「人工知能（AI）が浸透するデータ駆動型の経済社会に必要な AI セキュリティ技術の確立」に関する研究開発構想（個別研究型）

令和４年１０月

内閣府

文部科学省

目次

1　構想の背景、目的、内容

1.1 構想の目的

1.1.1　政策的な重要性

人工知能（AI）の社会実装は、民生部門・公的部門双方において着実に広がり、広範な産業領域や社会インフラなどで AI 技術は大きな影響を与えている。しかし、不正アクセスにより秘匿性の高い学習データが復元されて漏洩するリスクや、AI アルゴリズムが窃盗・改ざんされることで AI の判断が意図的にゆがめられてしまうリスクなど、AI そのものを守るセキュリティ（Security for AI）に関する脆弱性がどのようなものなのか、国際的にもまだ十分に理解されていない。

また、AI を活用したサイバーセキュリティ対策（AI for Security）に関しては、実際に AI を活用したセキュリティ製品やサービスの商用化が進んでいる。一方で、攻撃そのものに AI 技術を活用した新たな攻撃手法が広まるなど、年々複雑化・巧妙化するサイバー攻撃に対処することが求められている。

このように我々の経済社会の基盤の保持の観点から、AI セキュリティの研究は近年最も注目度の高い分野の一つであり、我が国としても、サイバーセキュリティ戦略（令和 3 年 9 月 28 日閣議決定）や AI 戦略 2022（令和 4 年 4 月 22 日統合イノベーション戦略推進会議決定）において、AI セキュリティの必要性が述べられており、国として独自の技術力を確保することが必要である。

本構想は、個別研究型として、こうした背景の下、自由、公正かつ安全なサイバー空間の確保に資する支援対象とする技術として研究開発ビジョン（第一次）において定められた「AI セキュリティに係る知識・技術体系」の整理・獲得を目指すものである。

1.1.2　我が国の状況

ここ数年 AI セキュリティに関する分野に参入する研究者は増加傾向にある。学術面では、例えば、AI の誤認識を誘発し得る敵対的サンプルの生成を試みる研究や、AI により自動的にマルウェアの機能分析を行う研究など、個別研究課題が進みつつある。しかし、社会における AI の活用が試みられて日が浅く、かつ、AI とセキュリティの境界領域として捉えられてしまい

個別の領域の専門家が積極的に参入しづらいことから、研究者・技術者のコミュニティ醸成は十分ではない。

また、「Security for AI」に関しては、情報セキュリティの３要素であるCIA（Confidentiality（機密性）、Integrity（完全性、保全性）、Availability（可用性））に相当するセキュリティの基本的な考え方や社会的側面への影響に関する知見が十分に蓄積されていない。

「AI for Security」に関しては、AIによるセキュリティ検知の精度向上に向けた取組が行われているが、日々新たな攻撃手法が開発され、裾野が広がっていくサイバーセキュリティのリスクへの対応や、高精度のAIモデル実現のために必要な産業界や行政からの具体的な課題や機械学習に必要なデータの収集・共用が難しいといった課題がある。

1.1.3　世界の取組状況

米国では2013年ごろからGoogleやAppleを筆頭にSecurity for AIに関する新たな概念が次々提案されるなど、この領域での学会での採択論文数の著者の約半数が米国である。サイバーセキュリティの中では現在最もホットな研究分野の一つであり、どの国も注力してきている。

国防高等研究計画局（Defense Advanced Research Projects Agency：DARPA）は2018年9月に、次世代AI開発を行う「AIネクスト（AI Next）」キャンペーンを開始し、新規・既存プログラムに対し、複数年に亘って総額20億ドル超を助成することを明らかにした。対象となる主要分野には、AIシステムのロバスト性及び信頼性の改善や、機械学習・AI 技術のセキュリティ及び耐障害性の強化といった課題が含まれており、Security for AIに関する取組を推進している。

