AIセキュリティ分科会（第6回）議事要旨

１．日　時）令和7年12月5日（金）10：30～12：00

２．場　所）中央合同庁舎 2 号館 8 階 第 1 特別会議室（ハイブリッド開催）

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、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】

堀内総務副大臣、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(敬称略)】

関根聡（国立情報学研究所（NII））

４．配布資料）

資料6－1 生成 AI LLM の安全性の確立（NII）

資料6－2　AI セキュリティ分科会取りまとめ（案）

5．議事概要）

（1）開会

（2）議題

◆議題（１）「LLM の安全性ベンチマーク構築の取組」について、NII 関根氏から、資料６－１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。

綿岡構成員）

AI セキュリティに関する課題感を伺いたい。私個人としては特に、静的なベンチマークデータセットにおいてどのように AI セキュリティのリスク観点を評価するかという点に課題を感じている。関根先生の中で、そういった評価の AI セキュリティへの課題感があれば共有いただきたい。

NII 関根氏）

セキュリティと一言で言っても、人によって定義がかなり異なるが、特にサイバーセキュリティは、静的なデータだけでは評価しきれないと思っている。そういったものも含めて、あり得る形を模索しつつ、静的なデータだけでやろうという前提ではない。色々な形があってよく、また、なくてはならないと思っているくらいである。そういった自由度を持って、その部分についても今後、特にセキュリティのグループを中心に考えていただきたいと思っている。

石川構成員）

Safety という点について、2 つの段階があると思っている。いわゆる爆弾の作り方はダメという話と、例えば業界固有の、社外に出してはいけない個人情報や専門用語といったものがあると思う。こういった業界固有のものについても、やはり同じようなアプローチが有効なのか、その辺りの見解をいただきたい。

NII 関根氏）

資料 6－1 の 38 ページに「分野依存」という項目があり、農業、エンタメ、ヘルスケアが立ち上がっており、分野依存は非常に重要である。各分野に特化したインストラクションを作る方向で考えてはいるが、もしかしたらそこも動的な評価が必要かもしれない。

私たちもその点は非常に重要だと考えており、コンテキストも本当に重要である。ただし、今回作るのは「最小公倍数」、最低限これだけは守らなければならないというものを作りたいと考えている。

森主査）

資料 6－1 の 38 ページに「安全性」に関わる項目がいくつか記載されているが、それぞれに対して個別のベンチマークが作成されるというイメージで良いか。

NII 関根氏）

同ページに記載されている安全性のうち、Jailbreak 以外の項目は Answer Carefully でカバーしている。それぞれに対して数百個ずつのデータを作成済みだが、現在の Answer Carefully の安全性データではほぼ満点が取れてしまい、評価対象間の差異が生じないため、これらをより高度化させる必要がある。したがって、より適切な評価ができるよう検討してもらっているのが現状である。

◆議題（２）「AI セキュリティ分科会取りまとめ（案）」について、事務局 中村補佐から、資料６－２を説明。

◆構成員の意見・コメント

森主査）

これより、（２）について、ご質問や自由討議の時間を設けたい。

披田野構成員）

資料 6－２の４ページ 図１の右側にある「外部連携システム」について、これ以降のページではデータベースが含まれているため、本ページのみ含まれていない事に違和感がある。含めた方が良いのではないか。

同ページの「対象とする AI」について、本文内に図１に関する説明が無いように見受けられるのが気になった。

中村補佐）

指摘通りデータベース等の事例を含めた方が、今後の事例との平仄も合うため修正をしたいと思う。

図１の説明についても、指摘を踏まえて検討したい。

綿岡構成員）

資料 6－２の 28 ページ 2.2.1 入力プロンプトの検証の具体例について指摘したい。

”ガードレールによる検証”の中に「入力するプロンプトの内容を評価する役割を与えたガードレール(LLM-as-a-Judge と呼ばれる別個の LLM)」との記述がある。一般的にガードレールとは、入出力を検査するシステム全体を指すものと考えられるが、当該記

