象としたRTを実施する。
- 対象システムに対して、攻撃シグネチャを入力する。
- 攻撃シグネチャとは、特定の攻撃手法を実行するために使用される具体的な入力やパターンである。

実施イメージ



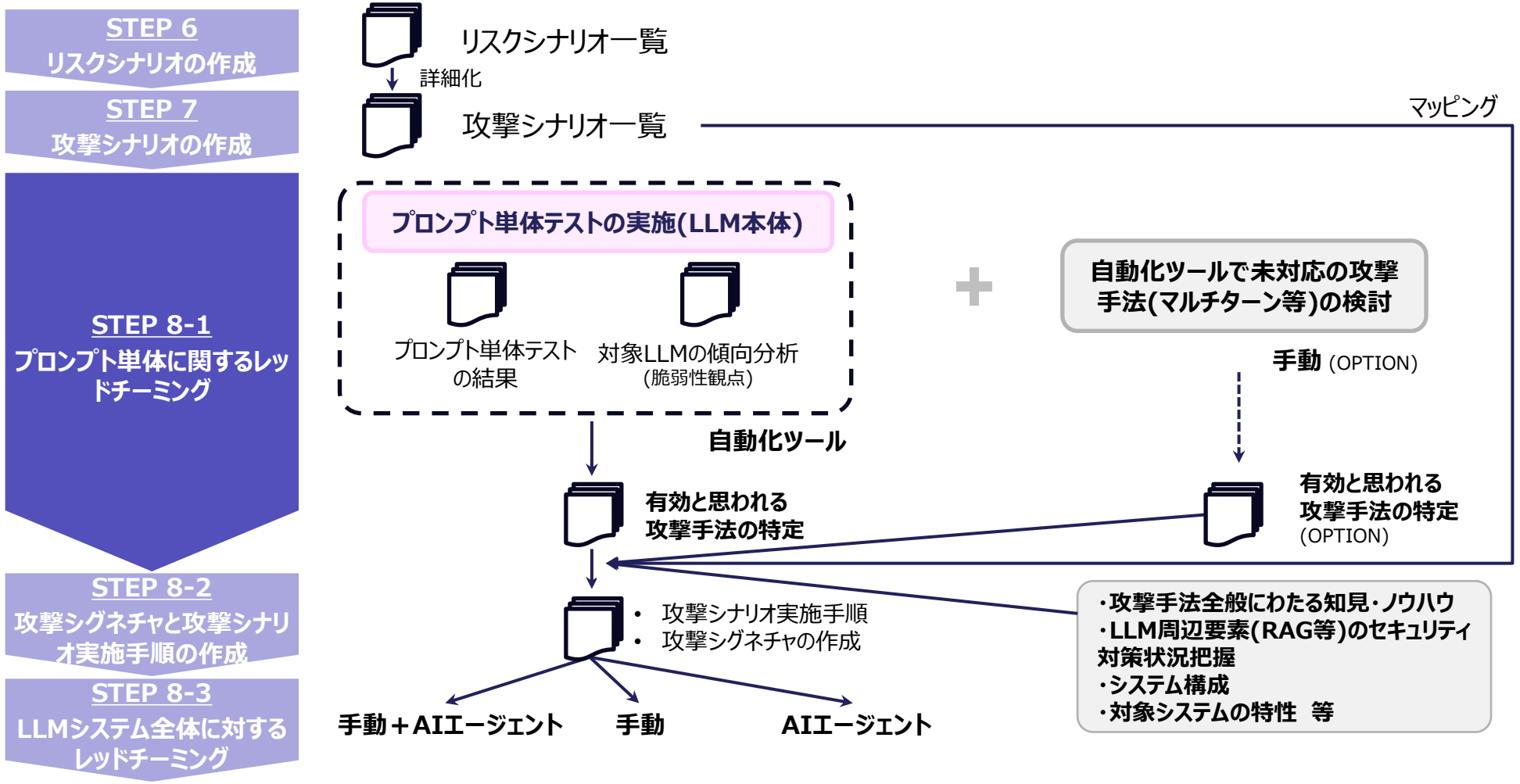
実施ポイント

- プロンプト単体に関するRTでは、対象システムに依存せず準備できる膨大な数の攻撃シグネチャを入力し、攻撃が有効なものを特定する。次頁でプロンプト単体に関するRTの位置づけを説明する。
- 自動化ツールを活用しプロンプト単体に関するRTを行った結果を【成果物例：リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果】※に示す。

※ RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果を参照

【参考】プロンプト単体に関するRTの位置づけ

- プロンプト単体に関するRTでは、対象システムに依存せず準備できる膨大な数の攻撃シグネチャを入力し、攻撃が有効なものを特定する。
- 膨大な数のプロンプトインジェクションを実施するため、自動化ツールを活用することが望ましいが、マルチターンなど、ツールが未対応の攻撃手法は手動での実施も検討する。



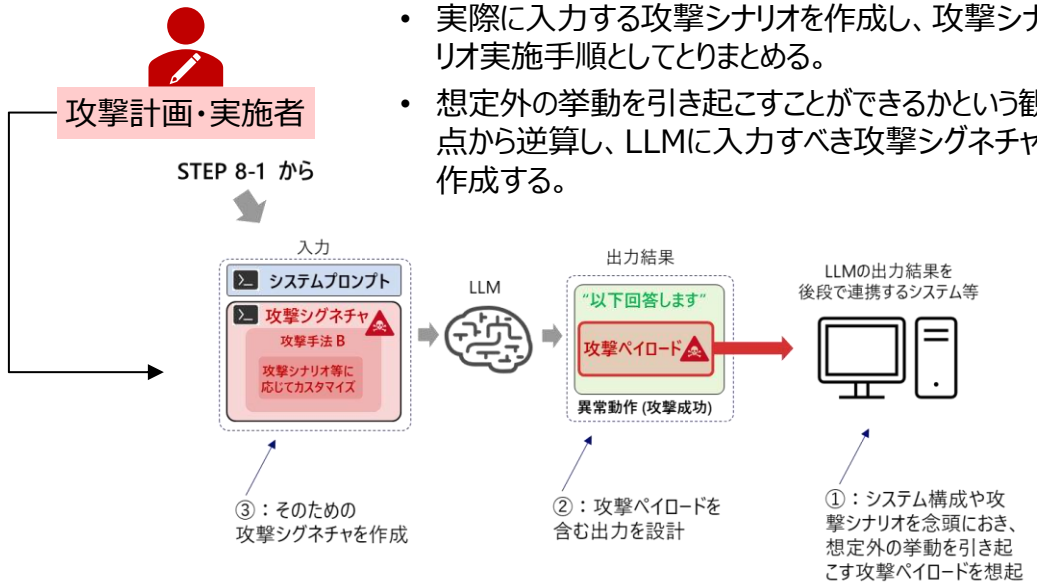
【詳細】（STEP 8-2）攻撃シグネチャと攻撃シナリオ実施手順の作成

実施事項

- プロンプト単体テストの実施結果を踏まえて、実際に入力する攻撃シグネチャを作成する。
- 各攻撃シナリオに対して、一連の攻撃シナリオ実施手順を作成する。
- 攻撃シナリオ実施手順とは、具体的な攻撃シグネチャの入力や環境設定、攻撃の実行方法を整理し、再現可能な手順としてまとめたものである。

実施イメージ

レッドチーム



- 実際に入力する攻撃シナリオを作成し、攻撃シナリオ実施手順としてとりまとめる。
- 想定外の挙動を引き起こすことができるかという観点から逆算し、LLMに入力すべき攻撃シグネチャを作成する。

実施ポイント

- システムにリスクを引き起こす出力を検討し、攻撃計画・実施者にとって望ましい出力を引き出すための入力（攻撃シグネチャ）を逆算して設計する。
- プロンプト単体テストで使用した攻撃シグネチャをそのまま利用するのではなく、修正する場合もある。修正の定まった手順はなく、攻撃手法に関する最新動向を把握することが重要となる。
- 攻撃シグネチャの作成例を【成果物例：レッドチーミング実施結果報告書】※1に示す。また、実際に作成した攻撃シナリオ実施手順を【成果物例：リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果】※2に示す。

※1 RT手法ガイド＞別添（成果物例）＞レッドチーミング実施結果報告書を参照
※2 RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果を参照

