AI セキュリティ分科会（第５回）議事要旨

１．日　時）令和７年 11 月 21 日（金）10：00～12：00

２．場　所）WEB 開催

３．出席者）

【構成員】

森主査、秋山構成員、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、高橋構成員、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、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、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)】

石川太一（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５－１　AI セキュリティ分科会取りまとめ骨子・論点整理（案）

資料５－２　AI セキュリティの評価基盤構築に向けて（NICT）

資料５－３　LLM に対する脅威への対策(三井物産セキュアディレクション株式会社）

資料５－４　画像識別 AI（CNN）に対する脅威と対策（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 セキュリティ分科会取りまとめ骨子・論点整理（案）について」、事務局 中村補佐から、資料５－１を説明。

◆構成員の意見・コメント

森主査）

これより、（１）について、ご質問や自由討議の時間を設けたい。

北條構成員）

形式的な点だが、資料５－１の 14 ページの下の表は開発者に対する対策なのか、提供者に対する対策なのか、どちらも含まれるのか。そのあたりを書いていただく事は可能か。

神谷企画官）

資料５－１の 18 ページに示しているものを簡潔に記載しているため分かりにくい部分もあるが、対策の主体をもう少し明確にできるかどうかについては、取りまとめに向けて検討させていただく。

福田構成員）

今の議論にもつながるが、AI 開発者・提供者・利用者という 3 つの枠組みで責任の分界点を定義するのは難しい側面があると思う。今後 AI エージェント等が出てきた際に、資料 5−1 の 26 ページの「今後の動向を踏まえた対応」では次の議論となっているが、AI 事業者からすると、きちんと考えていかなければならない点にはなってきている。AI エージェントと人の責任分界点についても議論が必要になると思う。AI エージェントが人間の指示を超え、想定外の動きをした際に責任の所在が問題となるため、役割の定義や機能の可視化、監査ログを残す等の準備が必要である。責任分界点はある程度柔軟に読み解けるようにしておける事が望ましいと考える。

神谷企画官）

AI がもたらした被害に関する責任の線引きは今回のスコープには含まれていないが、重要な点であると認識。ご指摘について今後参考とさせていただきたい。

森主査）

将来的な課題になると私も認識している。AI のエコシステムは非常に複雑であるため、特に第 4 章で記述されている将来的な課題、すなわち AI エージェントに関する部分については、そのエコシステムがどのように構成されるか、また責任分解点がどこになるかという点は、将来検討すべき重要な課題であると認識した。

綿岡構成員）
チャットにて送付した文献「Safety Pretraining: Toward the Next Generation of Safe AI」に関し、資料 5 － 1 の 14、15 ページ付近について意見を述べる。
AI 開発者における対策として、現状は事後学習に焦点を当てて議論されているようだが、送付した文献にある通り、事前学習における安全性対策やプロンプトインジェクションへの効果も見直され始めている。そのため、一言でも良いので追記した方が良いと考える。特にデータポイズニング攻撃に関しては、事前学習段階での対策が根本的な部分となる。主に事前学習の大規模コーパス内にデータポイズニングが仕込まれるものであるため、資料にある「データの信頼性の確認」にとどまらず、フィルタリングといった対策にまで踏み込んで記述すべきであると考える。

中村補佐）
事後学習を対策として記載しているが、事前学習における対策についても、取りまとめに向けて備考欄等への記載を検討したい。

綿岡構成員）
難易度の高さや適用可能な人が限られるため、備考等にとどめておいても良いかもしれない。

森主査）
指摘にあったように、備考等でもこれを入れておくと良いと考える。事後学習の前に、事前学習の部分でも実は対策ができるという NeurIPS の論文であるため、ぜひ盛り込むと良いと考える。

本ガイドラインは多岐にわたる内容であり、本日は全体構成や、スコープ、用語整理、優先度に関してこのような形で記載するということが示された。資料 5 － 1 の 13 ページの法的整理への言及、対策の整理や想定事例等についても示されている。想定事例においてもシステムとしてのパイプラインが示されており、具体的な事例となっていると思う。ヒアリング等も通じて、具体的な対策等がかなり分かってきたところを、ガイドライン案として、盛り込んでいる状況になっているかと思う。

◆議題（２）「関連の報告等」の「NICT における AI セキュリティ確保に向けた取組」について国立研究開発法人情報通信研究機構(NICT) 高橋構成員から、資料５－２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、質疑時間を設けたい。

