AIが信頼できるものであるために

AIが信頼できるものであるためには「倫理・品質」に加え「セキュリティ」を確保する必要がある。AIにはAI特有の脅威や攻撃手法が存在するので、セキュリティを確保するためには、AIを活用するシステムにおける脅威を把握して個々のリスクを評価した上で、適切な対策を実装する必要がある。

AIそのものを守るための Security for AI

社会でのAI利活用がさらに進むためには法制度の整備が進むこと、AIが信頼できるものであることの2点が必要な条件となると言われている。後者は、AIサービスを利用する企業やAIを活用したシステムを開発するすべての企業が考慮すべき事項である。

AIが信頼できるものであるためにはまず、AIが返す結果に差別がない等の「倫理」、AIが期待通りの結果を返す等の「品質」の確保が必要となる。倫理や品質を確保するために必要な性質はAIの利用者にとって想像しやすいものだと考えるが、AIが信頼できるものであるためにはもう1つ、「セキュリティ」の確保が必要となる。

AIとセキュリティの関係には2つの側面がある。1つは、システムのセキュリティ対策高度化のためのAI活用で、もう1つは、AIそれ自体を守るためのセキュリティである。

セキュリティ対策としてのAI活用は、AI for Securityと言い換えることができる。例えば、セキュリティ確保のためのプロセスである予防、検知、対処等にAIを活用することによる、対策の高度化や自動化・効率化が該当する。

かたや、AIに対する攻撃等の脅威からAIを守るためのセキュリティは、Security for AIと言い換えることができる。ここでは後者に焦点を当てて、問題の所在を概括的に示してみたい。

AIにおける脅威にはどのようなものがあるだろうか。どこに、何があるかを識別する「物体検知AI」を対象とした事例を挙げてみる。

図表　物体検知モデルにおけるSecurity for AIの例

図表は弊社オフィスで撮影した映像を対象に、AIで物体検知を行っている様子である。3人のうち左の2人は赤枠で「person」として検知されているが、右端の人物は、personとして検知されていない。これは、Tシャツに細工した特殊な模様をプリントして着用することで、模様がカメラに写っている人らしき存在がpersonとして検知されなかったことを示している。つまり、AIに対する攻撃が成功したというデモである。

ここで仮に、同じ仕組みのAIがビルの監視システムとして使用されている状況があったとしたらどうか。人として検知されることなく、ビルに侵入することが可能となるのである。AIを用いた監視システムは機能しないことになる。これは物体検知を対象とした攻撃の一例だが、その他AIを含めAIに対する攻撃にはどのようなものがあるのか、攻撃をどのように回避するのかということは大きなテーマとなっている。

生成AIにおける脅威

では、多くの分野で利用が進む生成AI、特に大規模言語モデル（LLM：Large Language Model）を想定するとどのような脅威があるのか。

LLMでは、モデルが出力を生成するための指示や条件

をテキストデータで入力するが、この入力は「プロンプト」と呼ばれている。LLMに対する攻撃として、このプロンプトを細工することによる攻撃テクニックが存在する。代表的なものとして、モデルから予期しない、または不適切な情報を取得したり、LLMに設定されている指令を盗み出したりするプロンプトインジェクションがある。

このようなLLM固有の攻撃手法は、新たな攻撃手法がインターネット上で紹介されても、LLM開発者側で対策が行われ、同じ攻撃そのものは、幸い、いまのところすぐ通用しなくなっている。ただし、対策が行われる一方で、若干のプロンプトの修正で攻撃が成功する事例が報告されることもあり、現時点で完全な対策は難しいとの見立てもある。

他にも、利用者に見せてはいけない機微な情報（個人情報、認証情報等）の漏洩や、LLMの出力が精査されないことによるバックエンドの侵害、LLMを活用するシステムに過剰な権限や機能を付与することによる意図しないアクションなどLLMを活用するシステムとしての脆弱性も脅威となる。

生成AI技術の目覚ましい進歩は、多くの分野において、それらを活用する利用者に創造的なコンテンツを効率的に生成できるようにした。また、生成AIを活用して業務効率化に取り組む企業、サービスの開発を進める企業も増えている。このことは企業、広く言えば社会が生成AI特有の脅威に直面する可能性があるとも言える。

AIシステムはリスクを起点に評価する

AIを利活用する企業は、AIを活用するシステムにおける脅威を把握して個々のリスクを評価し、適切な対策を実装する必要がある。AIはAI単体でサービスとして機能することはなく、周辺機能と連携することで具体的な提供サービスを構成する。そのため、利用・開発するAIシステムによって存在する脅威やリスク、求められる対策が異なるので、画一的な対策で完結させることは難しい。AIシステムのアーキテクチャや利用形態等を踏まえ、具体的にどういう脅威が存在するか、またそのリスクはどの程度かを評価するリスクベースでのアプローチが必須となる。

評価後は、システムに存在するAI特有の脅威を含め、対策を実装していくことになる。AI特有の脅威は、Webシステムを守る従来のセキュリティ対策であるWAF（Webアプリケーションへの不正なWeb攻撃を防ぐために開発された専用防御ツール）やIDS/IPS（ネットワークにおいて不正侵入を検知・防御するシステム）等では防ぐことが難しい。また、プロンプトインジェクションのような攻撃に対しては様々な対策が考案されているが、現時点では完全に防ぐことが困難な状況にある。AIにおける脅威への対策は、セキュリティ対策の鉄則である多層で用意して、リスクを低減していくことが重要である。

AIにはAI特有の脅威や攻撃手法が存在する。AIの利活用において信頼性を確保するには、AIとセキュリティの両方を理解し、アップデートに追随し続ける必要があり、今後企業が取り組むべき課題となるだろう。

Writer's Profile

西田 助宏 Sukehiro Nishita
NRIセキュアテクノロジーズ　研究開発センター
サービス開発推進部長
専門はサイバーセキュリティ関連の事業企画
focus@nri.co.jp