AI 時代の安全保障と主権・ガバナンス：我が国が採るべき二段階戦略

佐藤　達夫
株式会社 Citadel AI
代表取締役　小林 裕宜

はじめに

安全保障領域における先端技術、とりわけ高度な情報処理基盤の導入を巡る議論が、近年急速に具体性を帯びてきている。その中核に位置するのが「人工知能（AI）」である。大量のセンサーからリアルタイムでデータを収集し、最適な行動計画を瞬時に提示する高度な意思決定支援システムは、我が国の国家安全保障をも左右する重要な要素となりつつある。

急速な人口減少に伴う人的資源の制約が、安全保障分野を含めた我が国の構造的な社会課題となっているが、それ以上に深刻なのは、現場で求められる分析や決断のスピードが、人間の認知や判断の限界を遥かに超えつつあるという現実である。膨大な情報をリアルタイムで処理し、瞬時に情勢を分析する高度な情報処理基盤や意思決定支援システムの導入は、もはや避けて通ることができない。

しかし、こうした先端技術の適用には、技術的な不確実性やセキュリティリスクという深刻なガバナンス上の課題が伴う。さらに、基盤となる最先端のフロンティアモデルや AI エージェントシステムを巡る技術力において、我が国は米国や中国に対して圧倒的な格差をつけられており、この技術的格差を短期間で挽回することは極めて困難な状況にある。

このような現実を踏まえた時、我が国が採るべき安全保障分野における最善の AI 戦略とはいかなるものか。我が国が目指すべきは、他国製の先進的な AI システムの導入という不可避の現実を受け入れつつも、その外側に「我が国独自の主権的検証・制御システム」を構築することで、世界で最も信頼できる AI ガバナンス国家として、AI の信頼性、安全性、説明責任及び主権を保証する能力を確立し、将来の国産 AI 開発の足がかりとすることである。本稿では、その実現に向けた「二段階の主権確保戦略」を提言する。

AI を巡るグローバル潮流と不可避の現実

米欧の主要国が公開している政府報告書や調達情報を分析すると、安全保障領域における情報処理システムの刷新は、単なる業務効率化の域を

完全に超え、事態の推移に応じた意思決定の迅速化と最適化のフェーズへと移行していることが窺える。膨大な衛星画像、電波情報、物流データ、さらには SNS や国内外の報道をはじめとするオープンソースの情報（OSINT）を瞬時に統合した上で、全体の状況を多角的に可視化し、複雑な状況下でのリスク評価・複数の対応選択肢を自動提示する等、先端の高度情報処理基盤は現場での実用フェーズへ既に移行しつつある。

こうした高度情報処理基盤がもたらす技術的優位性は、事態の展開スピードが人間の認知限界を遥かに超えていく中で、状況変化に対応するタイムフレームを劇的に短縮する点にある。いわゆる OODA ループ（観察・情勢判断・意思決定・行動）を極限まで加速させるシステムを配備した組織に対し、従来のアナログな人間による情報処理プロセスだけで、適時適切に対応することは実質的に困難である。高度な情報処理基盤の整備は、純粋に技術的な選択肢というだけではなく、もはや現代の安全保障インフラにおける死活的な要素となっていると考えるべきである。

先端 AI が内包する安全保障上のリスク

一方で、フロンティアモデルや AI エージェントシステムのような最先端 AI システムを次世代意思決定支援システムに組み込むことには、伝統的なソフトウェアとは本質的に異なる固有のリスクが内包されている。

第一に、「判断プロセスの不透明性」と「ハルシネーション」である。
先端 AI の判断プロセスは確率論的であり、なぜその結論に至ったのかというロジックを人間が完全に追跡し説明することは困難である。さらに、ハルシネーション（尤もらしい嘘）は、1%の誤認が致命的なリスクの拡大を招きかねない安全保障分野において、極めて重大な不確実性をもたらす。

第二に、「サイバー・セキュリティリスク」である。
次世代意思決定支援システムは、意図的なプロンプトインジェクション（悪意ある指示の混入）やジェイルブレイク（安全制限の無効化）等による敵対的な攻撃対象となりやすい。仮に意思決定支援システムがこうした攻撃を受けた場合、機密情報が漏洩したり、あるいは最悪の場合、システムの制御権が乗っ取らたりする事態が発生する。

第三に、「他国製ブラックボックスの採用に伴う主権上の課題」である。
他国製の高度な AI プラットフォームを採用する場合、そのコアとなるアルゴリズムや学習データ、学習モデルの内部パラメータ（重み）は供給国の最高機密、あるいはベンダーの知的財産として「ブラックボックス」化されるのが常である。
AI システムは継続的にアップデートされる。アップデートにより、判断

ロジックが僅かに変化するだけで、AI が提示する優先順位、リスク評価、行動提案が徐々に変化する可能性もある。
システムがどのように学習し、アップデートされ、どのようなロジックで判断を行っているのかを自国で検証できない状態は、安全保障上の判断を他国の技術インフラに全面的に委ねることを意味し、主権の維持という観点から重大な課題を提起する。

直面する開発力格差、時間と主権のジレンマ

これらのリスクに対処する最も理想的な解は、自国で完全に制御可能な「国産の先端 AI システム」を開発することである。しかし、ここで我々は技術的・資本的な現実の課題に直面せざるを得ない。

