AIセキュリティ分科会（第1回）議事要旨

1．日　時）令和7年9月18日（木）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、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】

内閣府、デジタル庁、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)、道方サイバーセキュリティ統括官室参事官（政策担当)、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】

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1－1　「AI セキュリティ分科会」開催要項
資料1－2　AI セキュリティ分科会について
資料1－3　AI セキュリティに関する検討の進め方について
資料1－4　AI システムに対する脅威の事例
参考資料1　「サイバーセキュリティタスクフォース」開催要綱
参考資料2　AISI「AI システムに対する既知の攻撃と影響」
参考資料3　AI システムに対する脅威の事例の詳細（構成員限り）

5. 議事概要）
（１）開会

（２）三田サイバーセキュリティ統括官、森主査、構成員挨拶
省略

（３）議題

◆議題「（１）AI セキュリティ分科会について」事務局から資料１－１、資料１－２を説明。

◆構成員の意見・コメント
特段の意見なし

◆議題「（２）AI セキュリティに関する検討の進め方について」事務局から資料１－３、資料１－４を説明

◆構成員の意見・コメント
綿岡構成員）
対象 AI について、議論の範囲を LLM に絞り込み、一部に画像識別 AI を含める方向で進める点に賛同する。AI セキュリティはシステムの複雑化により脅威範囲が拡大するため、AI に関するすべての問題を網羅的に議論する必要があると考えるが、それらを纏めることは難しいと考える。そのため、まずは範囲を絞って議論を進めていくことが良いと考える。

森主査）
ご指摘の通り、スコープを広げ過ぎるとスコープが定まらなくなるため、明示的に定めた方が良いと考える。具体的にスコープがどこになるのかは更なる議論の余地はあるが、承知した。

綿岡構成員）
RAG がスコープに入っているか確認をしたい。

中村補佐）
RAG はスコープの範囲内である。

秋山構成員）
想定読者について、開発者と提供者は AI 事業者ガイドラインからの抜粋と理解している

が、この定義について根拠はあるか。EU の AI Act でも「Provider」の表現があり、モデルを開発・公開した段階で Provider とみなされる。そのため、本ガイドラインと EU AI Act の Provider の定義が異なっているように見受けられる。文脈により違う意味で使われる可能性があるため、用語の定義を明確にし、意味の違いをはっきりとさせるべきだと考える。

中村補佐）
本ガイドラインは、AI 事業者ガイドラインを踏まえて作成するものであり、定義については AI 事業者ガイドラインを引用している。表現について明確化が必要と認識したので、補足するなどの対応を検討したい。

秋山構成員）
AI 開発者による対策と AI 提供者による対策が分けて記載されているのは良いと考える。AI 提供者から見ると AI 開発者による対策が、AI 提供者自身で実施出来ないものがあるのは納得感があるが、一方で、AI 開発者の対策の中には AI 提供者が実施できるものもあると考えている。AI 開発者は対策しなくても良いというように解釈すれば良いのか、完全に役割を分けて、AI 開発者はこれだけをやればよい、AI 提供者はこれだけをすればよいとする考えなのかを確認したい。

中村補佐）
AI 開発者、AI 提供者がどういった対策をするべきかというより、対策を例示するもので、どの対策が実施するかについては、各事業者側で判断していただき、出来る範囲で実施していただくものだと考えている。

神谷企画官）
補足すると、本ガイドラインの記載により、AI 開発者、AI 提供者のどちらでも実施出来る対策をどちらかに寄せたり、そもそも出来ない対策の実施を求めていると捉えられることがないよう留意していきたい。

秋山構成員）
大量にある脅威に対する対策を実施するのは大変である。トリアージするための優先度のような情報があり、自身が開発した AI モデルやアプリケーションが影響する脅威と対策が特定しやすくなると使いやすいガイドラインになると考える。

中村補佐）
優先度があった方が分かりやすいという点についてご指摘のとおりと考える。
リスクや影響・対策の取りやすさを含めて優先度的なものを示せるかについて検討した