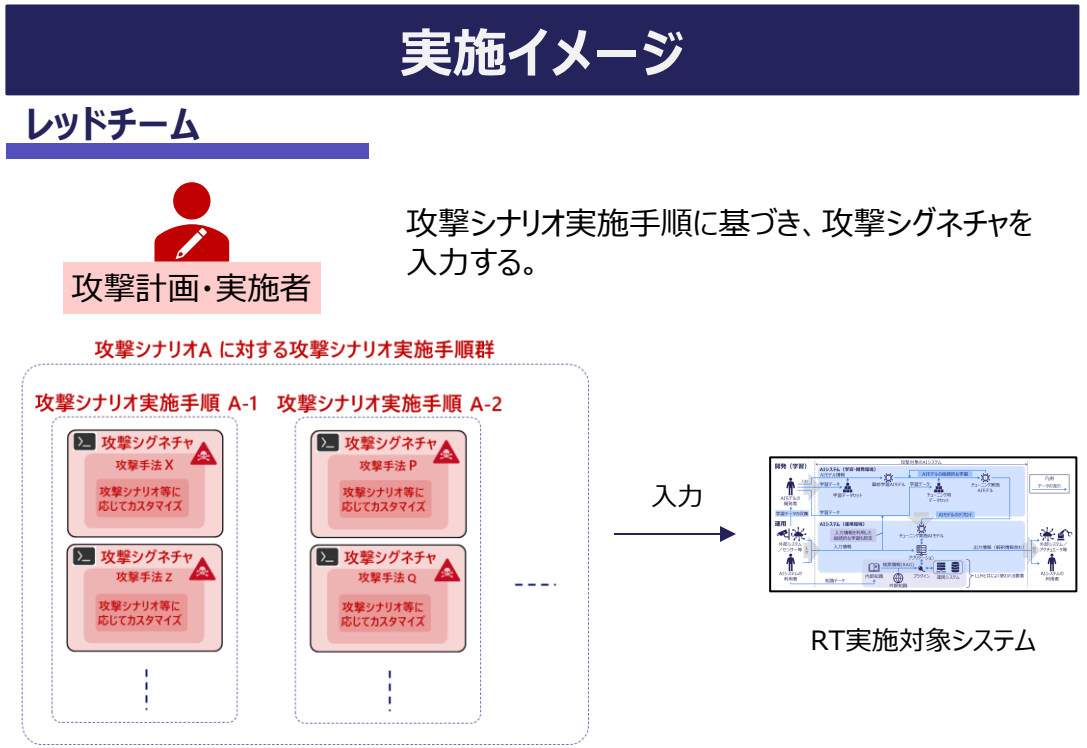
【詳細】（STEP 8-3）LLMシステム全体に対するレッドチーミング

AISI Japan AI Safety Institute

実施事項

- **STEP 8-2**にて作成した攻撃シナリオ実施手順に基づき、一連の攻撃シグネチャを対象AIシステムに入力し、結果を検証する。
- 攻撃シグネチャに対する出力結果を確認し、攻撃計画・実施者にとって望ましい出力になるよう攻撃シグネチャを修正し再度入力する。
- 攻撃計画・実施者は攻撃シグネチャの入力を複数回実施する。

実施イメージ



実施ポイント

- LLMシステムでは、その挙動・振る舞いが確率的・非決定的であるため、複数回繰り返すことで攻撃が成功する場合がある。そのため、同一の攻撃シグネチャの場合でも複数回攻撃試行することが望ましい。
- 攻撃シナリオを実施した結果を【成果物例：リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果】※に示す。

※ RT手法ガイド＞別添（成果物例）＞リスクシナリオと攻撃シナリオの作成及び攻撃シナリオの実施結果を参照

【概要】（STEP 9）実施中の記録、（STEP 10）実施後の処理

【凡例】 ：攻撃計画・実施者 ：AIシステムに関連する有識者 ：対象AIシステムの開発・提供管理者 ：その他関連ステークホルダー ：経営層 プロジェクトチーム レッドチーム

第2工程：攻撃計画・実施

レッドチーム

STEP 9 実施中の記録取得

- 実施されたRTの詳細を証跡として保存するため、RTの実施中の記録を取得する。

レッドチーム

・実施中の記録を取得し、報告書に記載して関係者に共有
・手動によるRTの場合はすべての攻撃シグネチャを取得
・攻撃成功した場合はスクリーンショットを取得

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RTの攻撃終了の連絡を行う。
- RT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。

レッドチーム

・関連するステークホルダーに対して以下を依頼
  ・RTの終了連絡
  ・一時的に作成したアカウントの削除や設定変更

【詳細】（STEP 9）実施中の記録

実施事項

- 実施されたRTの詳細を証跡として保存するため、RTの実施中の記録を取得する。
- 取得した記録については、組織の文書管理規程や機密情報の取扱規程等に従い、適切な保護対策を施し、決められた期間、保存する。

実施イメージ

レッドチーム

攻撃計画・実施者　AIシステムに関する有識者

以下に留意してRT実施の記録を取得する。

- 自動化ツールによるRTの場合、ツールのログ取得機能によりログを取得する。
- 手動によるRTの場合は、経路途中にプロキシを設置し、通過するすべての攻撃シグネチャを取得する方法もある。
- 攻撃成功した際には、LLMの出力結果などスクリーンショット(画面キャプチャ)を取得する。