欧州では、Horizon Europeの中で、AI for cybersecurity reinforcement（AI 技術コンポーネントとツールを用いたサイバーセキュリティを強化）の研究開発に取り組んでいる。主な内容としては、(i)システムの堅牢性の向上、(ii)システムの回復力の向上、(iii)AIベースの方法とツールを開発、(iv)システムレスポンスの向上、(v)AI が攻撃に使用できる方法に対する対抗策が挙げられている。

1.1.4　構想のねらい

本構想では、我が国においてAIセキュリティに係るリスクが今後顕在化した際に、自らの技術力で課題の理解・解決ができるよう、国として独自のAI セキュリティの知識と技術力の確保を目指し、産学官の技術力向上を図ることを目的とする。そのために、我が国の技術の優位性の獲得に繋がり得る自律性の確保も念頭に、産学官の人材層を幅広く糾合し、様々なアプローチによる研究開発を進めることで、AI セキュリティに関する必要な知見蓄積や、知識・技術体系の整理・獲得を目指す。

1.2　構想の目標

1.2.1　アウトプット目標

AI セキュリティに関する必要な知見蓄積や、知識・技術体系を整理するとともに、具体的なシステムを対象としたプロトタイプの実証実験を実施する。併せて、産学官の複数の研究チームにより基盤的な研究開発を行う中で、全国的な情報発信や研究者間の連携を促し、その後の展開あるいは社会実装に繋げていく。

1.2.2　アウトカム目標

AI セキュリティのユースケースの横展開、研究コミュニティの醸成、研究開発人材の確保、国や関係機関による AI セキュリティ技術の理解促進、幅広い事業主体（政府機関や民間企業、インフラ関係など）による国内機関からのセキュリティ技術調達が可能になることを目指す。

1.3　研究開発の内容

1.3.1　研究開発の必要性

- Security for AI

  AI に関するセキュリティの基本的な考え方を構築していくためには、主だったAI技術（画像認識、音声認識、自然言語処理等）を対象に、それらを活用する際のセキュリティリスクを分析しつつ、CIAを確保するために必要な知識や技術、社会への影響を明らかにするとともに、防御技術を獲得していくための研究開発が必要となる。

- AI for Security

  攻撃そのものに敵対的生成ネットワーク等の AI 技術を活用した精巧なフェイクなど新たな攻撃手法が日々開発される状況に受け身にならないよう、将来の実装が見込まれる社会システム（自動運転など）を視野に入れながら、具体のユースケースをもとにした産学官の連携体制を構築し、オフェンシブセキュリティ（攻撃者の視点から知見を得る）のアプローチを考慮した研究開発を進めることが必要となる。

1.3.2 研究開発の具体的内容例

- Security for AI
  - 海外を含めた最先端動向を収集しながら、主だった AI 技術（画像認識、音声認識、自然言語処理等）について、活用環境や様々な攻撃・悪用等に応じて想定されるリスク（社会的影響を含む）とその対応に必要な知識・技術の分析・整理
  - 各 AI 技術を対象とした、敵対的サンプルやポイズニング攻撃をはじめとする悪意のある入力・攻撃の検知の高度化や、それらの無毒化、およびそれらに耐性のある AI モデルなど AI 防御技術

- AI for Security
  - 実際に高度なサイバー攻撃を受けた事例を募り、当該システムを対象にセキュリティシステムについて要件定義を行うことによる、必要となる防御の核となる要素技術の特定
  - 先端的な攻撃技術の知識・技術を取り込みながら、攻撃者の視点から知見を得るとともに、これらを活かした革新的な AI 活用によるセキュリティ技術
  - 一つの組織で学習データを収集する困難を克服するための、プライバシー保護にも配慮しつつ領域・業種ごとに複数の組織が学習データを連携・利活用することができる連合学習技術

1.3.3　研究開発の達成目標

- Security for AI

  主要な AI 技術を対象に、CIA の確保や社会的影響への対応に関する研究開発を進めるための考え方・方向性を整理する。当該整理に基づき、AI が活用された具体的なシステムを対象として、AI 防御技術のプロトタイプの開発・実証を行う。

