AI セキュリティ分科会（第 3 回）議事要旨

１．日　時）令和 7 年 10 月 17 日（金）14：00～16：00

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、高橋構成員、披田野構成員、福田構成員、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】

国家サイバー統括室、デジタル庁、経済産業省、文部科学省、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、赤阪大臣官房サイバーセキュリティ・情報化審議官、水間サイバーセキュリティ統括官室参事官(総括担当)、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）】

齋藤真一、平手勇宇（楽天グループ株式会社）

高江洲勲（三井物産セキュアディレクション株式会社）

４．配布資料）

資料 3－1 楽天グループにおける AI Safety 対策について（楽天グループ株式会社）

資料 3－2 LLM 開発段階における安全性・セキュリティ対策（SB Intuitions 株式会社）

資料 3－3 AI 開発者の想定する脅威・対策（三井物産セキュアディレクション株式会社）

5．議事概要）

（１）開会

（２）議題

◆議題（１）AI 開発者における対策について、楽天グループ齋藤氏から、資料３－１を説明。

◆構成員の意見・コメント

森主査）

ここでは、ただいまの発表につき、事実関係の確認について質疑時間を設けたい。

なお、本分科会のスコープは「AI のセキュリティの確保」すわなち「不正操作による機密情報の漏えい、AI システムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策となっており、質問に当たっては、その旨を念頭に置いていただきたい。

綿岡構成員）

ガードレール等の機能は、サービス毎にポリシーや機能を変えているのか、それとも統一されたものを使用しているのか。

楽天グループ 齋藤氏）

複数の基盤ツールがあり選択し、組み合わせて利用することが可能である。サービスが多岐にわたるため、ガードレールのポリシーはサービス内容に応じて調整が可能である。また、顧客の質問は広範にわたるが、サービスによっては、入力されたプロンプトに許可されないキーワードが含まれるケースにおいても、サービスを提供する上で提供な必要なケースもある。例えば、トラベルサービスの予約の際、地名に許可されないキーワードが含まれるケースもあり、予約する際には正しい動作としてキーワードを許可する等、サービス毎に判断する必要があると考えている。

森主査）

差し支えない範囲で、ガードレールの実装に当たって、例えば、そういった機能をチューニングによってモデルの中に組み込んでいく方法や、入出力に対して直接的に検査をする方法などがあると思うが、どのような方法をとられているか。

楽天グループ 齋藤氏）

複数のレイヤーで対応している。モデルをファインチューニングする段階で答えないように鍛えるということもあれば、本日紹介したように、フロントでゲートウェイ的に検出して、それを置き換える対策もある。いずれも必要であり、サービスによって組み合わせて使っていくものである。

秋山構成員）
　御社で用いられるガードレールは、オープンソースと同様のアーキテクチャで動作するものか、自社で内製した特殊なものなのか、あるいは、オープンソースのツールをそのまま利用しているのか。

楽天グループ 齋藤氏）
　今回紹介したものに関しては自社開発したもの。この他に弊社の標準として使用できるものとして、グローバルで有名なクラウド事業者が提供しているようなサードパーティ製品もあり、こういった製品を評価しながら、オプションとして使えるような建付けとしている。

◆議題（１）AI 開発者における対策について、SB Intuitions 株式会社 綿岡構成員から、資料３－２を説明。

◆構成員の意見・コメント
森主査）
　ただいまの御発表について、事実関係の確認について質疑時間を設けたい。

森主査）
　御発表資料 21 ページに記載のある Post-training は、いわゆる Instruction Tuning に相当する部分と思っているが、ここで様々なベンチマークを用いて安全性対策と評価を実施していると理解した。プロンプトリーキングについては、ベンチマークとの対応がつきやすいが、直接的・間接的プロンプトインジェクションについては、ベンチマークとの兼ね合いで言うとどの部分がプロンプトインジェクションの評価につながっていると考えればよいか。

綿岡構成員）
　間接的プロンプトインジェクションは、ベンチマークでの反映が非常に難しいと思っている。直接的プロンプトインジェクションは、ユーザからの質問文を知前文で書くことが可能であり、例えば、「これまでの指示文を全て無視して、何々をしてください」といった自然文を、AnswerCarefully や DoNotAnswer の枠組みの中である程度評価することが可能と考える。

