AIセキュリティ分科会（第4回）議事要旨

1．日　時）令和7年11月4日（火）16：00～18：00

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】

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)】

小西達也、湯浅潤樹、齋藤耕平（サイボウズ株式会社）
河野省二（日本マイクロソフト株式会社）
瀧澤与一（アマゾンウェブサービスジャパン合同会社）

4．配布資料）

資料4－1　サイボウズにおけるAIセキュリティ対策（サイボウズ株式会社）
資料4－2　MicrosoftのAIサービスにおけるセキュリティ対策（日本マイクロソフト株式会社）
資料 4－3　生成 AI 技術を用いた製品のセキュリティ・リスク評価・ガバナンス体制（株式会社 Preferred Networks）
資料4－4　AWSの生成AIサービスとセキュリティ対策（アマゾンウェブサービスジ

ャパン合同会社）
資料４－５　海外動向の調査報告（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 提供者からのヒアリングについて、サイボウズ 小西氏、湯浅氏から、資料４－１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）
資料 4－1 の 6 ページで説明された多層防御の 5 つの層の区分けと、14 ページのガイドラインに記載された気を付けるべきポイントとの間には、対応関係があるか。

サイボウズ 小西氏）
多層防御の各層と、気を付けるべきポイントは明確に区分けしているわけではない。ただし、ガイドラインにおいて、それぞれの多層防御の方法で対策を行うべきであることや、どのような対策があるかといった点を提示している。

森主査）
区分けはないが、対応がつくようになっているとの理解で良いか。

サイボウズ　小西氏）
然り。

◆議題（１）AI 提供者からのヒアリングについて、日本マイクロソフト河野氏から、資料４－２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。

綿岡構成員）
　コーディングアシスタントのようなアプリケーションでは、ユーザがプライベートキー等の機微情報を誤ってアップロードするリスクがあると思うが、このリスク対策として、例えば、ユーザのローカル上で危険な情報を検知しブロックやマスキングを行うようなフローが導入されているか。あるいは、これから議論されていくのか。

日本マイクロソフト 河野氏）
　「ローカル」とは、各テナント(クラウド上の顧客専用領域)も含むローカルか、それとも PC の中のどちらになるか。

綿岡構成員）
　後者の PC の中での方である。

日本マイクロソフト 河野氏）
　PC の中で検知やブロック処理を行うのであれば、アプリケーションがこちら側(ユーザの PC 側)にないと処理が出来ないため、アプリケーション設計が必要になる。設計を手助けする機能については、我々のツールの中で提供できる形になっている。

北條構成員）
　資料 4－2 の 3 ページの箇条書きの下から二つ目に Web 検索中に成人向けサイト等を Bing のウェブブロック機能で除外する旨の説明がある。これは攻撃とは若干違うような気がするが、例えば、業務上必要なサイトをブロック機能から除外することは可能か。もしくはそのような機能があるか。

日本マイクロソフト 河野氏）
　どこまで外すことが出来るかは、フィルタの作り方にもよるが、アクセス権をポリシーとして設定出来る。ただ、どれが成人向けサイトか、どれがあまり評価の良くないサイトかについてはレピュテーションデータをマイクロソフトサイト側で持っており、これを使うか使わないかによる。これらを全て外して使いたいという場合には、自身で何らかのデータ収集をしながら制御していかなければならない。一部のこのサイトだけは見ても良い、といったものであれば制御が用意かもしれないが、北條先生がおっしゃっているような細かくとなると、難しいと思う。

北條構成員）
　特定のサイトだけ許可する設定が可能であると理解したが、正しいか。

日本マイクロソフト 河野氏）

特定のサイトだけ許可する設定は可能である。

◆議題（１）AI 提供者からのヒアリングについて、Preferred Networks 福田構成員から、資料 4－ 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、事実関係の確認について質疑時間を設けたい。

石川構成員）

資料 4－3 の 18 ページに記載のある「責任の分担・共有」に関して伺いたい。ユーザ側に「何を見せたいか／見せたくないか」といった要件を決めると思うが、御社の設計思想は、ユーザがそうした要件を詳細に設定出来るよう自由度高く設計しているのか、誰でも簡単に設定出来るよう機能を狭めて設計しているのか、どちらの側面もあると思うが、確認したい。