現在の先端 AI、とりわけフロンティアモデルや自律型エージェントシステムの開発には、天文学的な規模の計算資源（GPU）と、膨大なデータ、そして世界トップレベルの AI 人材の集積が必要不可欠である。この分野において、米中の巨大テクノロジー企業や国が主導するプロジェクトでは、年間数十兆円という、文字通り「桁違いの巨額投資」が常態化している。これは単なる資金力の多寡に留まらず、先端半導体の圧倒的な調達力、巨大なデータセンターを稼働させる電力・インフラの確保、そして膨大な試行錯誤から得られたノウハウの蓄積という、一朝一夕には埋めがたい「構造的・時間的な格差」である。安全保障の現場において、技術の遅れはそのまま抑止力の低下を意味する。

AI の性能は、どれだけ良質で大量の「本物のデータ」で学習させ、常に最新のデータで改善したかで決まるといっても過言ではない。実戦経験のない我が国には当然のことながら学習・改善に必須な「実戦データ」が致命的に不足しており、我が国が独自に米国あるいは中国のレベルに匹敵するだけの最高峰の先端 AI システムを開発することは、技術的にも運用的にも極めて困難である。

さらに開発した AI を一回限りではなく、継続的にバージョンアップしていくだけの底力をつけるためには、少なくとも今後数年から十数年の歳月を要すると考えられ、その間の技術的空白期間を放置することは許されない。すなわち、「国産開発には時間がかかるが、目の前の脅威に対抗するためには今すぐ最先端の AI が必要である」という、時間と主権のジレンマを如何に解決するかが、現在の我が国に課せられた最大の論点である。

我が国が採るべき「外側からの主権的検証・制御」戦略

このジレンマを打破するための現時点におけるベストシナリオ、それこそが、他国製の先進的 AI プラットフォームを活用しつつ、そのシステム

の外側に「我が国独自の主権的検証・制御システム（ガードレール）」を構築する戦略である。

最先端の AI プラットフォームの多くは、オープン API やモジュール化された拡張アーキテクチャを採用しており、サードパーティ製のツールや独自のシミュレータを接続できるように設計されている。このため、最先端 AI の内部構造そのものには触れずとも、その入出力ならびに途中のコミュニケーション・フローを捕捉することが可能である。

こうして得たデータに対して、歴史・文化・地政学ならびに政策・戦略等「我が国固有の評価観点」に基づき、AI の行動を「外側」から分析することで、不適切な入出力や異常な振る舞いをフィルタリングし、その安全性やセキュリティを外部から独自に制御する仕組みを構築する。こうした外部構造を取り入れることで、機密情報の漏洩やシステムの乗っ取りを防ぐと同時に、我が国としての主導権を保ちつつ、主権的な検証・制御体制を確立する。

逆照射的アプローチが拓く、我が国の勝ち筋

この「外側からの検証・制御」戦略は、単なる当面の安全性の確保やリスク回避の手段に留まらない。我が国が安全保障分野の AI の開発・導入において長期的な自立性を確立するための、極めて地道かつ確実な「布石」となる。

第一に、他国製の最先端システムを運用しつつ、その入出力を我が国独自のシステムで検証し続けることで、「実際の運用データ」と「検証・評価の知見」が国内の先端技術メーカーや研究機関に安全に蓄積される。安全保障分野において、どのようなデータが入力され、AI がどう判断し、どのような誤りを犯しやすいのかという実戦的な運用知見は、いかなる計算資源よりも貴重な資産である。さらに機密情報の漏洩・流出の防止にも繋がる。

第二に、この「検証・制御システム」自体を我が国独自の強み（コア・コンピタンス）として育成することができる。AI の品質やセキュリティリスク、安全性を評価・制御するテクノロジーは、これからの AI 社会において「信頼性の担保（トラスト）」という巨大な一分野を形成する。プラットフォームそのものは他国製であっても、その安全性を外側から保証する「高精度な評価・制御システム」を我が国が独自に保有することは、欧米諸国等に対する強力なバーゲニング・パワー（交渉力）となり、真の意味での相互運用性と主体性の確保に寄与する。

そして第三に、ここで蓄積された運用データ、評価ノウハウ、およびガードレール技術を統合・発展させることにより、将来的には特定のドメ

イン（例えば、我が国固有の地形や周辺環境に特化した意思決定の立案等）における国産独自モデルの開発へとスムーズに繋げることが可能となる。まずは「検証・制御」という勝機のある分野に加えて、信頼性、安全性、説明責任、運用規律、品質保証という分野で我が国独自の AI ガバナンスの生態系を築き、そこから国産の先端 AI モデルの開発へと逆照射的・長期的にアプローチしていく戦略こそが、資源と時間に限りがある我が国にとって勝率の高い道筋である。

おわりに

最先端 AI システムを丸ごと保有せずとも、その入出力と信頼性を自国の手で制御できれば、安全保障の最後の砦である「人間の意思決定」と「運用の主体性」は揺るがない。我が国が誇る精緻な品質と信頼性を重んじる文化を、この「検証・制御」の分野に集中投下し、強固な次世代意思決定支援システムのガードレールを構築すること。それこそが、岐路に立つ我が国の安全保障を実質的に担保し、将来的な技術的主権へと道を拓く、これからの我が国の主権を支える新たな防衛基盤構築のベストシナリオであると確信する。