R&Dフェーズ｜研究　**実用化時期**｜FY28-

Mehr Licht!
More Photonics for Intelligence

#C18

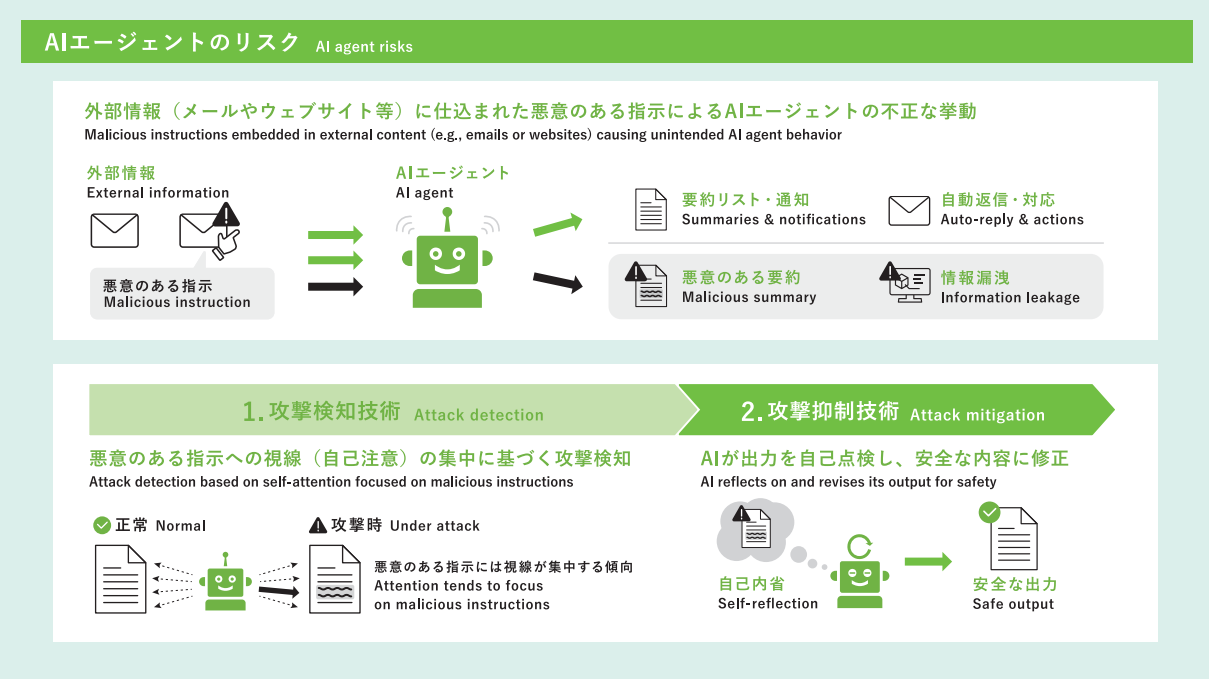
#安心・安全#業務効率化#DX・AX（デジタル変革）

外部情報に仕込まれた指示からAIを守る技術

AIエージェントへの外部情報を用いた攻撃を検知して危険を抑制し、安全な業務活用を実現します

技術課題・ゴール

生成AIの業務利用が広がり、AIエージェントがWebサイトや文書に埋め込まれた悪意ある指示に従う「間接プロンプトインジェクション」の問題が顕在化しています。既存対策は入力文面の判定に偏り、巧妙な攻撃を見逃す恐れがあります。本技術はAIの内部状態から異常反応を検知し、危険な行動を抑制します。これにより、情報漏洩や誤操作のリスクを低減し、企業がAIエージェントを安心して活用できる社会を実現します。



技術ポイント

要素技術

・攻撃検知技術：AI内部の自己注意の偏りを監視し、外部情報に埋め込まれた悪意ある指示へ過度に注目している状態を検知

・攻撃抑制技術：検知した注意の偏りをAI自身に振り返らせ、攻撃の影響を除去して再判断し、安全な応答・行動へ修正

市中技術差異点

・既存技術による入力文面の分析では見逃してしまう攻撃も、タスク処理中のAI内部の自己注意の偏りから検知

・既存技術では攻撃を受けると処理を中断するが、AIの自己点検で攻撃の影響を除去することで、タスクを安全に継続

適用ビジネス/ユースケース

情報技術

事務処理、情報収集、システム開発など、外部のWebサイトや文書を参照しながら自律的に判断・実行するAIエージェントに適用できます。AIエージェントの内部状態を常に監視することで、外部情報の指示に従おうとすることを検知し、危険な振る舞いを抑制します。機密情報や業務システムを扱う企業利用で有効です。

出展組織

NTT株式会社 社会情報研究所