AI セキュリティ分科会（第 2 回）議事要旨

１．日　時）令和 7 年 10 月 9 日（木）15：00～17：00

２．場　所）WEB 開催

３．出席者）

【構成員】

森主査、秋山構成員、新井構成員、石川構成員、篠田構成員、披田野構成員、福田構成員、北條構成員、綿岡構成員

【サイバーセキュリティタスクフォース構成員】

岡村構成員

【オブザーバー】

内閣官房国家サイバー統括室、内閣府、デジタル庁、文部科学省、経済産業省、AI セーフティインスティテュート（AISI）

【総務省】

三田サイバーセキュリティ統括官、水間サイバーセキュリティ統括官室参事官（総括担当）、道方サイバーセキュリティ統括官室参事官（政策担当）、赤阪大臣官房サイバーセキュリティ・情報化審議官、神谷サイバーセキュリティ統括官室企画官、中村サイバーセキュリティ統括官室参事官補佐、藤本国際戦略局国際戦略課 AI 政策推進室課長補佐

【発表者（敬称略）】

斎藤邦子（NTT サービスイノベーション総合研究所）

近森啓太郎（デジタル庁省庁業務サービスグループ参事官補佐）

４．配布資料）

資料２－１　第 1 回の御議論について

資料２－２　「行政の進化と革新のための生成 AI の調達･利活用に係るガイドライン」について（デジタル庁）

資料２－３　プロンプトインジェクションの事例（三井物産セキュアディレクション株式会社）

資料２－４　プロンプトインジェクションに対する取組（NTT）

資料２－５　プロンプトインジェクションの対策（三井物産セキュアディレクション

株式会社）

参考資料１　行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン

参考資料２　プロンプトリーキング攻撃への対策の実機検証結果（構成員限り）

5．議事概要）

（１）開会

（２）議題

◆議題（１）第１回の御議論について、事務局から資料２－１を説明。

◆構成員の意見・コメント

特段の意見なし

◆議題「（２）「行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン」について、デジタル庁省庁業務サービスグループ近森参事官補佐から資料２－２を説明。

◆構成員の意見・コメント

福田構成員）

AI 事業者の立場から質問がある。セキュリティの観点から逸れるかもしれないが、著作権を無視した生成を行うような事業者もいると認識している。こういった第三者の著作権を侵害するような出力を生成する事業者は、将来的に政府調達から除外もありえるか。

近森参事官補佐）

著作権がクリティカルな問題である旨、認識している。知的財産関係については各種ガイドラインの取り込み方の検討が進んでいないが、対策を組み入れていきたいと考えている。LLM が画像や音声の生成を出来るため、生成出来る範囲の広がりを捉えてガイドラインとして対策していきたいと考えている。個々の事業者に対してどのような対応をするかについて、ガイドラインはあくまでも基準を示す方針のため、それ以上は別の検討が必要になると認識している。

綿岡構成員）

AI セキュリティ分科会のガイドラインとデジタル庁のガイドラインの関係で混乱している。デジタル庁のガイドラインを共有いただいた狙いやどう関係性を持つべきかについて説明いただきたい。

近森参事官補佐）

デジタル庁のガイドラインは政府の情報システムが対象である。本ガイドラインの対象読者は、政府情報システムを調達する方向けである。調達する方々は、AI セキュリティに関する細かい知見がないため、守ってもらいたい事項をデジタル庁で取りまとめて整備しているところである。そのため、総務省の分科会において、AI セキュリティリスクでプロンプトインジェクションが顕在化しているのを示せるのであれば、その内容を踏まえて守るべき事項として我々のガイドラインで示す必要があると考えている。ガイドライン検討時には、AI セキュリティの細かい議論ができていないところがあったので、この分科会に非常に期待している。

綿岡構成員）

具体的な関わり方として、分科会に継続して出席し、分科会構成員からの意見をガイドラインにフィードバックする機会を設ける等を予定しているか。

近森参事官補佐）

デジタル庁は本分科会にオブザーバーとして参加している。分科会の成果物を発表頂く形ではなく、分科会での議論の内容をデジタル庁で取り纏めて政府内で守るべき文書として整備しようと考えている。