- AI for Security

  AI 活用によるセキュリティ技術の需要が高いいくつかの領域・業種において、複数の組織が学習データを連携・利活用する連合学習技術のプロトタイプの開発・実証を行う。また、具体的なシステムを対象として、最先端の攻撃技術への対応を盛り込んだ革新的な AI 活用によるセキュリティ技術のプロトタイプの開発・実証を行う。そのほか、仮想システムにおいて攻撃・防御を行う模擬対戦による技術の高度化と人材育成、コミュニティ拡大に取り組む。

以上について、より具体的には、提案者の設定した個別の達成目標を基本としつつ、文部科学省及び JST のサポートの下、採択後、研究開発を開始するにあたって行う研究計画の調整にて定めるとともに、研究開発開始後においては、協議会における意見交換（脅威情報等の共有・分析、対策情報等の作出・共有等を含む）の結果も踏まえ、必要な場合、見直しを行う。

2　研究開発の実施方法、実施期間、評価、社会実装に向けた取組

2.1　研究開発の実施・体制

研究開発のフレームワークを構築するため、研究開発対象となり得る技術・システム等についての情報収集・調査研究を行う。その結果を踏まえ、プログラム・オフィサー（PO）、関係府省等による意見交換を経て研究開発課題を決定し、産学官の複数の研究代表者による AI セキュリティの基盤的な研究開発を行う。当該研究開発開始から 2、3 年目を目途に、それまでの成果をもとに、プロトタイプ開発・実証に向けた新たな研究機関又は研究代表者の追加など研究チームの再編成を必要に応じて行う。

PO の指揮・監督の下、研究代表者（研究開発課題の実施責任を法人が担

う場合は当該法人を含む。以下同じ。）が研究開発構想の実現に向け責任を持って研究開発を推進する。JST 等の助言に基づき、研究代表者は、適切な技術流出対策を行うよう体制を整備するとともに、研究インテグリティの確保に努め、適切な安全保障貿易管理を行うよう、これらを推進するとともに、研究開発に必要な事項を行う。

研究開発成果を民生利用のみならず公的利用につなげていくことを指向し、社会実装や市場の誘導につなげていく視点を重視するという本プログラムの趣旨に則り、研究代表者は PO 及び研究分担者との協議の上、知的財産権の利活用方針を定めることとする。その際には、研究開発途中及び終了後を含め、知的財産権の利活用を円滑に進めることができるように努めることとする。

なお、研究開発成果の利活用にあたりその成果にバックグラウンド知的財産権が含まれる場合には、その利活用についても同様に努めることとする。

また、当該分野における民間企業等における処遇水準を踏まえ、研究開発に従事するリサーチ・アシスタント等人件費の支弁を受ける者には、その報酬等について、これに相応しい水準を支弁する。具体的には、担当する PO が研究計画を踏まえ調整した上で、JST が決定するものとする。

2.2 研究開発の実施期間

各研究開発課題の実施期間は原則 5 年以内とする。構想全体で最大 25 億円程度の予算を措置する。

2.3 評価に関する事項

自己評価は毎年実施する。外部評価については、原則、研究開発の開始から 3 年目に中間評価、研究開発終了年に最終評価を実施する。具体的な時期については、担当する PO が採択時点でマイルストーンを含む研究計画とともに調整した上で、JST が決定するものとする。

2.4 社会実装に向けた取組

本構想は、知見蓄積、知識・技術体系の整理、プロトタイプの実証実験の取組を通じて、国内機関からのセキュリティ技術調達が可能になること等

を目指すものである。このためには、研究代表者と潜在的な社会実装の担い手として想定される関係行政機関や民間企業等との間で、情報セキュリティのインシデント情報や脅威情報、それらの分析結果、対策手法等の情報共有や、社会実装イメージや研究開発の進め方を議論・共有する取組等の伴走支援が有効である。

したがって、今後設置される協議会を活用し、参加者間で機微な情報も含め、社会実装に向けて研究開発を進める上で有用な情報の交換や協議を安心して円滑に行うことのできるパートナーシップを確立することが重要であり、関係者において十分にこの仕組みの運用を検討する必要がある。なお、協議会の詳細は別に示す。また、PO は研究マネジメントを実施する際には、協議会における意見交換の結果も踏まえるものとする。