研究開発構想（個別研究型）

人工知能（AI）が浸透するデータ駆動型の経済社会に必要な AI セキュリティ技術の確立「安心安全な AI 利活用の為の知識・技術の体系化と知識集約環境構築」

研究開発実施報告書（年次）

令和6（2024）年度

研究代表者
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

１．当該年度における研究開発の実施概要

（１）研究開発概要

本研究開発では、AI セキュリティを社会に普及させる ことを目的とし、最新動向に配慮しながら AI セキュリティに関する知識・技術を網羅的に体系化します。また、体系化された知識に基づき論文や記事などの文献を効率的に収集・分類し、注意喚起や対策の普及啓発を効果的に行うための知識集約環境を構築します。本研究開発では、実際に構築した知識集約環境を利用して AI セキュリティに関する情報を発信しながら、その効果を実証します。

（２）実施内容と成果の概要（研究開発開始から当該年度末まで）

令和 6(2024)年度

AI セキュリティに関する情報を総合的に発信する AI セキュリティポータルを新たに開設し、運用を開始しました。本サイトの主なコンテンツは、AI セキュリティマップ、文献データベース、解説記事であり、これらは当該年度の研究開発の成果に基づいて作成しています。AI セキュリティマップに関する研究開発では、AI セキュリティに関する論文やガイドラインなどを網羅的に調査し、AI に対する攻撃や防御に加えて、AI の毀損や悪用が人や社会に与える負の影響を体系的に整理しています。文献データベースに関する研究開発では、AI セキュリティに関する Web 上の文献を自動的に収集する環境を整備するとともに、収集した文献に自動的に分類ラベルを付与する技術を開発しています。本分類技術は、新たなトレンドが出現した際に、LLM を利用してそのトレンドに合わせて新しい分類ラベルを創出する機能を特徴としており、従来技術のように分類モデルを再構築することなく、分類ラベルを付与できます。解説記事については、一般ユーザと専門家を対象としながら、専門性や嗜好性などに配慮し、発信対象毎に項目や発信方法を変えて執筆しています。今後はユーザ調査などを通じて、前述のコンテンツの妥当性を評価するとともに、改善点についても整理しながら、継続的にサイトを更新していきます。

２．主たる研究分担者一覧

なし