福田構成員）

セキュリティのために機能は狭めたいというのが正直なところだが、リスクと利便性のトレードオフのような判断に関しては、現時点では、ユーザ側・AI 利用者側に持ってもらうのが良いのではないかと考えている。ただし、絶対にダメな領域、弊社としても許容できない領域はある。そうではない部分についての判断は、利用者側が行うべきと考える。

森主査）

資料 4－3 の 17 ページ右側にセキュリティ課題と対応例が示されているが、サイボウズ社は資料 4－1 の 6 ページで説明された多層防御のように体系的にまとめていたが、御社でも内部的にそういった体系的な形でセキュリティ課題と対策をまとめているか。

福田構成員）

体系的にまとまっているが、最終的には様々なリスク評価委員会であるとか、セキュリティ委員会で持っている部分はあるが、弊社の場合、プロダクト数が多いため一概にきっちりまとめきれているものではなく、プロダクトごとにセキュリティ課題と対策を考えている状況である。

◆議題（１）AI 提供者からのヒアリングについて、アマゾンウェブサービスジャパン瀧澤氏から、資料４－４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、事実関係の確認について質疑時間を設けたい。

北條構成員）

資料４－４の 11 ページに顧客データは一切記録されないというのと、16 ページに包括的なモニタリングとログ機能があると説明があったが、これらは異なるものか。

アマゾンウェブサービスジャパン 瀧澤氏）

まず、モデルプロバイダーに対してユーザからのデータ等を共有しないというところが１つ目にある。また、プロンプトそのもののデータ自体を保管しないという機能もある。ログそのものに関しては２種類あり、AWS の API に関してはもちろん AWS として記録はとるが、ユーザがどういうプロンプトを入れていたのかに関しては、データはあくまでユーザ側にある位置づけであり、ユーザ側で AWS の必要なサービスを使ってログを記録でき、記録しないということもできる。AWS では責任共有モデルという考え方があり、データ自体を顧客が管理し、所有するという考え方に基づいているので、顧客が入力したデータに関しては、取り扱いはあくまで顧客の管理という位置づけで考えている。一部細かなサービスで、データを見ないと処理出来ないもの、例えば画像認識ですと、画像そのものに意味があるため、こういうものに関しては AWS のサービス条件で、契約書でそのデータをどう取り扱うかということは書いており、その内容が不適切だと思えばそのサービスを使わずに、別のサービスを使っていただくことになると思っている。

北條構成員）

画像認識 AI では、画像データを見る必要があり、契約の中でどう取り扱うかが書いてあるということだが、生成 AI のプロンプトで入れる文章については、そういうことはないと理解すればよいか。

アマゾンウェブサービスジャパン 瀧澤氏）

基本的に文章や画像も、例えば Amazon Bedrock がサポートしているモデルに入れる際には、顧客が選択するため、Amazon の社員が中のコンテンツを見るための権限もないという位置づけになっている。一部のマネージドサービスの中において、入力されたデータそのものに意味がある場合においてのみ、取扱いに関して AWS サービス条件という契

約書の中に取り扱い方法が書かれている。今回ご説明した Amazon Bedrock に関しては、データに関して AWS は感知していない。

北條構成員）
　そうだとすると、例えば利用者によってプロンプトインジェクションのような攻撃が発生した際に、そのような利用者が入力したデータのログも見ることができないということか。

アマゾンウェブサービスジャパン 瀧澤氏）
　利用者が、Amazon Bedrock Guardrails のプロンプト攻撃フィルターを利用している場合において、拒否した旨の内容が含まれたレスポンスが返ってくるので、そこで拒否されたと理解できる。そのエラーコードの内容に基づいて、何によって失敗したのかが分かる。

森主査）
　Amazon Bedrock のガードレールは、資料 4－4 の 14 ページに書かれているフィルタなどの様々な処理は、顧客ごとに調整ができるのか。

アマゾンウェブサービスジャパン 瀧澤氏）
　顧客ごとに調整が可能である。

森主査）
　それぞれのポリシーに応じて、フィルタの強さ等は調整できるということか。

アマゾンウェブサービスジャパン 瀧澤氏）
　主観的な項目について、強度を調整できる。利用規約に基づいて、犯罪に関するものなど禁止されている項目がある。犯罪に関わる行為は、AWS は認めていないので、例えば詐欺のやり方を教えてくれ、と言ったプロンプトは、AWS の利用条件に反することなので、答えない、ということになる。それ以外の調整に関しては、ユーザ側で行うことができる。

