資料47－４

AIセキュリティに関する検討について

令和７年９月

サイバーセキュリティタスクフォース事務局

AIセキュリティ分科会（案）

背景・目的

- 生成AIの社会実装が急速に進む中、AIのセキュリティ確保が重要な課題となっており、「デジタル社会の実現に向けた重点計画」（令和７年6月13日閣議決定）では、総務省が、今年度末までに、生成AIとセキュリティのガイドラインを策定・公表することとされている。
- また、AIの安全安心な活用促進については、「AI事業者ガイドライン」（総務省・経済産業省）において「セキュリティ確保」が共通指針の一つに位置付けられ、これを踏まえ、関係省庁・関係機関により構成される「AIセーフティインスティテュート(AISI)」※が、AIに対する脅威の特定等を行っている。
- 本分科会は、このような状況を踏まえ、「サイバーセキュリティタスクフォース」の下に開催される会合として、AIに対する脅威への技術的対策について検討を行う。

※　AIの安全安心な活用が促進されるよう 官民の取組を支援する機関。統合イノベーション戦略2024に基づきIPAに設置。

主な検討事項

- AI開発者及び提供者における、AIに対する脅威への技術的対策の在り方
- 上記対策の普及啓発の在り方

構成員（敬称略・50音順）

秋山　満昭	NTT株式会社　社会情報研究所　上席特別研究員
新井　悠	株式会社NTTデータグループ　技術革新統括本部 品質保証部情報セキュリティ推進室 NTTデータCERT担当 エグゼクティブ・セキュリティ・アナリスト
石川　朝久	東京海上ホールディングス株式会社 IT企画部サイバーセキュリティグループ Distinguished Cyber Security Architect
篠田　佳奈	株式会社BLUE　代表取締役
高橋　健志	NICTサイバーセキュリティ研究所 AIセキュリティ研究センター　研究センター長
北條　孝佳	西村あさひ法律事務所・外国法共同事業 パートナー弁護士
披田野　清良	株式会社KDDI総合研究所 セキュリティ部門 エキスパート
森　達哉	早稲田大学　理工学術院　教授

※以上に加えて、AI開発者・提供者の構成員を調整中

オブザーバ：国家サイバー統括室、内閣府、デジタル庁、経済産業省、AISI

スケジュール

令和７年９月　　第１回分科会（以降、月１回程度の開催を想定）
令和７年12月頃　分科会とりまとめ
令和７年度内　　総務省ガイドラインの公表

本分科会における検討のスコープの方向性について①

- 生成AIを含むAIの安全安心な活用については、総務省・経済産業省が、**「AI事業者ガイドライン」**（令和6年4月第１版、令和７年3月第１．１版）において、**AIの事業活動を担う主体が取り組むべき事項**を10の**「共通指針」**として整理。
- **AISIは**、これらの事項の中から、**ＡＩセーフティを向上する上での重要要素を特定**し、各重要要素について、**ＡＩセーフティを評価する観点を設定**（**「ＡＩセーフティに関する評価観点ガイド」**（令和6年9月第1.00版、令和7年3月第1.10版））。
- 本分科会では、これらの要素・観点のうち、**「セキュリティ確保」に関するものを中心的に取り扱う。**具体的には、**「不正操作による機密情報の漏えい、AI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」に対する脅威への対策**について中心的に取り扱う。

○AIセーフティにおける重要要素・AIセーフティ評価の観点

AIセーフティにおける重要要素 ＼ 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
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		●	●				●	●	●	●

本分科会で中心的に取り扱う領域

（出典）AISI「AIセーフティに関する評価観点ガイド」から一部抜粋し、事務局にて追記・加工

本分科会における検討のスコープの方向性について②

- AISIは、AIセーフティ評価の観点の整理とともに、AIシステムの脅威を攻撃者の観点から評価するための基本的な重要事項を設定（**「AIセーフティに関するレッドチーミング手法ガイド」**（令和6年9月第1.00版、令和7年3月第1.10版））。ここで、**LLMシステムへの代表的な攻撃手法を整理**し、脅威を評価する枠組を提示。また、令和7年3月には、**「AIシステムに対する既知の攻撃と影響」を整理、公表。**
- 本分科会では、このような脅威への技術的対策について検討。総務省は、対策例をガイドラインとして公表予定。

