

東北大学言語AI研究センターと共同でAIエージェントの安全性・信頼性に関する研究を開始

東北大学言語AI研究センター（センター長：鈴木 潤）と株式会社Acompany（アカンパニー、愛知県名古屋市中区、代表取締役CEO 高橋 亮祐、以下Acompany）は、共同でAIエージェントの安全性・信頼性に関する研究（以下、本研究）を開始することをお知らせします。本研究により、AIエージェントに対する脅威の理解と対策技術の確立を目指します。



共同研究の背景

AIエージェントに様々なツールやシステムへのアクセス権限を与え、複雑なタスクを自動化する取り組みが広がっています。一方で、自律的に探索や判断を行いながらタスクを遂行するAIエージェントには、従来の機械学習や生成AIとは異なる新たな脅威が存在します。

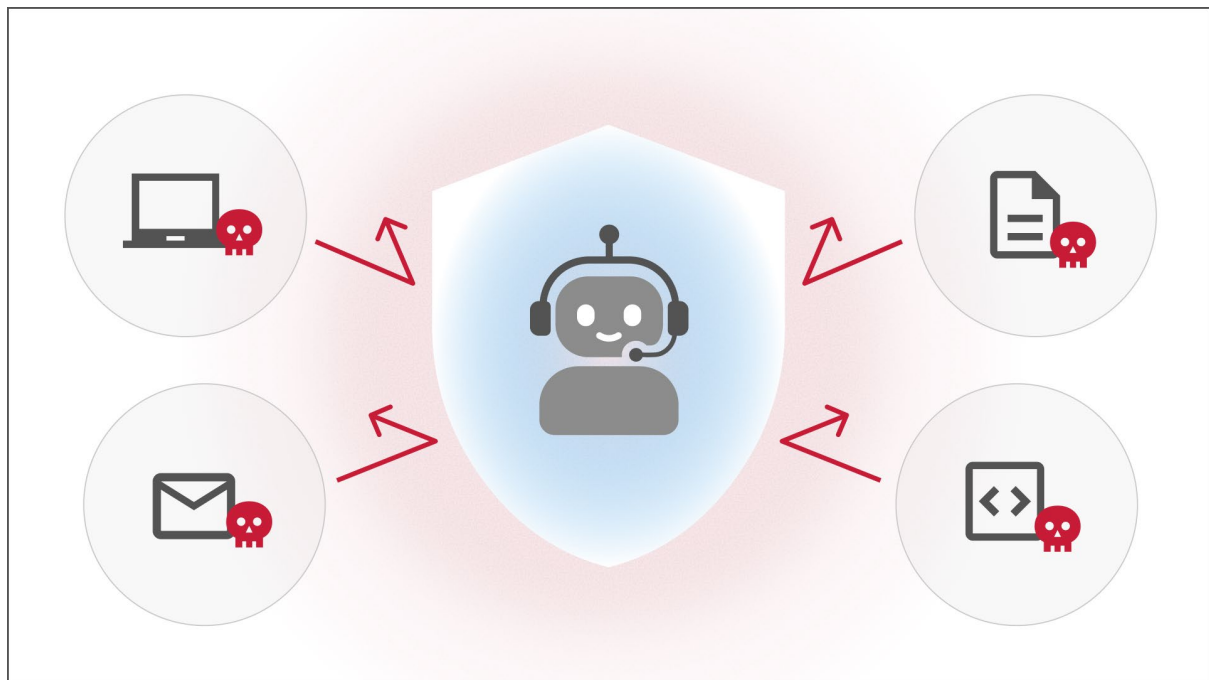
例えば、AIエージェントが、Webページや文書、メール、外部ツールから得られる情報などに含まれる悪意のある指示に接触することで、意図しない挙動へ誘導されたり、与えられた権限や想定された制約を逸脱したりするリスクが指摘されています。また、AIエージェントが置かれる環境や利用可能なツールは動的に変化するため、個別・静的な対策だけでは十分とは限りません。