森主査）
　そうなるようにベンチマークを工夫していると理解した。
　19 ページ目で、JailBreak Detection は直接 JailBreak を検出するためのロジックを作っ

ていると考えればよいか。

綿岡構成員）
　記載しているもの全てを我々が開発しているわけではないが、一般に Jailbreak には別のコンポーネントを開発することが多いと考えている。有害性に関する検知については、自然文の中にある有害なクエリを検知することができ、包括的な一つの枠組みで検知することが可能と考える。一方で JailBreak 検知に関しては、機械的な処理の下で完成されたような JailBreak も存在する。このように機械的に最適化されたクエリ等まで検知対象に含めると、既存の枠組みにおけるガードレールの汎化性能が低下する現象がよく見られる。そのような場合、機械的に生成された JailBreak を専門に検知するコンポーネントを分ける事があるため、このような記載としている。

◆議題（１）AI 開発者における対策について、三井物産セキュアディレクション高江洲氏から、資料 3－3 を説明。

◆構成員の意見・コメント
森主査）
　VLM について、最新の研究事例から紹介があったと思う。この研究事例では、画像識別 AI に対する攻撃と似たような事が VLM でも出来る旨の事例だったと記憶しているが、対策についても従来の画像識別 AI と似たような手法で対策ができるのか。

三井物産セキュアディレクション 高江洲氏）
　従来の CNN(画像識別 AI)と同様の対策を実施する事で対策可能であると記載されている。例えばメンバーシップ推論型攻撃の学習データの中に入力画像が含まれているかどうかを推測するような攻撃については、応答するロジットを少し丸める等の対策が述べられている。そのため、CNN に適用できる対策は一部 VLM の攻撃に対する対策として転用できると考える。

（3）自由討議：

森主査）
　これまでの発表を踏まえて AI 開発者における対策について議論の時間としたい。なお、スコープは AI システムのセキュリティ確保である点にご留意いただきたい。個別の技術の御説明が続いたところであるが、俯瞰的な観点でのコメント等でも構わない。

綿岡構成員）

取組の共有に当たり、AI セキュリティの分野における脅威にどの程度明示的に対応出来るかについて考え直す良い機会となった。これから作成するガイドライン等に含めるかはまた別であるが、脅威には、AI の他の安全性に関するものや、広く信頼性に関わるものなど、様々な分野がある中で、AI セキュリティに関しては、実応用上でどの程度対策が取られているかという点と、どの程度脅威が存在するかという点に関して、ギャップが激しい分野だと感じた。例えば、データポイズニング攻撃に関して事業者が完璧に対策を取れているかは疑問であり、モデル反転攻撃やメンバーシップ推論攻撃についても実応用上しっかりと対策が取られているかというとかなりのギャップがあると感じる。

森主査）
脅威と、実際の対応との間にギャップがあるとのご指摘だったと思う。それぞれの脅威に対してしっかりと対策ができればもちろん言うことはないが、第２回分科会の議論でもあったように、優先度等を踏まえた上で、それぞれの脅威に対してどのように対策をしていくか、本分科会においても議論が必要ではないかと、個人的に感じたところである。

綿岡構成員）
性的な表現の誘発や犯罪の幇助等の有害性に比べて、攻撃者の悪意や攻撃を高いレベルに仮定しないと達成できない脅威がいくつかあると理解している。開発者側としては、その辺の優先順位をしっかり議論していない場合でも優先度を暗示的に持っているため、後回しになりがちなものが多いと感じる。

森主査）
非常に本質的な指摘と思う。安全性の確保という意味では、考え得るヘイトスピーチ等については既に検討されていると考える。セキュリティの場合は、もともとの訓練データを細工する等、攻撃者としてもケイパビリティが高くないと成功しないものも対象にしている。そのため、その辺りの関係や順位付け等を今後しっかり議論していく必要があると理解した。

福田構成員）
綿岡構成員が触れられた内容に同意する。インシデントデータベースやリスクリポジトリ等を見ているが、具体的なリスクに関して、今日発表されたものに関しても、想定できるもの、想像できるものと、できないものがある。例えば攻撃後に何が発生し、どれぐらいのリスクになるのか分かりにくい部分もあると思う。そのため、対応の優先度を決める上でも、事例のようなものを示せると良いと考える。事例は、想像出来るものでも良いかもしれないし、具体的なグローバルの事例を併記する形だと、事業者側も対応する優先度を決めやすくなると考える。