森主査）

資料５－２の４ページで、Do Not Answer や Answer Carefully のように、LLM のコミュニティで作られている安全性評価モデルでは対応しきれていないプロンプトインジェクションに対して、まず評価をするというご説明だったかと思う。指標としては、評価軸が資料５－２の７ページ目にあったと思うが、ここでいう Safety というのは、8 ページのプロンプトインジェクションのようなセキュリティ的な観点のものを意味すると理解すればよいか。

高橋構成員）

概ねその通りである。一般的な「Safety」の定義にはセキュリティ以外の要素も混在するが、本取組においては Misuse(悪用)やプロンプトインジェクション等を扱っている。

森主査）

既存モデルがカバーしている Fairness や Ethics 等の領域も、ベンチマークとして整備されていくのか。

高橋構成員）

技術的な基盤として対応可能であり、拡張性はある。しかし、弊機構での活動の目的がサイバーセキュリティであるため、注力する領域はセキュリティである。興味のある方に機能を拡張していただくことで、他のフェアネス等のパラメータについても評価ができる形になっている。

森主査）

作成されたベンチマークは公開されるのか。

高橋構成員）

完全な公開になるか分からないが、例えば弊機構には CYNEX（サイバーセキュリティネクサス）という組織がある。日本のコミュニティの方々に利用いただく、あるいは

共にデータを提供していくといった形で進めていきたいと考えている。

◆議題（２）「関連の報告等」の「LLM に対する脅威への対策」について三井物産セキュアディレクション 石川氏から、資料５－３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、質疑時間を設けたい。

中村補佐）

事務局より補足する。本資料は三井物産セキュアディレクションによる調査内容である。今後、本日ご議論いただいた骨子・論点整理案に基づき、事務局で取りまとめ案を作成することになるが、この調査内容は、取りまとめ案の別紙として対策の詳細を整理するに当たって参考とする予定である。こうした位置づけのものとして、内容等について意見をいただきたい。

森主査）

先ほど、綿岡構成員より事前学習段階においてもセーフティーアライメントが可能であるという研究事例のご紹介があったが、これについても、Appendix 等への記載になるかもしれないが、対策の一つになり得るか。

中村補佐）

調査内容としてどのようなものがあったか、事前学習に対する対策について、三井物産セキュアディレクションから紹介していただきたい。

三井物産セキュアディレクション 高江洲氏）

事前学習に関わる対策として、学習データ自体の内容を精査して不正なデータを混入させないといった手法は、論文等で確認している。

森主査）

事前学習段階における実践的な対策について、先ほど綿岡構成員からご提示いただいた論文「Safety Pretraining」は、9 月に公開され 10 月に更新されたばかりの非常に新しいものであると認識。このように技術は日々進展しており、現時点でのスナップショットで新しいものを反映することしかできないと思われるため、ガイドラインにおいてどこまで網羅すべきかの線引きは困難ではあるが、現時点での最新情報を反映し

ていくことは重要であると考える。いずれにしても、事前学習におけるデータフィルタリング等の手法については、LLM を開発する際の有用な参考情報となり得るので、何らかの形で記載することが望ましいと考える。内容の詳細な精査には時間を要するかと思うが、「Safety Pretraining」のような事例が存在するのであれば、参考情報や脚注といった形で構わないので、記載を検討いただければという印象を持った。

神谷企画官）
　ガイドラインに組み込むにあたり、技術的な成熟度や現実的な導入可能性といった観点も考慮した上で、記載について、また記載する場合の表現方法なども含め、検討したい。

秋山構成員）
　「脅威への対策として、プロンプトの例まで掲載するか否か」という観点でコメントがある。この点について事務局から個別に相談を受けた際、具体的に提示しすぎてしまうと、利用者が思考停止に陥り、対策を単にコピー・アンド・ペーストして使用するような懸念があり、また、個別具体的な対策は陳腐化していく事を考慮すると、詳細を記述するのではなく、ある程度の方針や指針にとどめるべきか、という点について議論を行った。「対策を具体的に示すべきか、あるいは指針にとどめるべきか」という点について、他の構成員の皆様がどのようにお考えか、この場でご意見をお聞かせいただけると有難い。

神谷企画官）
　別紙の対策詳細について、秋山構成員のご指摘にもあった具体的なプロンプトを記載するか否かは、指摘の通り悩ましく感じている。ガイドラインとして具体例を挙げると、読者の利便に資する一方、「これさえ実施すれば免責される」、「安全である」といった誤解を招きかねないと危惧もしている。構成員の皆様の御意見を伺えれば有難い。