こうした背景から、AIエージェントの脅威を理解・評価し、その変化にも対応可能な安全性・信頼性の確保が重要になっています。

研究内容

本研究では、東北大学言語AI研究センターが有する自然言語処理および機械学習の知見と、Acompanyが蓄積してきたAIセキュリティ・プライバシーの知見を組み合わせ、AIエージェントに対する脅威の理解と対策技術の確立を目指します。

AIエージェントは、探索や評価を繰り返しながら自律的にタスクの達成を目指します。その過程では、悪意のあるアーティファクトによって意図しない挙動へ誘導されるリスクや、与えられた権限、あるいはハーネス ※1 によって定められた制約を逸脱するリスクなどが考えられます。



AIエージェントを取り巻く新たな脅威

本研究では、このようなAIエージェントの脅威について、その発生条件や影響を明らかにするとともに、安全性・信頼性を高めるためのモデルおよびハーネスの構成方法、脅威の検知技術、影響を抑制するための技術などを幅広く検討してまいります。

期待される成果と意義

本研究に先んじてAcompanyが公開したホワイトペーパー ※2 では、「秘密を守れるAI」を実現するための重要な要件の一つとして、AIの「振る舞いの信頼性」を挙げています。

AIを安全な実行環境で動作させるだけでは、AI自身が攻撃や悪意のある情報によって意図しない行動を取るリスクまで防ぐことはできません。特に、自律的に外部の情報やツールと関

わるAIエージェントの普及に伴い、その振る舞いをどのように安全かつ信頼できるものにするかは、今後ますます重要な課題になると考えています。

本研究を通じて、AIエージェントの振る舞いの信頼性を毀損し得る脅威やその成立条件を明らかにするとともに、それらに対する対策技術の確立を目指すことで、「秘密を守れるAI」の具現化と社会実装に貢献してまいります。

共同研究に関するコメント

東北大学言語AI研究センター センター長 鈴木 潤

AIエージェント技術の急速な発展に伴い、その安全性・信頼性の確保は、社会実装に向けた最重要課題の一つです。東北大学言語AI研究センターでは、大規模言語モデルをはじめとする言語AIの構築と原理解明に取り組み、モデルの振る舞いに関する知見を蓄積してきました。本共同研究では、Acompany社が培ってきたAIセキュリティ・プライバシーの知見と、私たちの言語AIの知見を持ち寄り、AIエージェントに対する脅威とその振る舞いへの理解を深め、安全性・信頼性を高める新たな技術の創出を目指します。また、学術的な知見の深化に加え、その成果を実社会で活用できる技術へとつなげたいと考えています。

Chief Scientist 兼 執行役員 VP of Research&Development 高橋 翼

企業がAIエージェントに業務を委ねるとき問われるのは、「何ができるか」だけでなく「どこまで任せられるか」です。安全な実行環境を用意するだけでは、AI自身が悪意のある情報に誘導され、意図しない行動を取るリスクは残ります。振る舞いそのものをどう信頼できるものにするか—これはAIの社会実装における本質的な課題です。言語AIの研究を牽引されてきた東北大学言語AI研究センターと知見を持ち寄り、技術と実務の両面から答えを出していきたいと考えています。

※1 ハーネス：AI エージェントが利用できるツール、権限、操作対象を制限する仕組み。AIエージェントへの動作ポリシー、指示などを含む

※2 Acompanyが公開したホワイトペーパー：「Confidential AI：Confidential Computingで実現する秘密を守れるAI」 (<https://www.acompany.tech/service/white-paper/confidential-ai/>)

株式会社Acompanyについて

社名：株式会社Acompany

代表者：代表取締役CEO 高橋亮祐

所在地：愛知県名古屋市中区栄二丁目1番1号 日土地名古屋ビル7階

設立 : 2018年6月

URL : <https://www.acompany.tech/>

事業内容 : Confidential Computingに関連した製品・技術と、機密データ活用に関するコンサルティングサービスの提供

本件に対するお問い合わせ

Acompanyに関するご質問・ご相談などはお気軽に下記からお問い合わせください。

<https://www.acompany.tech/contact>

東北大学大学院情報科学研究科広報室

koho_is@grp.tohoku.ac.jp