◆議題（１）AI 提供者における対策の自由討議

森主査）
　これまで 4 件の発表を踏まえて、AI 提供者における対策について議論頂きたい。

新井構成員）

サイボウズ社の資料 4－1 の 14 ページにある「レベル感」はサイボウズ社独自のものか。それとも、何かの基準を参照しているのか。

サイボウズ 小西氏）
こちらのレベル感はサイボウズの内部で決めているレベル感、つまり社内基準を参照している。外部の特定の基準があって、それを基準にしているというものではない。主には、その攻撃の観点や、それに対するエクスプロイトに対してどれぐらいのレベル感があるかという点を、PSIRT 側で設定しているものとなっている。

新井構成員）
個別の脅威について、人間が議論してレベル感として高い・低いを判定し、独自に決めているという認識で合っているか。

サイボウズ 小西氏）
その認識で合っている。

新井構成員）
本分科会でガイドラインを出していくという点で、こういったリスクのレベル感は非常に重要だと認識していたため、改めて質問させていただいた。本分科会の最終的なアウトプットにレベル感を盛り込むことで、有効なガイドラインになると思う。

綿岡構成員）
AWS 社では、AI セキュリティ対策に関して非常に先進的な取り組みをされていると認識しているが、その上で未だに抱えている課題感や対策の限界等はあるか。

アマゾンウェブサービスジャパン 瀧澤氏）
良い質問であるとともに、AI はまだ進化の過程にあるので、今後を予測することは難しく回答に悩む難しい質問。
例えば RAG のようなソリューション等、単純な質問に対して答えを返すシチュエーションは、世の中においてもリスクが認識されて、対策が取られ始めていると思っている。そういったリスクが明確になっているものに対しては、Amazon Bedrock のガードレールなどにより、カバー出来る範囲が増えていると思う。ただ、新しい脅威が認識された時にいかに早く対応するかということだと思っている。AWS の中でもボードメンバーのミーティングで、細かいレベルの討論を実施していて、セキュリティに対して最優先事業として取り扱っている。
生成 AI のトレンドの観点においては、発表の最後にも触れたが、エージェンティック

AI 等、新しい使われ方の部分に関して予測ができていない。開発 AI エージェントもそうだが、非常に短い指示をすると、色々な情報、例えばリポジトリなどにアクセスして、コードの改変等をしてアプリケーションを返していくことになるが、使うユーザ側への出力に対してある程度チェック機能がないと、出来上がったアプリケーションが、一見、要求を満たしているように見えるが実は意図しないものになるというリスクがあると思う。エージェンティック AI や AI エージェントは、自律的といえど、ある程度のトラストの実現と、重要な部分は人間がチェックできるようにすることが重要になる。この点については、色々考えなければならない部分があると思う。例えば AI エージェントも、一連の決まりきったフローがあるようなものは、何が呼ばれるか分かりやすいが、自由に色々な情報源やリソースにアクセスして改変を加えてくるようなものは、新しいリスクがあるかと思う。これに関しては、使われ方・使い方を顧客と一緒に学んでいきたいと思う。

北條構成員）
　AWS 社ではユーザの入力したデータが攻撃内容を含むものであったとしても、その入力データは取得しておらず、エラーとしての出力から攻撃内容の判断が可能とのご説明だったと思う。他の 3 社ではユーザが入力したデータが攻撃であって、これを検知した場合、事業者側はその入力データを確認することが出来るか、伺いたい。

サイボウズ 齋藤氏）
　約款等より、顧客の入出力データはどこにも記録するようにしていない。そのため、エラーの原因となるプロンプト、あるいはその出力内容を確認することはできない。どういった入力、あるいは出力かは分からないが、どういった理由でエラーとなったかは記録している。

北條構成員）
　攻撃内容を詳細には把握することが出来ないと理解した。

日本マイクロソフト 河野氏）
　マイクロソフトではユーザのやり取りを 2 つのレベルで捉えている。マイクロソフトが提供するサービス全体の中で攻撃を検出する場合と、顧客のテナントの中で記録する場合がある。マイクロソフト側ではプライバシーデータは取らないため、アクティビティに対してのフィルタリングは、システム側で行うことがあるが、どのようなリクエストをしたか、といったことについては顧客のテナント内での処理になる。テナント内の処理の中で、先ほど Microsoft Purview を紹介したが、Microsoft Graph というアクティビティが全て記録されているデータベースがあり、それらを用いて、顧客自身で何らかの不正に