森主査）
　具体的に記述するか否かについては、私も事前の議論において、特定のシステムに特化した内容を記述するよりも、概念的かつ汎用的な内容にとどめておく方が、様々なケースに対応できて望ましいのではないかと申し上げた。私はそのような観点を持っているが、他の構成員の皆様におかれてはどうか。

石川構成員）
　ご指摘の通り、利用者が「これだけやれば免責される」と誤解する懸念があるという点には同意する。一方で具体例は必要ではないかと考える。新規に参入される方々にと

っては、「どのような観点で記述すればよいか」という指針や例が必要かと思われ、「これさえ実施すれば十分である」といった趣旨のものにすべきではないという点はご指摘の通りであるが、サンプルを示す方が有意義ではないかと感じた。総論については賛成である。

福田構成員）
　総論としては賛成である。ただ、実務の立場としては、やはりある程度の手がかりとなる事例がほしいところではある。また、以前にも触れたかと思うが、どのような脅威やリスクが存在するのかという点が重要であると考えており、「このような必要性があるため、このような対策を講じるべきである」といった背景を、読み手が理解できるよう記述するのが望ましいと考える。

森主査）
　ご意見をまとめると、実務上で対策を講じる際には、やはり具体的な記述があった方が有用であり、あくまで一例であるという注釈を付した上で、何らかの形で記載しておくことが有益であるということだと理解している。

綿岡構成員）
　私の意見としては、当初の総論と同様であるが、現在画面に表示されているような（注：資料５－３　の４ページのような)、ハイレベルな(抽象度の高い)指示内容の例を記載する事がより望ましいと考えている。以前の調査結果では具体的なシステムプロンプトの内容まで記載されていたかと思うが、そこまで詳細に記述すると、読者がそれのみを使用する懸念がある。また、システムプロンプトはあくまで 2025 年時点でのベストプラクティスに過ぎず、すぐに変化する可能性があるため、政府のガイドラインの指標として記載する事に懸念がある。したがって、詳細な記述は避け、現在提示されているような箇条書きにとどめるのが良いのではないか。

森主査）
　ガイドラインの大枠（トップレベル）においては、ご指摘のような抽象化された内容が望ましいと私も考えている。一方で、実務における活用を考慮すると、抽象的すぎると具体策が不明瞭となるため、具体的な例示が必要であるとのご意見もあったと理解している。そのような事例を別紙に掲載する場合には、あくまで一例に過ぎない点、また、ご指摘の通り技術は変化するため未来永劫有効なものではないという点を、注釈等で十分に周知・理解いただけるよう記述するのが適切ではないか。

神谷企画官）
　本件については、過去の会合において、LLM によってフォーマット等のルールが異なる旨のご指摘があったかと思う。具体的にどこまで記述可能かは、実際に記載しながら検討する必要もあると考えている。一度試案を作成した上で、取りまとめに向けて構成員の皆様にご議論いただきたい。

森主査）
　私の前提とするイメージが十分に説明できていなかったかもしれない。具体的に記述すると言っても、特定のシステムにのみ有効なものではなく、プログラミングにおける「擬似コード」のような形式を想定している。特定のプログラミング言語そのものではないが、参照すれば概ねのアイデアが理解できるもの。技術者であれば、実際のプロンプトがどのようになるか推測できる程度のものがあれば望ましいと考えている。具体的な案が作成されれば、構成員においてその適切性を判断できるかと思う。

◆議題（２）「関連の報告等」の「画像識別 AI（CNN）に対する脅威と対策」について
三井物産セキュアディレクション 石川氏から、資料５－４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、質疑時間を設けたい。

中村補佐）
　三井物産セキュアディレクションの発表について事務局より補足する。
　資料５－３の際に補足したものと同様の位置づけであるが、取りまとめ案の別紙として画像識別 AI の脅威と対策について取り込んでいくに当たり、この調査結果を参考とする予定である。こうした位置づけのものとして、専門家の皆様よりコメントをいただきたい。

披田野構成員）
　LLM への攻撃ではスポンジ攻撃の優先度が高かったが、画像識別において資料５－４の 10 ページ記載にある通り「その他の脅威と対策」に含まれている。この扱いの違いは何か。