い。システム毎にどのような影響があるかの具体例があることで、わかりやすくなると考えるので、具体例を示すなどについても検討したい。

森主査）
　今の議論を更に深めると、本構成員の中にも AI 開発の立場で関わっている人、ヒアリングという形で AI 開発をしている人もいる。AI 開発のワークフローの中で見出す対策やどこまで何ができるかをこの分科会で議論出来ると考えている。

石川構成員）
　対象とする AI の範囲は記載されているもので良いと考えるが、対策の取り扱いで AI 開発者と提供者を分けるのも良いと考える。
　クラウドの責任共有モデルのように、可視化することで、曖昧な文言による解釈の難しさや Provider といった広範な用語の指す範囲を明確にできるため、より理解しやすいガイドラインになると考える。

中村補佐）
　AI 事業者ガイドラインで示されている開発から提供の流れの段階において、それぞれどういった対策がとれるかを図示することを検討したい。

石川構成員）
　対策をどの開発フェーズで実施するかに関しても明記いただきたい。また、その際の留意点なども観点として入れていただきたい。

中村補佐）
　検討させていただく。

石川構成員）
　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 ファイアウォールをどう使い分けるかについて、その考えをコラム等で示せるようであればそうしてほしい。

中村補佐）
　ツール使用などに関する考え方ついても、どういったものが示せるか含めて検討させていただく。

北條構成員）

優先度の意味は、重要性であるのか、それとも対策の容易性なのかを確認したい。法的観点では「べき」なのか「望ましい」かによって随分違う。ガイドラインとはいえ、ガイドラインに沿った対策や開発をしていなければ、被害が発生した際に責任を負うのかという指標にもなり得る。絶対やらなければならない対策があるのであれば「べき」と記載すべきであり、どれもが同じレベルであれば全部「望ましい」とすることになるのか、濃淡を付けていただければ、ガイドラインを読む側として、実施すべき対策が明確になると考える。

神谷企画官）
本検討会では、あくまで技術的対策例を示す想定をしており、法的に実施すべき規律を設けるのは難しいと考える。他方で優先度の面では、リスクや対策コスト見合いで、容易に対策でき、かつリスクを大きく軽減できるものは優先度を高くするという考え方も一例としてあるのではないかと考えている。

森主査）
リスクの箇所は、真面目にするとリスクアセスメントという形でそれぞれの攻撃、脅威がもたらす被害の大小をある程度見積もった上でスコアみたいな形で実施する方法があると思うが、分科会でのスコープとしてどう考えているか。

中村補佐）
リスクアセスメントや被害の大小を見積もったスコアリングなどの手法に関しては、限られた時間の中でどこまで検討できるかは考慮する必要がある。また、現時点でよく知られていない脅威をどこまで含めるかなど、進めるにあたっては、粒度感についても検討が必要である認識である。

森主査）
承知した。厳密なアセスメントは現時点では難しいと考えているが、プロンプトインジェクションのような現実的な脅威と、モデルインバージョンのようなまだ現実的ではない脅威など、濃淡を付けながら整理していければと考える。

新井構成員）
１点目として、濃淡に関して、サイバーセキュリティの観点から、AI と連携するシステムにおける脆弱性が喫緊の課題として指摘されている。特にプロンプトインジェクションは、従来のセキュリティ対策では防ぎきれない新たな攻撃経路として最も大きなインパクトを持つとされている。サイバーセキュリティの世界は黎明期、25 年以上前から様々なセキュリティ対策が講じれているが、AI という新たな攻撃経路が出来たことで、過去積み上げてきたセキュリティ対策が通用しなくなり、そこから一気にやられてしまう、というケー

スが考えられているものである。このため、ガイドラインでは開発者および利用者がプロンプトインジェクションの脅威を最優先で認識し、濃淡という意味では色濃く対処する必要があると考える。