◆議題（３）プロンプトインジェクションの事例について、三井物産セキュアディレクション株式会社から資料２－３を説明。

◆構成員の意見・コメント

秋山構成員）

攻撃の体系化であまり知られていないサイドチャネルのような攻撃やサービス拒否攻撃が言及されていない。事業者にとってサービス拒否攻撃はサービスが停止するため対策を検討することは必須と考えている。サービス拒否攻撃であればレートリミット等の対策があると考える。私が LLM のプロンプトインジェクションを体系化し、ドキュメント化したものがあるので参考にしていただきたい。

https://www.rd.ntt/sil/project/LLMInjectionTaxonomy/document.html

三井物産セキュアディレクション 衛藤氏）

サービス拒否攻撃について含める必要がある認識のため、いただいた情報を確認したい。第 1 回の発表資料でお示しした脅威一覧ではサービス拒否攻撃について言及しているが、不足があれば情報を更新する。

綿岡構成員）
　資料２－３の 6 ページ目の記述内容について確認したい。間接プロンプトインジェクションはユーザにメールを送信してそのメールボックスを検索することで発生するのであって、Copilot にメールを送信するという記述は誤解を招くと考える。

三井物産セキュアディレクション 高江洲氏）
　ご指摘の通り、資料の記載が誤解を招く内容であるため、誤解を生まないよう留意したい。

◆議題（４）AI 開発者からのヒアリングについて、NTT 斎藤氏から資料２－４を説明。

◆構成員の意見・コメント
森主査）
　安全を評価する基準は NTT の独自のものを使用して評価しているのか、それとも参考にしているものがあるのか。差支えのない範囲で教えていただきたい。

NTT 斎藤氏）
　AnswerCarefully が一番活用しているベンチマークである。ベンチマークを独自に作成するにはまだ難しい点がある。そのため幅広い観点で統一的なものが用意されると活用できるのでありがたい。

森主査）
　AnswerCarefully は、プロンプトインジェクションを含めた様々な評価を実施するのか。

NTT 斎藤氏）
　プロンプトインジェクションを網羅的に把握できているわけではなく、専門家の見解をいただければと思うが、ある程度含まれている認識である。

◆議題（５）プロンプトインジェクション対策について、三井物産セキュアディレクション株式会社から資料２－５を説明。

◆構成員の意見・コメント
綿岡構成員）
　システムプロンプトの例でランダムな文字列タグを使用する点について、攻撃者に類推されにくくなるとのことだが、<|start|>等の特殊トークンを使って役割を記述し、ユ

ーザーからの入力ではその特殊トークンを利用不可にする方法が一般的ではないか。設定などで特殊トークンを書き換え不可と出来る認識であるため、タグが類推された場合でも問題ない認識である。ランダムな文字列を使用することは文献などで明記されているものであるか。

三井物産セキュアディレクション 高江洲氏）

ご指摘の通り、特殊トークン等を使った記述が一般的であると認識している。今回、特殊トークンやタグが悪用されない観点、および可読性の観点から、一例としてランダムな文字列を使った例を提示している。頂いたコメントをもとに理解しやすい表現を検討したい。

秋山構成員）

システムプロンプトの強化について、ランダムな文字列タグの効果について少し疑問がある。ランダムであることがセキュリティ的に優位であるか否かについて検証した方が良いのではないか。タグがランダムである必要があるのであれば、ランダムなタグを推測しようとする攻撃も想定される。毎回ランダムな文字れるを生成するアルゴリズムをどうすべきかという話にもなってくる。また、ランダムなタグにすべきか否かの是非についても検証した方が良いと感じる。

プロンプトの強化の例で、独自のフォーマットでタグや指示を記載しているが、OpenAI等はシステムプロンプトのフォーマットを公開しており、基本的には推奨されている形式に沿うべきと考える。対策が推奨された形式と競合しないか検討が必要と考えるが、どのように考えているか教えていただきたい。

OpenAI gpt-oss：https://cookbook.openai.com/articles/openai-harmony
Anthropic：https://docs.claude.com/en/docs/build-with-claude/prompt-engineering/use-xml-tags

三井物産セキュアディレクション 高江洲氏）

ランダムな文字列タグについては可読性を高める目的等で記載している。ただし、ご指摘の通り、ランダムな文字列タグによってどれほど対策の効果があるかについては検証の余地が十分にあると考えている。

