資料２－１

第1回の御議論について

令和7年10月

AIセキュリティ分科会事務局

第1回で頂いた主な御意見と方向性

ガイドラインで取り扱うAIモデル等について

- まずは範囲を絞って議論を進める観点から、案のとおり、LLMと画像識別AIを取り扱うこととしてよいのではないか。
- 画像識別AIについては、種別まで記載した方がわかりやすいのではないか。攻撃の対象が、LLMなのか、画像識別AIであれば具体的にどのモデルなのか明示できないか。
- ソフトウェアとLLMを連携させるMCPが普及し始めているが、認証・認可の仕組みが不十分といったリスクは明記すべきではないか。

検討の方向性案

- ガイドラインで取り扱うAIモデルのスコープは、案のとおり、主にLLMと画像識別AIを取り扱う。
  ※　マルチモーダルなAIのセキュリティ対策も見据え、検討会では、画像を扱う生成AIの脅威についての議論も妨げない。
- 画像識別AIについては、本ガイドラインにおいて、可能な範囲で、種別まで記載することを検討。
- MCPやAIエージェントなどの急速に発展している技術についても、現時点で判明している脅威をガイドラインのAppendix等で参考として示すことを検討。

脅威について

- AISIが示した脅威の全体像は、良くまとまっていると思うが、モデルマージ等、最近明らかになっている脅威をどこまで取り扱うのか検討が必要ではないか。
- 脅威については、攻撃者の目的と手段が分かりやすくなるように記載をした方がよいのではないか。

検討の方向性案

- 最近の研究で明らかになっている脅威については、ガイドラインのAppendix等で参考として示すことを検討。
- 脅威については、読者に攻撃者の目的や手段がわかりやすいよう、図解を交えて記載することを検討。

第1回で頂いた主な御意見と検討の方向性案

対策について

- 開発者向けと提供者向けで分類している案はわかりやすい。
- 誰がどのフェーズで実装する対策であるかを明示した方がよいのではないか。事業者が自分のAIモデルが対象になるのかわかりやすいようガイドラインが対象とし得るAIモデルの事例があるとよいのではないか。
- 対策の優先度を示した方がよいのではないか。プロンプトインジェクションは対応の優先度が非常に高い脅威ではないか。
- ガイドラインに記載する対策について、何らかの基準を設定することはできれば参考になるのではないか。

検討の方向性案

- 対策については、AIのライフサイクルの各フェーズにおいて、どの主体がとり得る対策であるかが分かるように明示する（開発者と提供者向けに分けた上で、どの対策がどのフェーズで実装されるかを明示）。事例の記載についても検討する。
- 優先度の考え方について、今後、検討。プロンプトインジェクションについては、重点的に議論をすすめる。
- AIのセキュリティを検証するための基準については、長期間の検討を要するものであることを踏まえ、今後、扱いを検討。
- 本ガイドラインのスコープは、AIシステムへの脅威に対する技術的対策が中心となるため、別観点のリスクと切り分けを明確化する形での記載を検討。

プロンプトインジェクションについて

- プロンプトインジェクションは対応の優先度が非常に高い脅威ではないか。（再掲）
- プロンプトインジェクションについては、システムに侵入されてしまうという技術的なリスクと、それがもたらすレピュテーションリスクは、担当部署も異なる場合があり、切り分けて整理すべきではないか。
- プロンプトインジェクションなどの対策について、ツールを使用してのテストや、AIファイアウォールの使い分けなど、具体的実務の参考になる事項をコラム等で示せないか。

検討の方向性案

- プロンプトインジェクションについては、重点的に議論をすすめる。（再掲）
- プロンプトインジェクションがもたらすリスクについては、技術的なリスクを中心に、関連するリスクにも留意しつつ、整理する。

  ※　本ガイドラインのスコープは、AIシステムへの脅威に対する技術的対策が中心となるため、プロンプトインジェクションがもたらすリスクのうち、システムに侵入されてしまうという技術的なリスクを中心に扱うことになると考えられる。
- ツールの使用等の扱い方については、今後、検討。

その他

- 本検討会で扱う脅威と対策については、国際的な事例と整合性を取ってほしい。

検討の方向性案

- 国際動向・事例を調査の上、大きな齟齬がないようにする。