電子情報通信学会総合大会2025 依頼シンポジウムセッション [AI-4]

LLMとコグニティブセキュリティによる安全な社会の構築

電子情報通信学会 光輝会特別研究専門委員会

📅 2025年3月26日 09:00 AM - 12:10
🏢 東京都市大学 世田谷キャンパス 7号館1階 71A

本セッションでは、生成AIをセキュリティの観点で俯瞰し、科学技術・産業・倫理・法学等のさまざまな研究者/専門家の見解を共有するとともに、セキュリティ技術により様々な世代が安心して利用できる知の道具としての AI の可能性についての議論を行います。

コグニティブセキュリティの確立に向けて：情報操作時代における認知能力防御の新たな戦略

秋山 満昭
NTT社会情報研究所　上席特別研究員

コグニティブセキュリティとは、情報操作や心理的攻撃といった認知プロセスを混乱させる脅威から、人間の認知能力を守る新たなセキュリティ領域である。近年、LLM（大規模言語モデル）の普及によってこれらの脅威が増大し、個人や社会に深刻な影響を及ぼしている。本講演では、コグニティブセキュリティの基礎概念を整理し、認知プロセスを防御するために必要な技術要素、解決すべき課題、そして実現に向けた研究アプローチについて、サイバーセキュリティ研究を出発点として各種研究分野を横断しながら模索する.

LLM時代のフェイクニュース問題と対策

笹原 和俊
東京科学大学　環境・社会理工学院　教授

近年、ChatGPTやStable Diffusionなどの生成AIの進化により、リアルなフェイクニュースやディープフェイクの作成が容易になり、コグニティブセキュリティを脅かす社会問題として深刻化している。本発表では、計算社会科学の知見に基づき、生成AIによる偽情報の特徴とソーシャルメディアにおける拡散メカニズムを分析する。また、情報環境の変容を踏まえ、AIによって高度化する偽情報の検出・抑制に向けた行動科学的・技術的対策の現状と展望について議論する。

LLMの安全性構築に向けて

関根 聡
情報学研究所　大規模言語モデル研究開発センター　特任教授

LLMの安全性構築に向けて、主に情報学研究所大規模言語モデル研究開発センターで行われている研究開発を紹介する。LLMの安全性とは、ユーザーや社会に対して悪影響を与えるような情報を提供してしまうリスクに対する対策である。リスクとは具体的にバイアス、差別、反公序良俗、対話リスク、情報漏洩、悪用、偽情報、誤情報などであり、それらに対して、インストラクションの構築とそれを利用したSFTでの学習、攻撃的プロンプトの収集、文書フィルタリングなどにより安全性の実現を目指している。また、国際的な安全性確保のための組織AISIにおける活動も紹介し、LLMの安全性構築の活動について紹介する。