ついてルールに則ってフィルタリング等を実施し、内部不正や、何か自分たちが入力してはいけないデータが入力されたのかどうか等を判断いただくイメージである。

北條構成員）
　今の説明は内部の不正なので、どちらかと言えば、マイクロソフトに対する攻撃の対策として行っているのではないということか。

日本マイクロソフト 河野氏）
　マイクロソフトのシステム全体に対する攻撃はマイクロソフトが検出するが、テナント内の攻撃は顧客側で何らかのブロックがされているのではないかと思う。無償版 Copilot Chat の場合はこの限りではない。

福田構成員）
　基本的に弊社も顧客データを見ることはしない。ただし、サービス利用規約に禁止行為が記載されており、これに該当するものがあった場合、ユーザの事前同意がある場合に限り、デバッグやトラブルシューティングにデータを利用する事が出来るようにはなっている。あくまで勝手にはできない形になっており、サイボウズ社等と同じような形であると理解している。

北條構成員）
　利用者が攻撃をしたとしても、攻撃内容は見られないということか。

福田構成員）
　その認識で合っている。

秋山構成員）
　AWS 社への質問として、資料 4－4 の 10 ページ右上に記載されている情報公開に関し、その目的は、ユーザが様々なモデルやサービスを選択できるようにするという趣旨で合っているか。

アマゾンウェブサービスジャパン 瀧澤氏）
　その通りである。安心と安全の考え方があると思いうが、AWS ではこの情報公開においては、「安心」を得ていただくための情報公開であると理解いただけばと思う。AWS AI サービスカードに関して、資料 4－4 の 10 ページ右上に項目として挙げているものは、ほぼ列挙している。アジェンダになっている項目については、それなりの文章量で AWS がどうやってプライバシー等に対して配慮しているかを、例えば Nova や Titan 等のモデ

ル毎に書いている。情報公開をしている内容に基づいて、コンプライアンスと合っていれば使う、そうで無ければ使わないという判断ができる。ただ、AWS 自体は安心を得られるような形でモデルを開発している。色々なリクエストを受けながら、情報公開をしているところである。サードパーティ(AWS が開発していない)のモデルに関しては、各モデルのプロバイダーが色々なポリシーに基づいて情報を提供しているものを、EULA という形で情報を提供している。それらは各モデルプロバイダーのリンク先と同じ内容が書いてあることが多いと思うが、こちらも、顧客の選択に資する観点で公開しているもの。

秋山構成員）
　書かれている情報の粒度は、ユーザが読んで、例えば安全性や信頼性に関して判断が出来るレベルのものであるとの理解で良いか。

アマゾンウェブサービスジャパン 瀧澤氏）
　そう考えます。例えば Amazon Nova に関しても、Nova Reel、Nova Canvas、Nova Lite、Nova Pro 等、複数あるが、それぞれに対して結構な文章量を情報開示している。ご関心があれば、「Nova と AWS AI サービスカード」で検索いただければ、文章量がそれなりにあることが分かっていただけると思う。

秋山構成員）
　他社で同様の情報公開をしている会社があるかどうか、あるとすればどういうものか、ご教示いただければと思う

サイボウズ 小西氏）
　弊社はモデルセットを作っているわけではないので、同じような情報公開レベルは難しいと考えている。AI 利用に関しての規約やサポートページ等で説明しているとは思うが、詳細な公開は行っていないという認識である。

日本マイクロソフト 河野氏）
　先ほどご紹介した「責任ある AI」のページで、全体としての情報公開を行っている。また「Microsoft Learn」で各サービスについての細かな情報を提供している。それぞれの製品ホームページもあり、その中でも提供している。進歩の早いこともあり、それぞれのページ間で表現の違い等があったり、製品名も頻繁に変わったりするので、ご不便をおかけすることはあるかもしれない。各ページから問い合わせができるので、そこにいただければ、応えることができると思う。回答可能な内容については全てホームページで公開している。

福田構成員）

弊社では、ポリシー等の大枠の部分は公開しているが、個別の部分についてはこれからである。政府からの補助を受けて作っている LLM に関しては、どういうデータで学習したかまで公開している。先ほど説明のとおり、ポリシーやガバナンス体制等は公開している。

森主査）

本分科会の終了時間が迫ってきているため、一旦事務局の方で、この先の進め方についてアナウンスがあるのでお願いする。

中村補佐）

進行上、この後に海外調査の報告を三井物産セキュアディレクション社からしていただく予定であったが、時間の都合もあるので、そちらについてご質問等あれば、今週中までに事務局へメールでお寄せいただききたい。