２点目として、攻撃手法のカテゴリについて MBSD の資料の通りだと思っているが、例えば GitHub へのプルリクエストを解釈してコードの修正案を作るといったところは、間接的プロンプトインジェクションなのか、直接的プロンプトインジェクションなのかで意見が分かれると思う。意見が分かれるのは当然で、現時点ではベストプラクティスが無いからである。時間が経つにつれて議論が深まり、決まっていくと思う。そのため、今の時点で正解を出すのはなかなか難しいが、例示はいろいろすべきかと考える。攻撃経路などどういうものがあるかをかみ砕いて説明する方が、読み手にとって親切ではないかと考える。

３点目として、プロンプトインジェクションという脆弱性が資料１－４の４ページから５ページにかけて 8〜13 のパターンで示されていたが、更に分かれていると考える。

１つは、レピュテーションリスク、すなわち、倫理的な規制を破り不適切な回答をすることで企業としてのガバナンスリスクがこれまで注目されてきた。このリスクと、、提供者側が侵入されてしまうリスクは切り分ける必要があると考える。ガバナンスも必要であり、一方でシステム開発の中で提供者側は侵入を防ぐということも両方しなければいけないが、観点が変わると考えられる。どちらの観点なのか、または両方の観点なのかを整理する必要がある。企業は担当部署が分かれている場合があるため、この観点を整理し、どこに影響するのか例示して見える形にしていかないと、ガイドラインを読むセキュリティ担当が混乱するのではないかと危惧している。

中村補佐）

１点目について、プロンプトインジェクションなど社会的に知られている脅威については重点的に扱い、議論が必要であると認識している。本分科会においても議論を深めていければと考えている。

２点目について、具体的な攻撃経路を示すことは、対策を考えるうえでは重要と考える。そのような図についても読者に分かりやすい形でお示しできるよう検討したい。

３点目について、ガバナンスなどの周辺のリスクについても検討させていただきたい。

福田構成員）

基盤モデルの開発や提供している立場から発言する。これまでの構成員のコメントは同意するところが多い。弊社としても資料に記載されている対策を行っているが、現状は個社で対応できる範囲となる。官が主導して分科会を開く背景としては、LLM や生成 AI の脅威が社会基盤に波及する可能性があると考えていると認識している。公開できるかは分からないが、可能であれば官にて評価基準などの事例が作れると、弊社のような事業者側も大変参考になる。脅威や対策に関しては、国際的な事例についても事業者として追っているが、

それらと整合性を取った形で、優先度などについて本分科会でディスカッションできると大変ありがたい。

中村補佐）
AI のセキュリティに関する評価基準については、この検討会の中で策定可能かという点はあるが、何らかの基準を基に事業者が対策をしていくことで、AI セキュリティが確保され、その上で AI が社会の中で使われていくということは重要だと考える。本検討会のスコープというよりは、将来的なところに向けて、議論が出来ればと思う。
国際的な整合性については、これまでの調査においても NIST のフレームワークや OWASP など民間の基準との整合性を取りながら進めている。それらに留意して検討を進めたいと考える。

森主査）
ガイドラインには入らないかもしれないが、評価基準の結果みたいなのを関連ポータルで公開するような道筋はあるか。

中村補佐）
今後、ガイドラインとして策定した対策をどう周知啓発していくか検討会の中で考えていきたい。検討の上ではあるが、将来的に何らかの基準のようなものをポータルサイトなどで公開することは可能性としてはあり得ると考える。

高橋構成員）
これまでの議論内容に賛同している。一方、取り扱う脅威の範囲について議論を深めたいと考えている。今回のスライドでは AISI が作成した図が非常に良く全体を捉えているが、先日の USENIX で発表されたモデルマージを悪用した攻撃など、図に当てはまりにくい脅威もあると考える。脅威の濃淡はあるが、研究開発の分野では淡にあたる脅威も多く存在する。そういった脅威についても今後議論しながら徐々にアップデートできればと考える。

中村補佐）
AI に関する様々な脅威が研究段階のものとして出ていると認識している。この点、ガイドラインのスコープとしてどこまで扱うかは議論の余地があるが、すでに認識されている脅威については、研究段階であったとしても、何らか留意が必要であるという形でガイドラインにおいて言及する可能性はあり得ると考えており、議論できればと考えている。