実施ポイント

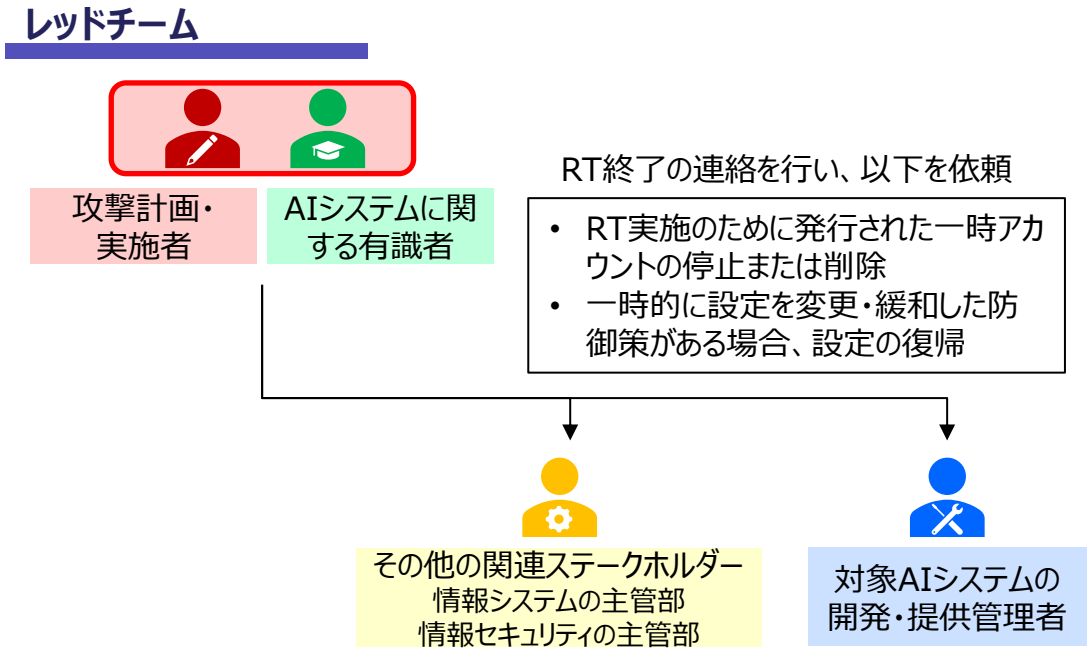
- 自動化ツールによるRTの場合、ツールのログ取得機能によりログを取得する。手動によるRTの場合は、経路途中にプロキシを設置し、通過するすべての攻撃シグネチャを取得する方法もある。
- 攻撃が成功した場合、スクリーンショット（画面キャプチャ）を取得する。

【詳細】（STEP 10）実施後の処理

実施事項

- 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攻撃シナリオの実施終了の連絡を行う。
- RT用の一時アカウントを削除する。また、一時的に設定を変更・緩和した防御策がある場合は、変更した設定を元に戻す。

実施イメージ



実施ポイント

- RT終了後は、RT実施のために一時的に発行したアカウント、変更した設定を元に戻す。

3. 各工程の解説
第3工程：結果のとりまとめと改善計画の策定

【概要】（STEP 11）実施計画の分析、（STEP 12）レッドチーミング実施結果報告書の作成

AISI Japan AI Safety Institute

【凡例】 ：攻撃計画・実施者 ：AIシステムに関連する有識者 ：対象AIシステムの開発・提供管理者 ：その他関連ステークホルダー ：経営層 プロジェクトチーム レッドチーム

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、RT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追加確認を行う。

レッドチーム

・実施結果の分析
・必要に応じて追加確認、協議 →

STEP 12 レッドチーミング実施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、RTの概要として示す。
- レッドチーミング実施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。