森主査）

本日の議論は、主に生成 AI、LLM を中心としたものであったが、本分科会では、深層学習、いわゆる機械学習モデルに関してもスコープに入れている。具体的には画像識別モデル等、また、それを利用したサービスであるが、AWS 社、マイクロソフト社では画像識別サービスを提供されていると理解しており、これに関して、セキュリティ対策としてどのようなことをされているかを簡単にご教示いただけないか。

アマゾンウェブサービスジャパン 瀧澤氏）

画像認識に関して AWS は Amazon Rekognition というサービスを提供している。顔の一致を見るような画像認識や、物体の認識、有名人やセレブリティの判断が出来るサービスである。Amazon Rekognition も、先ほどご説明した AWS AI サービスカードを提供しており、どのようなデータを取り扱うのか等を開示している。

画像認識においてリスクとして捉えられるのは、例えば似たような画像を与えて、同じ人と誤認させるものや、画像自体は似ていないが特徴点が似ているため同じものと認識させるような攻撃手法があると考える。AWS の Amazon Rekognition の認識率は、一致しているか否かをゼロイチでは返ってくるものではなく、例えば 80%や 65%という形で返ってくる。例えば、「ある%以下の場合は、攻撃の可能性が高いので、アプリケーションに対して一致するものがなかったと返すこととする」といった API の閾値を設定することが可能である。また、例えば、90%や 95%といった認識率など、悩ましいシチュエーションでは、人間に判断を委ねることとするアプリケーションを開発していただくことも可能である。例えば認識率が 99%、97%の場合は、「一致」として返すようにして、

90%ぐらいの時には、管理者に通知を行い、人間の判断で対応するようにする、という機能も用意している。

森主査）

敵対的攻撃と呼ばれているものとして、人間の目ではあまり差異がないように見えるものが、AI に与えると全く違う答えが返ってくるというものを意図的に作り出すことができるということが知られているが、これに対して何か対策はされているか。

アマゾンウェブサービスジャパン 瀧澤氏）

iBeta 等の認証等の基準に基づいて、Amazon Rekognition の評価をしてもらっており、iBeta の認証適合性テストのレベル 1、レベル 2 に合格していることで、顧客において、そういった攻撃への耐性があると判断いただくことになる。顧客側でこれでは足りないと判断される場合は、顧客側で追加的な対応や、チェック機構を入れていくということが必要になる。

日本マイクロソフト 河野氏）

マイクロソフトの場合、特に画像だから、文字だから、音声だからという事で認識を変えている事はなく、インプットする時にそれぞれのポリシーを踏まえる形になっている。画像について、AWS 社からも、「らしさ」の話があったと思うが、以前は、例えば猿か人間か、豚か人間かのラベルを付けて、1 回ラベルを付けたらその後はそれで運用していくフィルタリングのような形で学習等をしていたかと思うが、今は画像認識をした際に、データの保存の方法にもよるが、ベクターデータを活用している。ベクターデータというのは、例えばこれは猫らしいように見える、虎かもしれないけど、より猫らしいというような、方向性データを全ての画像等で持っている。Visual Question Answering といい、例えば一枚の絵の中に、猫が家の中にいると、虎が家の中にいる事はなさそうなので、これは猫らしさがより強いという関連性を持って判断をする形になっている。また、データが使われた経緯については、先ほどグラフデータという話をしたが、グラフという形で画像であったり、音声であったりといったものについて、「らしさ」を統計的に見ている。単純に画像をアップして比較をしていくときに、そういったベースデータをマイクロソフトでは沢山持っているので、「らしさ」というデータをもとに提供できる。皆さんがデータを自分たちで保有する場合にも、ベクターデータの扱えるデータレイクを用意しており、そこで管理していただければ、AI をより使いやすくなると思っている。セキュリティという観点で何か違反があるとか云々に関しては共通の見解ではあるが、皆様の組織における何らかの課題に対して、先ほどお伝えした通り皆様の中で行っていただく必要があるかと思う。マイクロソフトはパブリックなサービスも出しており、そういった開発環境も行っているので、様々な答えになってしまったが、このような形でお答えさせていただきたい。

森主査）

両社ともに、基本的にはサービスそのものが頑健になるよう検討されており、それ以外の対策も講じられていると理解した。

時間を超過してしまったので、他にも質問があろうかと思うが、申し訳ないが本日はここまでとさせて頂きたい。

◆その他

次回の会合は 11 月 21 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