森主査）
　ご指摘のとおりと思う。脅威の優先順位をつけるにあたり、実際のシナリオのようなものがあった方が分かりやすいと思う。

中村補佐）
　脅威に関し、今後の取りまとめに向けて、具体的なユースケースなど、各アプリケーションにおいてどういったことが代表的に想定し得るかといった点を分かりやすく示すことを検討したい。

森主査）
　事務局から説明があったように、脅威のシナリオや、どのような被害が発生するかを整理した上で考える必要がある。セキュリティバイデザインは、脅威になる前に対策を実施するという考え方だが、対策の実施によってどれくらい脅威を抑えられるかが非常に重要だと理解した。今後のガイドライン整理にあたり、その点を留意したい。

披田野構成員）
　第１回目の分科会で、森主査からモデル反転攻撃やメンバーシップ推論攻撃は画像識別AIだけでなく、LLMも対象となるのではといったコメントがあったと思う。今回説明いただいた対策は LLM に対するモデル反転攻撃やメンバーシップ推論攻撃も対象としているか。

三井物産セキュアディレクション 高江洲氏）
　画像識別 AI に対するモデル反転攻撃やメンバーシップ推論攻撃が、全く同じ原理でLLMへ転用可能というものは、調査範囲では確認出来ていない。ＡＩの学習データを盗むという目的においては、プロンプトインジェクションの学習データ窃取型と同様と考える。

石川構成員）
　ユースケースについて補足コメントをする。ユースケースを割り当てるにあたり、どのようなデータの流れか、どのような処理が行われるかを可視化する脅威モデリングの手法がある。脅威モデリングに一般的なユースケースをどう当てはめていくかという点も併せてガイダンスに記載すると利用者にとって便利だと考える。

中村補佐）
　ユースケースを示す際は、具体的にどのようにデータを処理しているか、データがどの

ように流れるかという点も併せて示していくものと考えている。

森主査）
　整理する中で脅威のシナリオや、よくある脅威の分析フレームワークとどうマッチングするかという話も出てくるかもしれない。うまく整理出来れば良いと思う。

秋山構成員）
　一般的に AI の開発現場では様々な対策がとられていると思うが、その達成度合いをどのように図っているのかを開発に携わっている方がいらっしゃれば伺いたい。個人的には何かしらのベンチマークのようなものが使われていると思うが、それ以外に何かあるか知りたい。その上で、ガイドラインとしては、対策の必要性に加えて"達成度"の尺度を盛り込むべきかという点も、この場で議論したい。

森主査）
　開発に携わっている方への御質問であるので、福田構成員、綿岡構成員のほか、本日出席いただいている楽天グループ株式会社から御意見等があればお願いしたい。

綿岡構成員）
　おっしゃる通り、ベンチマークで測る形がメインである。具体的には、ベンチマークデータセットにはスタティックな入力サンプルが例えば千件、1 万件等用意されており、それを毎回全て入力し、有害な発言や対策の対象リスクからどれくらい誘発されるかの確率を見て達成されたかを定量的に図るアプローチとなる。ベンチマーク以外ではレッドチーミング等も含まれるが、人手でポリシーと照らし合わせながらリスクが発生するかどうかのチェックをするというフローも存在すると思う。後者は誘発率より実際の発生件数を見る方法が主となる。おそらくマニュアルチェックのようなものが楽天グループ社で行われていたと認識している。そのあたりを補足いただけると有難い。

福田構成員）
　綿岡構成員とほぼ同じ回答となるが、ベンチマークに従って実施するものと、レッドチーミングに関しては、弊社の場合、リリース前に社員が手動で実施したり、業務委託で実施したりしている。また、社員が実施する際は、意図的に、人手によるテストとして、問い合わせ時に変な事を言わないか等も試している。弊社で出来る限りの対策を実施しているが、どこまで実施すれば良いか分かりづらい領域と思う。ガイドラインや、他社がどこまで実施しているか等の話を聞ける良い機会を得られて感謝している。