レッドチーム

・ログや証跡等の収集・整理
・RTの概要の作成
・レッドチーミング実施結果報告書の作成

結果報告書 →

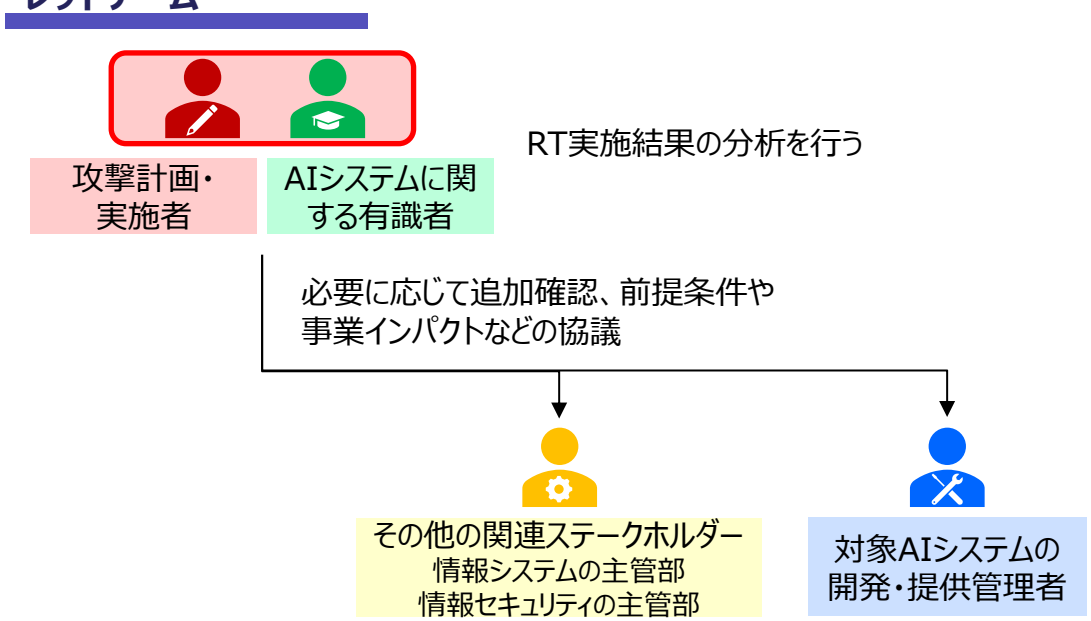
・関係者レビュー・報告

【詳細】（STEP 11）実施結果の分析

実施事項

- 攻撃計画・実施者は、RT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。その後、発見された脆弱性の前提条件の確認、引き起こされる想定被害や事業インパクトなどの協議を行う。
- 重大かつ緊急性の高い脆弱性が発見された場合は関係者に至急共有し、対策を検討する。

実施イメージ



実施ポイント

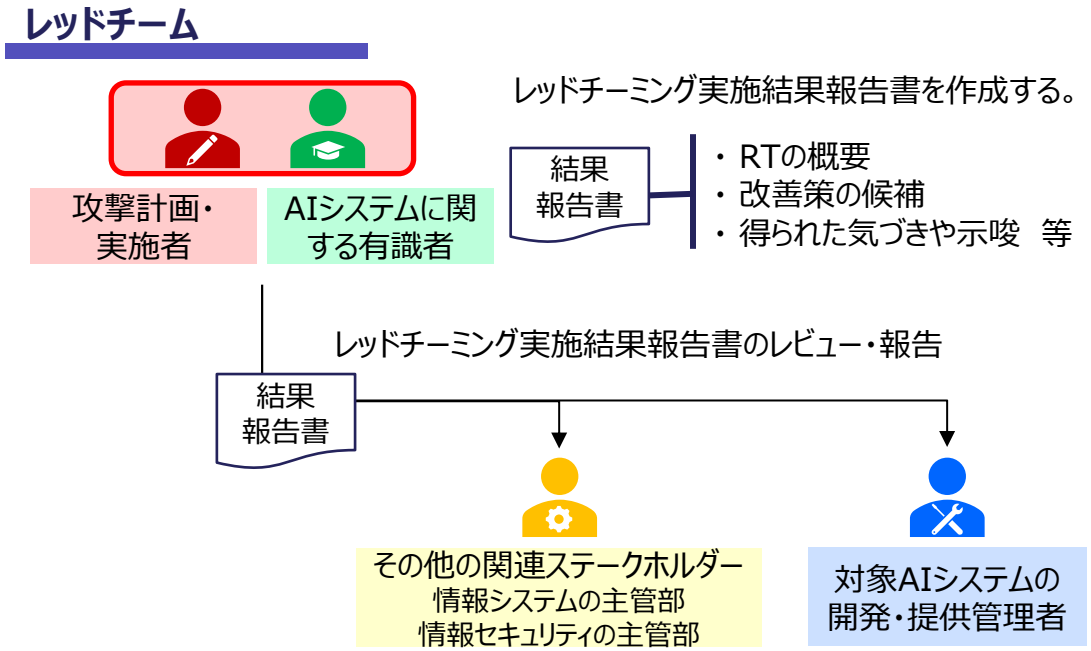
- 攻撃計画・実施者は、関係部署と連携し、必要に応じて分析に必要な情報の追加確認を行う。また、発見された脆弱性の前提条件の確認、引き起こされる想定被害や事業インパクトの認識を合わせる。