○LLMシステムへの代表的な攻撃手法

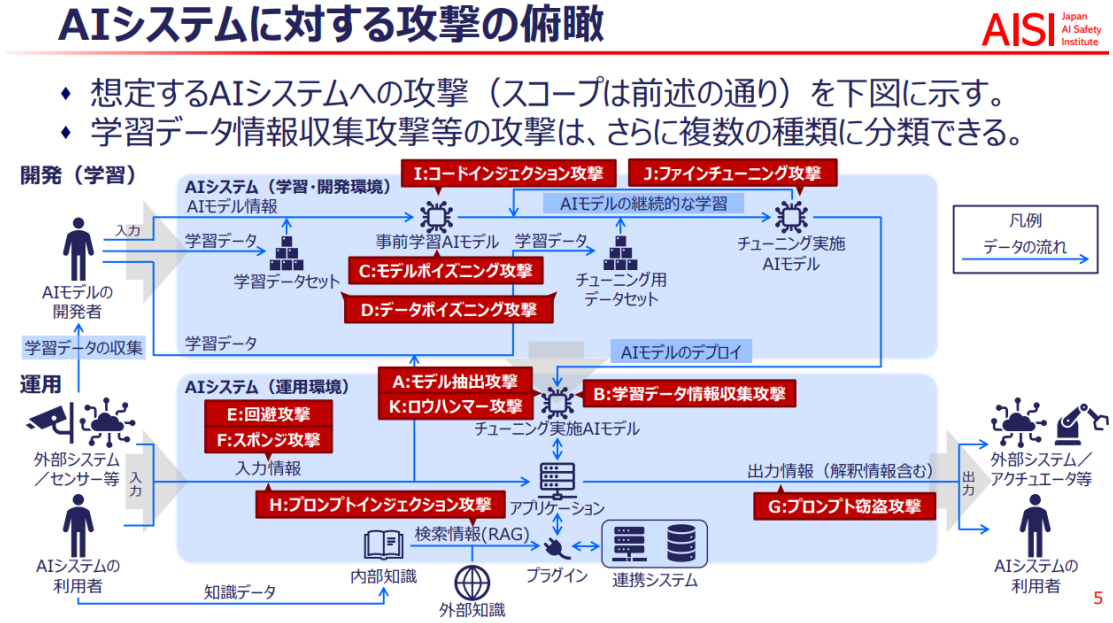
LLMシステムへの代表的な攻撃手法

➢ 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

（出典）AISI「AIセーフティに関するレッドチーミング手法ガイド」から一部抜粋

○AIシステムに対する既知の攻撃と影響



（出典）AISI「AIシステムに対する既知の攻撃と影響」から一部抜粋

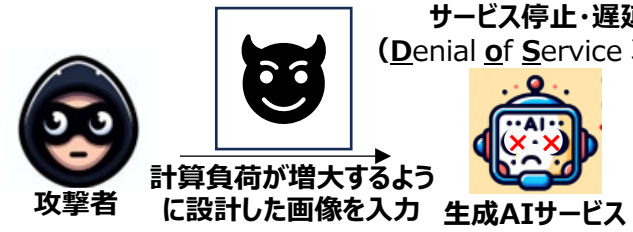
本分科会では、このような脅威への技術的対策について検討

（参考）AIセキュリティガイドライン（仮称）で扱う脅威例

脅威例

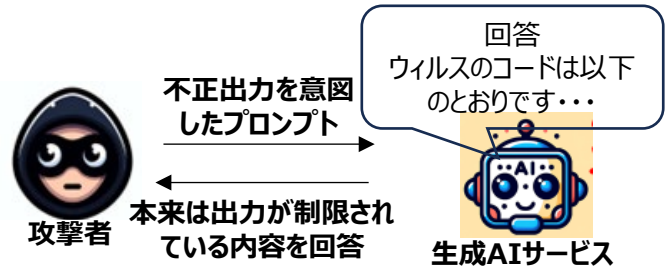
脅威例①：入力画像の細工によるDoS攻撃(スポンジ※攻撃)

- 生成AIへの入力画像に特定の細工をすると、巨大な計算負荷を発生させることができる。
- 攻撃者は、AIの利用のフェーズにおいて、この性質を悪用し、計算負荷の高い画像を入力することで、サービスの停止・遅延を引き起こす。



脅威例②：プロンプト・インジェクション

- 攻撃者は、AIへのプロンプト(指示)を巧妙に細工し、AIが出力を制限している事項（例：機密情報、ウィルスの作成方法など悪用可能な情報）を出力させる。



対策例

- 処理コストの高い画像の遮断やリサイズを実施するフィルタリング機構を設け、AIへの過剰負荷を防止

- 入力されたプロンプトを別のAIが検証し、不正出力を意図したプロンプトが検知された場合に、入力を拒否

※　「スポンジ」とは攻撃用データを指す比喩であり、悪意のあるデータがコンピューターの処理能力・消費電力を「吸い尽くす」性質を表す。

（参考）関連の政府決定

「デジタル社会の実現に向けた重点計画」（令和7年6月13日 閣議決定）

本文　４．取組の方向性と重点的な取組

（4）安全・安心なデジタル社会の形成に向けた取組

⑤ サイバーセキュリティの確保

・AI を積極的に活用したセキュリティ対策（AI for Security）を推進するとともに、セキュリティ分野における AI の安全な利用（Security for AI）に向けた環境整備の取組などを進めていく。