述では、LLM ベースで動作させた有害検知モデルをガードレールと呼んでいるように見受けられる。したがって、修正案として「ガードレールモデル」とするか、「LLM-as-a-Judge」や「有害検知器」といった他の用語を用いる方が適切ではないか。なお、この点は資料 6 － 2　32 ページ 2.2.4「出力データの検証」についても同様である。

中村補佐）
　いただいた修正案も含めて検討させていただく。

福田構成員）
　今回、スコープの定義等も記載され、大変分かり易いドキュメントになってきたと思う。資料 6－ 2 の 5 ページの「想定読者」について、AI 開発者・提供者・利用者の区分は重複している部分もあり、ガイドラインの想定読者を限定的な区分で定義するのは難しいと思う。エンタープライズ向けにシステムを提供する場合には、利用者側のデータベース等、社内情報を含む資産に対して、アクセスコントロールがきちんと適用されていることを前提として進められることが多い。そういった場合、利用者に関しても、ある程度セキュリティに関しては責任を持ってもらわなければいけない部分がある。将来的な話になると思うが、想定読者は広げても良いかなと思っている。
　また、Appendix に入っている AI エージェント等、自律的に動くものが出てきた時に、システム間、人間とエージェント間の責任分界点という論点も出てくる。そういった意味でも想定読者の部分は AI を利活用する事に関わる人達が、セキュリティに関して意識できるような形に持って行くことができればと思う。

神谷企画官）
　指摘のあった想定読者の範囲拡大等、将来的な課題になり得る点について、事務局としてよく認識しておきたい。

北條構成員）
　資料 6－ 2 の 12 ページの表 2 及び表 3 の記述において確認したい。表 2 を説明する文章には「主な対策の概観」とあり、表 2 のタイトルにも記載されている。他方、表 3 のタイトルにも概観と括弧で記載されているが、文章には「主な対策」としか記載されていない。文章からすれば、表 2 は「主な対策の概観」であり、表 3 は「主な対策」ということイメージなのだろうか。
　また、表 3 内の行タイトルにおいて「対策の例」となっている一方、表 2 の同じ行タイトルでは「AI 提供者における対策」と対策の例ではなく限定的に読み取れる表記になっている。脚注において表 2 及び表 3 ともに網羅的ではないと記載されているので、表現を統一した方が良いのではないか。

神谷企画官）
　平仄が合っているか確認の上、必要な修正を施したい。

高橋構成員）
　Safety の定義が欲しい。定義がないようであれば ITU-T や IETF 等の国際標準を参照すると良いのではないか。

神谷企画官）
　Safety に関しては、本ガイドラインがスコープとしているセキュリティよりも広い概念であると認識している。資料 6－２の 3 ページに記載されている AI セーフティに関する関連文書等を含め、引用できる定義がないか検討したい。

石川構成員）
　資料 6－２の 23 ページ以降に記載されている LLM に対する対策を実装する際、レスポンスタイムやユーザエクスペリエンスに影響を与えると思っている。この点に関して注意書きを入れた方が良いと思う。

神谷企画官）
　承知した。

森主査）
　構成員の皆様から意見をいただけたと思う。
　本日議論いただいた取りまとめ案の位置づけであるが、事務局からの取りまとめ案の説明でも言及があったとおり、取りまとめを年内に策定し、その確定版を踏まえた上で、総務省において「AI のセキュリティ確保のための技術的対策にかかるガイドライン(仮称)」案を年内に作成し、これをパブリックコメントに付した上で公表することが予定されている。
　この前提のもと、取りまとめの確定については、本日ご議論いただいた内容やご意見等を踏まえ、事務局とも相談の上、最終的に主査である私に一任いただければと考えているが、いかがか。
　異議がないため、そのように進めさせていただく。

◆その他
◆各構成員から、挨拶
◆堀内副大臣から、挨拶

（３）閉会

以上