また、OpenAI 等が公開するシステムプロンプトを強化するドキュメントがあるが、それらの効果や混在した場合など、対策の効果については検証が必要であると考える。

披田野構成員）

1 点目、機密情報の除外について、場合によっては「加工」といった対策もあるが、あ

えて外しているのか、追加する可能性はあるか。
2 点目、LLM の頑健性向上という表現について、「頑健性」は安全性以外の意味で使われることが多いと思う。読者が分かりづらいと思うので、「頑健性」という表現を変えた方が良いと思う。
3 点目、対策としてベンチマークの評価が含まれていないので追加されると良いと考える。

三井物産セキュアディレクション 高江洲氏）
1 点目、機密情報の除外について、加工の対策もあることは認識しているため、留意したい。
2 点目、「頑健性」については、セキュリティやセーフティなどの観点で幅広く使われるため、ミスリードが発生しないように留意したい。
3 点目、ご指摘の通り、AI Red teaming 等を通して評価する等の対応もある。

石川構成員）
資料２－５の 3 ページ目の冒頭に記載されている継続的見直しについて、どこを見直すかや頻度を明記出来ないか。また、継続的な見直しをする際の検証条件、例えば、定期的に同じリクエストを投げてレスポンスを確認していくなど、実施条件について触れてほしい。

三井物産セキュアディレクション 高江洲氏）
アプリケーションの特性、種類等によって変わるため、継続的見直しについて一概に明記することは技術的な観点から難しいと考えている。
例えば、外部公開しているシステムであればガードレール部分やブロック部分が最優先になると考える。社内向けのシステムで RAG を基に LLM が回答する仕組みであれば RAG に投入されるデータの精査等となる。アプリの公開範囲や仕組みによって変わると考える。

森主査）
対策は実際に何かが起きる前に様々に検討するものでもあり、実際に何が起きるか、どの程度の頻度として起きうるか、それによって引き起こされるダメージがどの程度かといったことが分からないことが多いと思う。分からないなりにリスクアセスメントを実施するに当たり、示していただいた各対策の重み、優先度についての考え方はあるか。

三井物産セキュアディレクション 高江洲氏）
現在までの調査を通して、一概に優先度を決めることは困難ではないかと考えている。

アプリケーションの外部公開の状況等に合わせて優先度を高くする等が一般的に考えられるが、アプリケーションの公開範囲や仕組みによって変わる認識である。

新井構成員）
　対策の詳細が表形式でまとめられているが、網羅的に全て実施する必要があるのか、優先度が高い箇所のみで良いのかを示すことは可能か。また、運用コストや難易度等対策の評価軸を示すことは可能か。こうしたことができれば、ガイドラインの読者にとって分かりやすいものになるのではないか。

三井物産セキュアディレクション 高江洲氏）
　対策の難易度や攻撃成功時の影響範囲などをもとに対策に濃淡をつけると言ったことはできると考えるが、想定する AI アプリケーションごとの特性によって優先度が変わる認識である。

北條構成員）
　LLM が応答拒否する対策は、攻撃者からプロンプトが拒否されていることが分かってしまうことが問題とならないか。SQL インジェクション攻撃では応答の挙動によって攻撃可否がわかるケースがあるが、LLM においては問題とならないのか。

三井物産セキュアディレクション 高江洲氏）
　探索フェースにおいて、どういったプロンプトを拒否する傾向があるのか、拒否しない傾向があるのか攻撃の成立可否が分かってしまう可能性はあると考える。ただし、現時点ではプロンプトを拒否する傾向等を示すような研究論文等はない認識である。

北條構成員）
　対策の中でブロックリストやガードレールによる対策例があるが、それらが標準的な対策になっていくのか。

三井物産セキュアディレクション 高江洲氏）
　市場に出回っている AI Firewall の中身を見たところ、ルールベースの手法と LLM を用いた LLM-as-a-Judge を併用するような製品が多く、これらについては標準になっていくと考える。それ以外にブロック手法も出てくる可能性もあると考える。

披田野構成員）
　ガードレールと Firewall の種類では、ブロック、検出以外に整形があると理解している。

三井物産セキュアディレクション 高江洲氏）

ご認識の通りである。そのような防御手法を備えたソリューションの事例等について調査したい。

◆その他

事務局から、次回の日程について説明があった。10 月 17 日(金) 開催で、『AI 開発者による対策』を予定している。

（３）閉会

以上