森主査）
先の件はモデルのエコシステムの問題と理解している。1 社だけで全モデル開発するので

はなく、配布されたモデルをファインチューニングして使うという使い方が最近増えている。モデルマージもその１つで、1 個の訓練したモデルを別のモデルとマージすることで非常に少ない計算コストで新しいモデルを作るというのもある。モデルは 1 か所に収まるというより色々なエコシステムを通じて実は使えるということを見せることができると、この図の中にも外部からモデルが入ってくるというのがあるかもしれない。いずれにしても今後ディスカッションできればと考える。

綿岡構成員）
資料 1－4 の「G のプロンプト窃盗攻撃」と「⑬のプロンプトリーキング攻撃」の違いについて教えていただきたい。

中村補佐）
資料 1－4 の G のプロンプト窃盗攻撃は画像生成 AI において画像を生成するためのプロンプトが攻撃者によって再構築されてしまう攻撃で、プロンプトリーキング攻撃は、プロンプトインジェクションなどでシステムプロンプトが漏洩する攻撃を指している。

綿岡構成員）
①のモデル抽出攻撃と類似していると思うが、商用利用不可のライセンスで公開したモデルを無断で商用利用される脅威は本ガイドラインに含まれているか教えていただきたい。

中村補佐）
商用利用不可で公開したモデルが無断で使われてしまう脅威については、AI 側でそれを防ぐ何かしらの対策が講じられており、それが破られてしまう脅威については検討の対象と認識している。一方、モデルを無断で使われるケースについては AI システムにおける脅威としてどこまで含めるかについては慎重に考える必要がある。

綿岡構成員）
資料 1－4 の J のファインチューニング攻撃について、「学習データ漏えい」と記載があるが、より一般的な脅威につながると考えている。例えば、画像生成の分野では、Uncensored モデルという名前で、性的な画像を生成させるチューニングなどが普及していたりする。同様に、Uncensored LLM のチューニングをすることで、兵器製造、犯罪幇助など、あらゆる悪用のリスクがあるかと考えており、広く記載が必要ではないかと考える。

三井物産セキュアディレクション　石川）
おっしゃる通り Uncensored モデルの脅威は、事業者様の視点では非常に重要と認識している。他方で資料１－３の１ページに記載の通り、本ガイドラインのメインスコープは「セ

キュリティ確保」にあるため、「有害情報の出力制御」などを「セキュリティ確保」に含めるべきかについては議論が必要と考える。

森主査）
　有害情報あるいは目的外利用になるので今回はスコープの範囲外だが、議論を妨げるものではないため、必要な事であれば議論を深める必要があると考える。整理としては、1 ページ目にあるようにセキュリティ確保というところでは、もしかするとやや対象外になるかもしれない。

披田野構成員）
　脅威について、手段と目的が入り混じっている。もう少し分かりやすくまとめた方が対策を講じる上で検討しやすいと考える。プロンプトインジェクションについて、ほぼ同じ手段がそれぞれの目的について書かれており、違いが分かりづらいと考える。

中村補佐）
　目的と手段については分かりやすく記載してほしいというご要望と理解した。どういった攻撃経路があるのか、攻撃者の意図があるかを含めて図などを示して、分かりやすくお示しできるようにしたいと考える。

披田野構成員）
　ガードレールやファイアウォールを入れるという議論において、ガードレールやファイアウォールを LLM で補正するというのも出始めている。アライメントなどではなく外付けのガードレールを外す脅威は対象となるのか。

中村補佐）
　外付けガードレールを外すものについても、AI システムが意図せず挙動をしないよう取り付けられるものなので、本検討会でもそのような脅威もスコープに入り得ると考える。

森主査）
　スコープについて LLM と画像識別 AI と記載されている。構成される技術要素について、LLM は比較的一意に定まるが、画像識別 AI といった際には Transformer ベースとしたものを指すのか旧来的な深層学習を指すのかその辺りを明確にした方が良いと考える。