【詳細】（STEP 12）レッドチーミング実施結果報告書の作成

実施事項

- 攻撃計画・実施者は、発見された脆弱性を元に、ログや証跡等を揃え、RTの概要として示す。
- **レッドチーミング実施結果報告書**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。

実施イメージ



実施ポイント

- レッドチーミング実施結果報告書作成の段階では事業インパクトの分析などは行わず、結果の事実確認を目的とする。
- レッドチーミング実施結果報告書では、入力した攻撃シグネチャの説明や、結果がどのように有害であるのかを補記することで、関係者と円滑に共通理解を深めることができる。
- レッドチーミング実施結果報告書の例を【成果物例：レッドチーミング実施結果報告書】※に示す。

※ RT手法ガイド＞別添（成果物例）＞レッドチーミング実施結果報告書を参照

【概要】（STEP 13）最終報告書の作成～（STEP 15）フォローアップ

AISI Japan AI Safety Institute

【凡例】 ：攻撃計画・実施者 ：AIシステムに関連する有識者 ：対象AIシステムの開発・提供管理者 ：その他関連ステークホルダー ：経営層 プロジェクトチーム レッドチーム

第3工程：結果のとりまとめと改善計画の策定

プロジェクトチーム

STEP 13 最終報告書の作成と報告

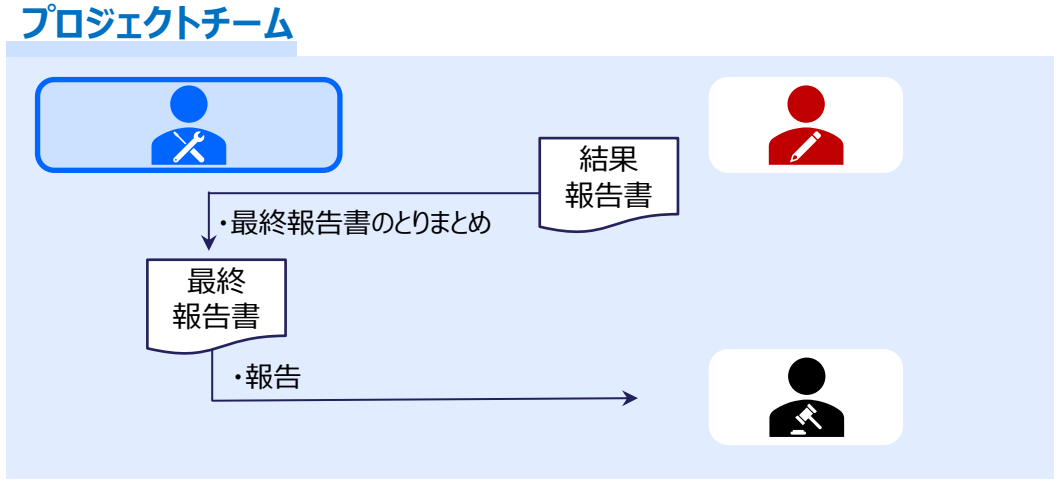
- 対象AIシステムの開発・提供管理者は、攻撃計画・実施者によるレッドチーミング実施結果報告書をもとに、最終報告書をとりまとめる。
- 最終報告書について、必要に応じて経営層へ報告する。

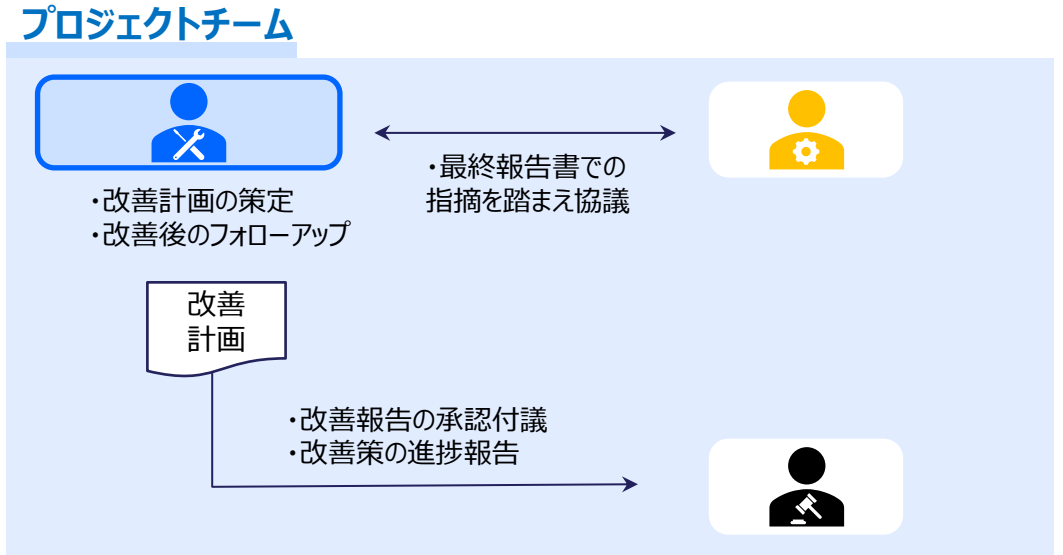
STEP 14 改善計画の策定と実施

- 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RTを実施し、発見された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。