重点政策一覧　４．取組の方向性と重点的な取組

〇[No.4-25] 生成 AI 等を活用したサイバーセキュリティ対策強化

具体的な目標：【「Security for AI」の取組】

・2025年度末までに、生成AIとセキュリティのガイドラインの策定・公表を 実施。2026年度には、2025年度に公表したガイドラインの更新等を実施。

主担当府省庁：総務省 関係府省庁：―

「サイバーセキュリティ2025」（令和7年6月27日 サイバーセキュリティ戦略本部決定）

第４部　2024年度のサイバーセキュリティ関連施策の取組実績、評価及び今年度の取組

4　横断的施策

4.1　研究開発の推進

（3）中長期的な技術トレンドを視野に入れた対応

＜2025年度年次計画＞

・生成AI等の技術は日に日に進歩を続けており、生成AI等を悪用したサイバー攻撃が国内外で発生している。そのため、令和6年度に調査・分析を行ったAIを悪用したサイバー攻撃の脅威への対策等の、開発者・提供者が留意すべきセキュリティリスクを取りまとめ、AIの安心・安全な開発・提供に向けたセキュリティガイドラインを策定する。

・引き続き、総務省において、生成AIをはじめとするAI技術がサイバーセキュリティに与える影響について、正の側面と負の側面の双方から、調査を実施し、必要な対策について検討を進める。

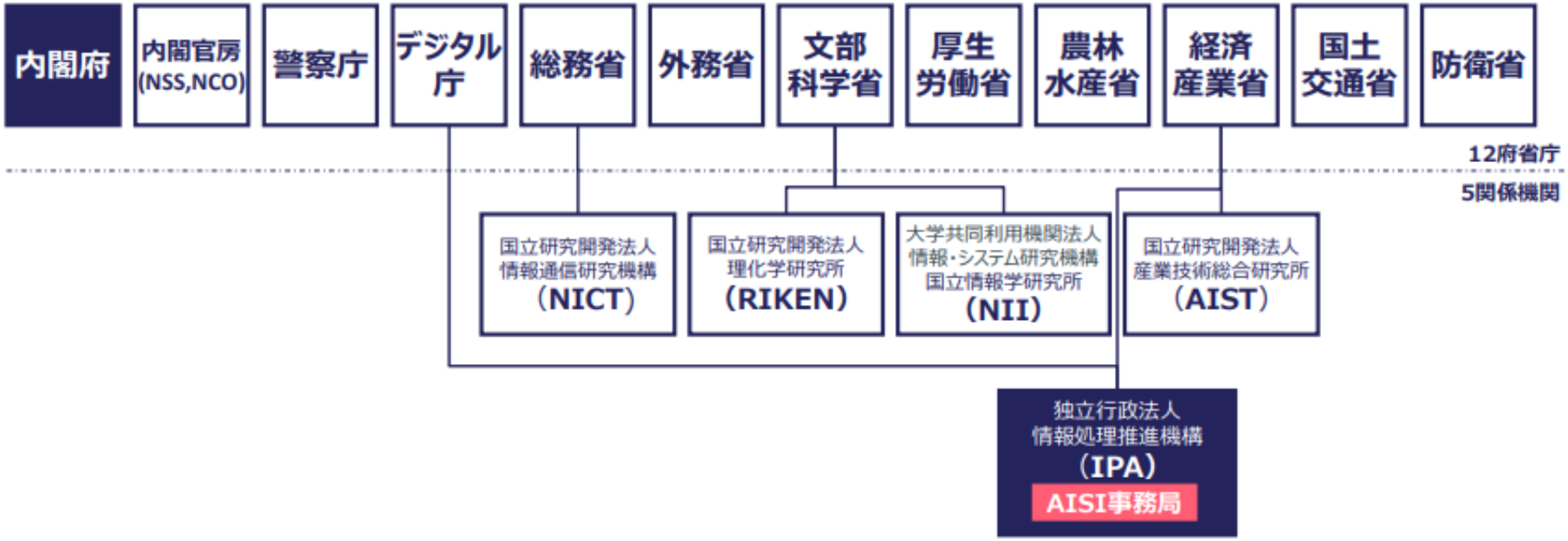
AISIの関係府省庁・機関

AISI Japan AI Safety Institute

AISIは、12府省庁・5関係機関が横断的に参画する**政府関係機関**

事務局は経済産業省とデジタル庁を所管官庁としている**IPA内に設置**

＊2025年4月時点



9

（出典）AISI「AIセーフティ・インスティテュート（AISI）について」から抜粋