中村補佐）

事務局より補足する。

LLM に対する DoS 攻撃は、基本的にプロンプトの入力によって引き起こされるため、実行が容易である。一方、画像識別 AI に対する DoS 攻撃は、調査の結果、攻撃者が開発・設計段階で決められた AI の処理メカニズムを把握する必要がある等、攻撃の成立に一定の前提条件が求められることを踏まえ、このような位置づけとしている。

披田野構成員）

そういった整理の根拠が読み手に伝わるような注釈を入れるべきと考える。

画像識別 AI の攻撃としてポイズニング攻撃があると思うが、発表に含まれていない理由について説明いただきたい。

中村補佐）

画像識別 AI へのデータポイズニング攻撃については、学習させる画像データの中に不正なデータを混入させ、その結果、モデルが意図しない挙動をするものと認識している。調査報告の中で具体的にどのような内容であったかについて、三井物産セキュアディレクションよりご説明いただきたい。

三井物産セキュアディレクション 高江洲氏）

指摘の通り、学習データを細工し、CNN ベースのモデルに事前学習させる事でバックドアを仕込む攻撃手法が存在するという調査結果が出ている。バックドアを仕掛ける方法としては、事前学習段階で汚染データを混入させる手法に加え、ファインチューニング段階で汚染データを混入させる手法もある。本調査において、これらは CNN モデルが登場した当初から存在する攻撃手法として整理している。

神谷企画官）

画像識別 AI に係る整理の考え方については、改めて整理させていただきたい。

披田野構成員）

敵対的サンプルへの対策について、今回は敵対的学習と防御的蒸留が挙げられているが、これ以外にも Feature Squeezing や Certified Robustness 等、多くの対策が提案されていたと記憶している。特に防御的蒸留は初期の頃から解析されていたが、その後、主要な対策として継続的に議論されてきたわけではないと認識している。様々な対策手法があり、それぞれに一長一短があるかと思うので、それを説明した上で提示した方が、活用する側にとって分かりやすいと考える。モデル抽出への対策についても、本資料ではシステム的な手法が重点的に挙げられているが、アルゴリズム的なアプローチや検知ベースの手法等、他にも様々な対策が提案されていると認識している。

神谷企画官）

今回調査し、まとめていただいた資料を参考としつつ、今後、取りまとめに向けて別紙の案として整理していく予定である。その際、指摘のあった手法等も含め、構成員のご意見を伺いながら、取扱いを検討させていただきたい。

森主査）

私の印象としても、本資料には特に実践しやすい対策が優先的に記載されていると認識している。特にモデル抽出については、画像認識に限らず LLM においても同様の課題が存在し、入出力を容易に取得され、模倣モデルの訓練に利用される事例も耳にしている。敵対的攻撃に関しては、ご指摘の通り Feature Squeezing などは比較的有効かつ容易に実装できるもの。敵対的学習は最も効果的であるというのが研究コミュニティの共通認識であると思う。そのような観点から、筆頭に挙げられているものと理解している。

最終的なガイドラインでは、調査結果に基づく内容が掲載されるものと理解している。

北條構成員）

後ほどの検討事項かもしれないが、本資料についても、提供者向けか開発者向けかの区分を明示、あるいは先ほどのまとめと同様、表形式等で整理する事は可能か。

福田構成員）

北條構成員のご意見と重複する部分があるが、今回の画像に関する資料は、主に AI 開発者向けの内容であると認識している。AI 提供者がどのような用途で提供するかによって状況は異なるが、基本的には、ドメイン知識を用いず、利用制限も設けない状態で単体利用されるケースは想定しにくいと考えられる。例えば、製品の分類や認識、あるいはセキュリティ目的での利用などが挙げられる。セキュリティ目的の場合、求められる強度にもよるが、知識・所持・生体情報のうち 2 要素を組み合わせる多要素認証などは一般的に行われている。例えば、AI による顔認証を導入する場合であっても、パスワードやセキュリティカードといった他の認証手段と組み合わせて運用されることが通常である。このように、実運用によってカバーされる側面も多々あると思われる。「AI を活用したシステム全体」としてのセキュリティを論じるのであれば、こうした具体的な運用事例と併せて提示できると、より望ましいと考える。

中村補佐）

事例や指摘事項を含め、今後の取りまとめにおいて検討したい。

◆その他

次回の会合は 12 月 5 日(金)開催を予定している。詳細は、後日事務局から改めて連絡する。

（３）閉会

以上