楽天グループ株式会社 齋藤氏）

これまでの議論の通り、ユースケースによって守られるべき対象やその度合いが大きく変動する点について同意する。

その上で、一般消費者向けも含めたサービスと組み合わせて生成 AI を提供している立場から発言する。システムの構成やサービスの中で特定される脅威に応じ、対策がどれくらい適切かを初めに大きくスコーピングする事で、対策の深さや強度が変動するため、その旨を意識している。その理由としては、開発の現場では守る事も大事だが、スピード感を持ってインパクトのあるものを世に出したい、あるいは開発の生産性も強く意識するためだ。今回の議論の対象外かもしれないが、システムのスコープ等をある程度特定されたリスクの中で組むとか、あるいはその中で特定されたリスクを狭めた状態から対策を始めるという事も有効だと考える。

その上で、ガードレール等の対策について手動でのテストも実施していると話したが、守りすぎると利便性が下がる部分も出てきてしまう。そのため、この点を実際に確認するために人手によるオペレーションを実施している。この点は数が必要になるため、より多くの試行を重ねられるようなテストデータの使用やそれによって適切に回答されたもの、あるいは守れたが誤って回答されたとみなされるものを比べながら、そのバランスを取っている。

楽天グループ株式会社 平手氏）

ガードレールのモデルを自作している立場から発言する。

よくあるテストデータセットを適用して、パフォーマンスのトラッキングを実施している。不正検知全般に言えることだが、何を尺度に 100%完璧なのかを定義しづらい領域と感じる。日々新しい攻撃にどのように対応していくのかが課題と思っている。我々は、ユーザの入力や攻撃内容を内部で保持できているため、そのようなログから一つ一つ手探りで実施する形を取らざるを得ないと考えている。

秋山構成員）

ガイドラインを作る中で、①アプローチ方法や対策の提示だけで開発者もしくは事業者として十分か、②対策が達成出来ているかを図る尺度ついても言及すべきか、③ベンチマークでは対策が不足しているためレッドチーミングが必要であれば、攻撃の再現方法も盛り込む必要があるか、という観点で対策方法を議論したいと考える。

森主査）

御説明を受けて、ベンチマークや標準的なデータを使用する等、各社で様々な工夫がされており、利便性とのトレードオフを考慮しながら、チューニングされていると理解した。一方で、福田構成員から、こうした対策がガイドラインとしてあると有用であるという示唆もいただいた。

今の秋山構成員からの質問も踏まえ、改めて出席されている事業者にお伺いしたい。各社が持つノウハウが競争領域にならない範囲で共有され、ガイドラインとして纏まったものは有用だろうか。各社で工夫され今の形になっていると思うが、その中で得られた知見等は相互共有する事で役立つ事があると考える。集合知のような形でガイドラインの中に盛り込まれる事は意義深いのか、共有する事でリスクになるという考え方もあるかもしれないが、忌憚のない意見を聞きたい。

綿岡構成員）
一長一短と思う。開発者側としても、指標が存在しないため物事が進まない現状がある。その辺りがガイドラインに含まれれば、開発者も参照して議論が円滑に進むというメリットがある。他方でガイドラインに記載されているものを盲信してしまう可能性もあると考える。それぞれの状況に応じた対策、もしくはその対策の達成度の評価は変えていくべきであり、変わるべきと考える。単一のガイドラインのみ達成できていれば問題ないという状況は考えにくい。これらの事から、ガイドライン上で明確に記載できるかどうかは、私では分からない。

森主査）
リファレンスとしては有益だが、実際に使う際に各社で背景や環境が異なるため、それらが同じものになるかは不明確だ、というコメントだと理解した。

福田構成員）
綿岡構成員の意見と私もほぼ同じ意見である。
リファレンスとして有益だと考える。一方、楽天グループ社が触れたように、用途によって大分異なる部分があり、そこまでカバーできないと思う。また、ベースラインとして事例があると有益だが、それが全てという形にはならないと考える。目的に応じたベンチマーク、テスト方法のリファレンス、事例があると有益だと考える。

楽天グループ株式会社 齋藤氏）
我々もリファレンスモデルとしての有用性は意識している。一方で、特定されているリスクや脅威に対し、より専門的なベンチマークがあれば、それは特定の領域の専門家や開発者にとっても有益なものになると思う。最終的に我々は消費者向けのものをたくさん世に出していく立場のため、モデル等にかかってくる脅威については、ある程度そのリスクを評価し受容するものも評価しながら、第一の優先事項として顧客を守っていくという観点で対策をしている。