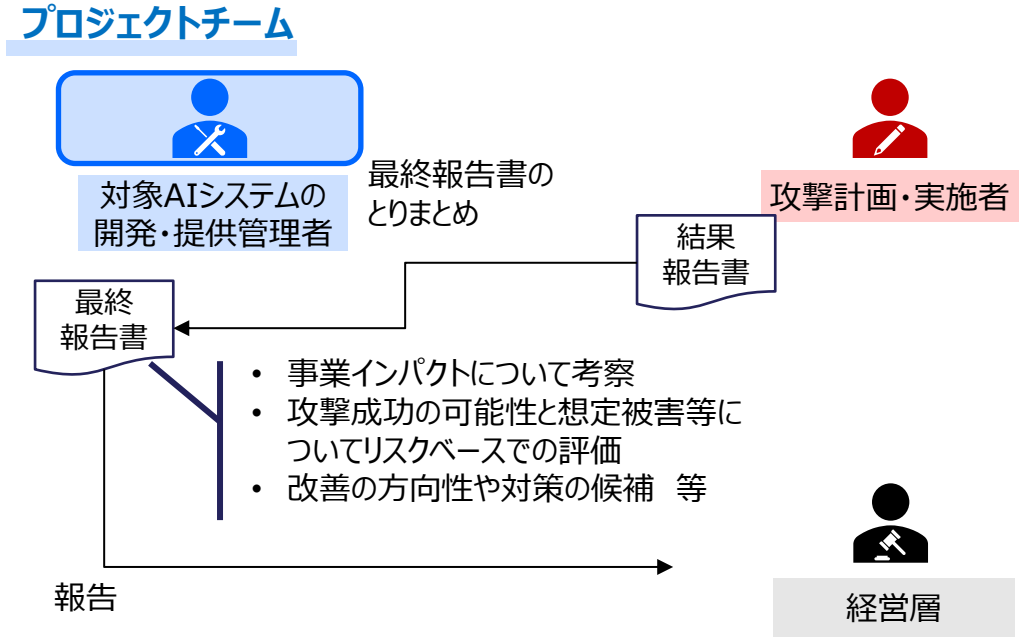


【詳細】（STEP 13）最終報告書の作成と報告

実施事項

- 対象AIシステムの開発・提供管理者は、攻撃計画・実施者によるレッドチーミング実施結果報告書をもとに、**最終報告書**をとりまとめる。
- 最終報告書では、レッドチーミング実施結果報告書に記載されたシステム観点からの想定リスクをもとに、実際の事業における観点に基づく事業インパクトについての考察を行う。その後、攻撃成功の可能性と想定被害等についてリスクベースでの評価を行う。
- 最終報告書について、必要に応じて経営層へ報告する。

実施イメージ



実施ポイント

- レッドチーミング実施結果報告書の目的が事実確認であったのに対して、最終報告書の目的は、実際の事業における観点に基づく事業インパクトについて考察し、リスクベースでの評価を行うことである。
- リスクベースでの評価の指標は、リスクシナリオ作成時に検討した総合的なリスク評価の指標が参考になる。
- 最終報告書の例を【成果物例：最終報告書】※に示す。