中村補佐）
　従来の機械学習を用いて画像を分類するような AI を想定している。従来からの蓄積があり対策も技術的に成熟していると理解している。生成 AI が画像を扱う場合の対策に従来型

の画像識別 AI の対策が一部転用可能であると研究があることを承知しており、今回のスコープとしては従来型の画像識別 AI を対象としている。将来的にマルチモーダルな AI の対策も検討し得ると考えており、今後、その扱いについて議論していただきたいと考えている。内容によっては年末の取りまとめにおいて対策例の定義まで進まず、中長期的な課題として整理することもありうると考えているが、出来る限り調査していきたいと考えている。

森主査）
攻撃の具体例を例示するところに関係すると思うが、この攻撃は LLM が対象なのか画像識別 AI であれば具体的にどのモデルなのか明示的に示せると何が対象となるか明確になると思う。最終的なガイドラインの中に具体例があれば良いと思うがいかがか。

中村補佐）
どういったモデルを対象として攻撃が実行されるか LLM、画像識別 AI は特にこの辺りは色々なモデルの下にあるアルゴリズムがあると思うので、そこの研究・脅威・対策が有るかの調査を行っているので、可能な範囲で分かりやすいように示せればと考えている。

森主査）
承知した。より具体的なもので言うと資料１－４でいくつか分類があるが、一番右の列に LLM か画像識別 AI か区別されている。これが研究分野だと他にもあることを伝えたかった。例えばモデルインバージョンなど画像識別 AI を対象と書いているが、研究分野では LLM や VLM もあり、ガイドラインの読者によって解釈が異なると良くないのではないかと懸念があり、この話をさせていただいた。それぞれの読者が見たときにそれが自分事として捉えられるようなっているかというところは意識していればと思った。それを判断するためには、具体例があれば良いと考えている。読者が自分の使っているモデルが何なのかが分かるような形にしたい。

披田野構成員）
Transformer を利用するような画像識別 AI に関しては取り扱わないのか。それとも Transformer に対して画像識別 AI に対する従来の攻撃が適用可能なのでスコープに入っているということか。

中村補佐）
今回扱おうとしている画像識別 AI は、あくまでも従来型の機械学習モデルの中で研究されてきたものに対しての脅威と対策がメインのスコープであるものと考える。画像識別 AI に対する脅威と対策が生成 AI において画像を扱う際に転用可能であるという研究があると承知している。この検討会においては、将来的にマルチモーダルな AI の対策も見検討し得

ると考えており、この観点から、画像を扱う生成 AI に関する脅威の扱いについても議論していただきたいと考えている。年末の取りまとめのスコープにどこまで含められるかは今後の検討とさせていただきたい。

森主査）

Adversarial Example のような研究は、従来的な深層学習に対してかなり枯れている。10年くらい前から研究例があるので、検討の余地はあるが対策についてはある程度固まっている印象だ。一方、新しい画像識別、自動運転では Vision Transformer を使って画像だけではなく他のデータも併せて処理している。議論の遡上にはあげるかもしれないが、ガイドラインには入れるものではないと理解した。

新井構成員）

様々なソフトウェアと 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である。サイバーセキュリティの専門家はすでにこの問題を共通認識として持っているため、ガイドラインにこのリスクを明記した方が良いと考える。今回は MCP に対して議論ができていなかったので私の方から問題提起とさせていただく。

中村補佐）

MCP に関しても実際に使われ始めており、そこに対する脅威というものもあると認識している。そのような新しい技術に対しては、対策も発展途上となる場合があり、明示的に対策を示せるかは課題があると考える。ただし、将来的にこのような技術に対しても留意が必要なため、現段階の情報として、Appendix などの形で示すのは可能と思っているので検討させていただきたい。

森主査）

指摘いただいたとおり、フューチャーワークとして明示的に書かれている事が重要であり、MCP については重要な問題であるので、何らかの形で記載出来ればと考える。

◆その他

事務局から、次回の日程について説明があった。

（４）閉会

以上