中村補佐）

事務局としては、ガイドラインに対してどのように達成度合いを図るかについて、様々な開発環境があり得る一方で、リファレンスとしてそういったものがあるのは有益であるという意見を受けたと認識している。

本分科会では、こういった対策があるという外延を決めた後、可能な場合は定性的に「こういった事をしていれば達成されている」という尺度を示す事もあり得ると考える。まずはどういった対策を示せるかを議論した上で、指標についても対策が固まった上でどのような達成度合いのベンチマークを示せるかを今後検討していくものと考える。

森主査）

本分科会では対策を決めた上で、達成度を評価するためのベンチマーク等も併せて検討していくものと理解した。

神谷企画官）

本日は開発企業にも参加いただいているので、事務局からも質問させていただければ有難い。

AI 開発者における AI セキュリティ確保のための対策は、基本的には、不適切な出力を防ぐために、開発時の安全性確保のための一般的な取り組み（例えば学習データの適切な事前処理、ファインチューニング、レッドチーミング等)に包含されているのではないかという感触を持っているが、その理解で大きな違いはないか。一般的な AI の取り組み以外に、AI セキュリティに特化して実施すべき対策や、一般的な AI セーフティの取り組みより更に深めて実施すべき領域等があれば示唆いただけると有難い。

森主査）

今回、各事業者から説明を受けた安全性確保の対策は、セキュリティに対する対策も含んでいるか、あるいはセキュリティに関しては別のものがあるのか。SB Intuitions 社からそのような指摘があったと記憶している。その点に関して事務局から意見がほしいという内容だった。

綿岡構成員）

端的に回答すると、AI セキュリティに関する脅威は一般的なチューニング等で行われる有害性対策とは異なる対策が必要な場合が多いと考える。例えばモデル反転攻撃、メンバーシップ推論攻撃等は、複数回のユーザからのクエリに対して総合的に判断し、どれほど敵対的な入力をされたのかという対策が必要になるため、過去のユーザログ等も考慮した上で、ユーザへの対処を考える必要がある。そのため、LLM のモデリングだけでは完結しない脅威が数多く存在していると考える。

楽天グループ株式会社 齋藤氏）

本日の議論は、生成 AI エージェントを介して提供されるサービスの水際対策に近いところを中心に議論していると認識している。セキュリティ全般では、サービスを提供する立場からすると、アクセス可能なデータや、そこから提供されるサービスの深さによって大きく変動すると考える。そのため、別のシステム構成も含めたレビュー等の対策が必要になると認識している。

福田構成員）

弊社でも AI 向けのセキュリティ、AI セキュリティ/AI セーフティの話と、一般的なシステムのセキュリティは完全に分けて考えている。

神谷企画官）

一般的なセキュリティとの比較というより、AI 開発者における AI セキュリティ対策とは、例えばガードレールに代表されるように、出してはいけない情報を出力する事を防ぐ対策がメインだと感じる。AI 開発者の視点で見た場合、その AI セキュリティを確保するための対策は、AI セーフティ確保のための対策とおおよそ似ている、あるいは包含されるという事か、そうではなく AI セキュリティのために別の領域としてきちんとやるべき事があるのか、そのあたりの感触をお伺いしたい。

福田構成員）

AI セーフティと AI セキュリティの境目は倫理的な話も含めて難しいと考える。提供するシステムの要件や、顧客からどういうものを求められるかによって、何をどこまで守らなければいけないかが決まる。弊社ではモデルを最初から全て作るため、ある程度学習データの部分はコントロール可能だと考える。RAG 等を合わせたシステム全体として考えた際にユーザからの入力の扱いや何を回答するかについても考慮している。AI を活用したシステムのセキュリティに関しては、要件も決められるため考え方が似てくると考える。要件定義が難しい部分もある。例えば、データポイズニングを受けて実際にどう困るのか、という点である。もちろんデータセットの管理はしているので、そういった事態は起こりにくい形にはなっているとは思うが、実際にそうした攻撃があると分かっても、それに対する優先度をこれ以上どうしたらいいかは考慮すべき点だと考えている。

◆その他

事務局から、次回の日程について説明があった。11 月 4 日(火)開催で、『AI 提供者による対策』を予定している。

（3）閉会

以上