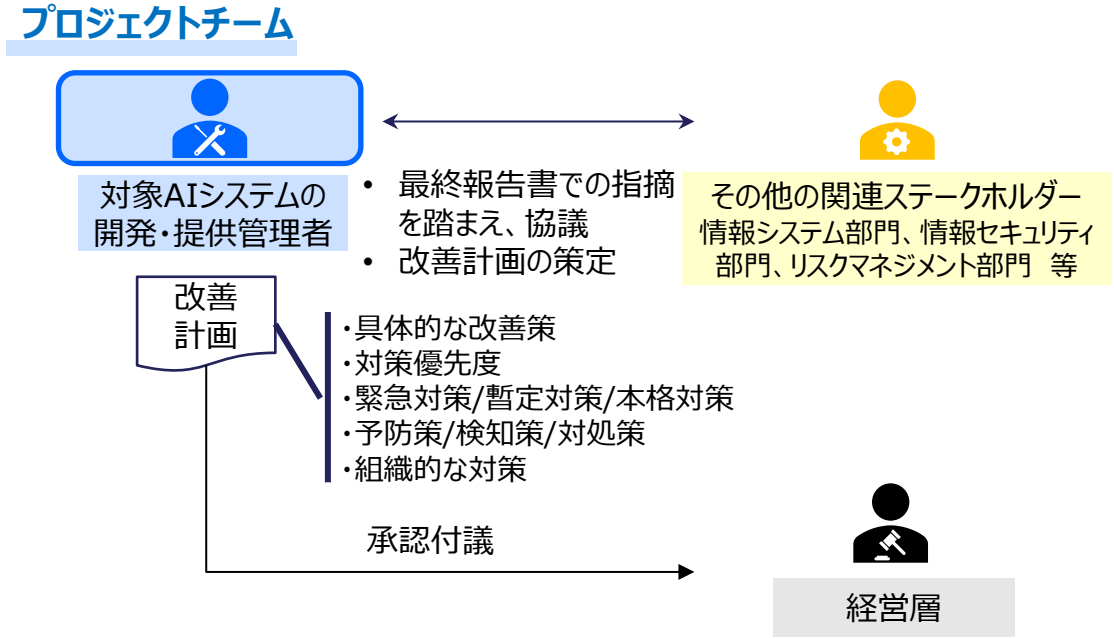
※ RT手法ガイド＞別添（成果物例）＞最終報告書を参照

【詳細】（STEP 14）改善計画の策定と実施

実施事項

- 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。
- 改善計画に対する経営層の承認を得たのち、改善計画を確定する。

実施イメージ



実施ポイント

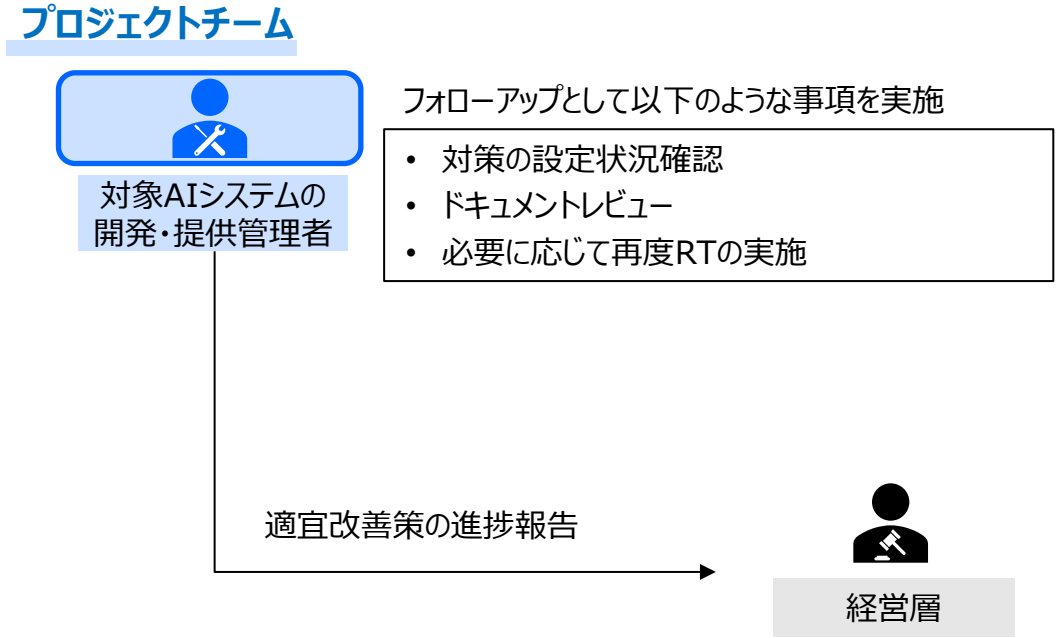
- 改善計画を策定する際は、最終報告書で挙げられた改善策について、事業インパクトに応じた緊急度やリスク度合いを考慮し、優先度付けを行う。
- システム面の改善策だけでなく、運用プロセスの見直しなど組織的な対策も含めて検討する。

【詳細】（STEP 15）改善後のフォローアップ

実施事項

- 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RTを実施する。その後、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。

実施イメージ



実施ポイント

- 進捗管理を徹底し、必要に応じて再度RTを実施するなど、継続的な改善サイクルを回すことで、実効性のある改善活動を推進できる。

AISI

Japan AI Safety Institute