ベンチマーク構築プロジェクトへの期待

令和8年5月

内閣府人工知能政策推進室

齊藤大地

人工知能戦略本部　総理ご指示（R7.12.19）

ＡＩは、産業競争力や安全保障に直結。信頼できるＡＩによる日本再起を実現するため、以下を指示（以下、抜粋）

- 第一に、『ガバメントＡＩ源内』の徹底活用です。2026年５月から１０万人以上の政府の職員が活用できるようにします。ＡＩ源内の活用により、創造的に業務を行い、国民の皆様に信頼できるＡＩの意義を示してください。

- **第二に、ＡＩセーフティ・インスティテュートの抜本的強化**です。ＡＩの安全性に対する不安が高まる中、**英国並みの２００人体制を目指し**て、小野田大臣と赤澤経済産業大臣は、**全省庁、産学から人材を集結させ、ＡＩセキュリティに万全を期してください**。

- 第三に、ＡＩロボットを始めとしたフィジカルＡＩに不可欠な信頼できる国産の汎用基盤モデルの開発です。赤澤経済産業大臣は、質の高い産業データを日本の競争力の中核に位置づけ、意欲ある企業としっかりと連携し、開発を進めてください。

- 第四に、信頼できるＡＩによる社会課題を解決できるサービスの開発・導入です。今般の経済対策で、４０００億円以上のＡＩ関連施策を措置したところです。これらを活用して、地域や中小企業の成長戦略を実現するとともに、世界各国にサービスを展開してください。

- 第五に、信頼できるＡＩを世界とともに創りあげるため、『ＡＩサミット』を可能な限り早期に日本で開催すべく、関係省庁を挙げて、取組を進めてください。

- 第六に、信頼できるＡＩを創る官民投資を日本成長戦略における危機管理投資として、力強く推進してください。政府としては、投資の予見性を高めるため、当面、１兆円超をＡＩ関連施策の推進に投資してまいります。また、大胆な投資促進税制を創設し、研究開発税制を深堀りします。これらの政府のコミットを、それぞれが所管する企業の皆様と共有し、政府の取組に呼応していただきＡＩ投資を強力に推進してください。

- 結びに、ＡＩをめぐる動向の変化は非常に速いです。小野田大臣は、今回の計画に基づく、官民の取組を直ちに実施するとともに、来年の夏を目指して、投資目標、制度改革、人づくり、データ戦略などを含む官民投資ロードマップを盛り込む形で、『ＡＩ基本計画』を更に充実させてください。以上です。

自民党デジタル社会推進本部　AI･web3小委員会提言(R7.12.19)
～ＡＩセーフティ･インスティテュート（ＡＩＳＩ）の機能強化に係る緊急提言～

我が国のＡＩに関するイノベーションの促進とリスク管理を両立させるためには、**「信頼できるＡＩ」の利活用及び開発の中核となるＡＩセーフティ･インスティテュート（ＡＩＳＩ）の抜本的な機能強化**を行わなければならない。**ＡＩモデルの技術的評価、広範な適正性に係る評価、セキュリティ面での対策を実行できる体制の構築**を行う必要がある。

このＡＩＳＩの機能拡充及び機能強化においては、**政府を挙げた取組み**が必要であり、特に以下の二つの目標を早急に達成しなければならない。

まず、世界のＡＩ開発事業者から、フロンティアモデルの発表、提供に先立ち、事前評価の実施を委託される機関となる。当面は他の独立行政法人や民間機関等との連携の下、将来的には自ら、技術評価能力の強化とそのための研究開発基盤を構築する。**世界の主要開発事業者との協力協定を積極的に締結**する。

また、顕在化する「ＡＩによるサイバー攻撃とＡＩによる防御」に対応できるよう、諸外国のＡＩＳＩや内外の関係機関と連携し**サイバーセキュリティの評価機能を強化**する。サイバーセキュリティに関する専門人材をはじめ人的基盤を強化する。

ＡＩＳＩを軸とした日本として安全性やセキュリティ確保に係る国際ネットワークをグローバルサウスを含めて構築し、AIサミットの日本での早期に開催も行うことで、日本の「信頼できるＡＩ」を世界に拡げていく。

そこで、ＡＩ･ｗｅｂ３小委員会･デジタル社会推進本部として、ＡＩＳＩの機能強化について、下記のとおり緊急提言する。

背景としては：日本AISI機能強化の必要性

■ 2023年11月、英国でのAI安全性サミットを契機に、英・米がそれぞれ国内にAISIを設立。
日本も、**2024年2月に日本AISIをIPAに設置。AISI国際ネットワークを形成。**

■ **日本AISIの予算、人員は英米に比べて圧倒的に少ない。**
国際ルール形成主導に向け、**産官学の人材、知見、資金を糾合して機能を強化する必要。**

	日本 AI Safety Institute	英国 AI Security Institute	米国 Center for AI Standards and Innovation
設立	2024年2月	2023年11月	2024年2月（25年6月にCAISIへ改名）
所管	経済産業省・デジ庁 IPA（情報処理推進機構）内	科学・イノベーション・技術省	商務省（NIST内）
所長	村上明子	Adam Beaumont	Austin Mayron
職員数・予算	**31人（併任含む*）** 令和6年補正：**3.8億円** ＊IPAや理研からの併任	**約200人以上*(うち専門家は90人)(目標300人)** 初期予算：**£1億（約200億円）** *2025年9月時点、大学教授、元Google，元Open AI等トップ人材を採用 *トップAIモデルへの特権アクセス及びコンピューティングへの優先的アクセス	**30人程度（目標80人）**** 2024年度:**予算＄1000万(約15億円)** **2024年時点情報
役割	• AI事業者ガイドラインの策定支援、米国ガイドラインとの相互比較を実施。 • 評価観点ガイド、レッドチーミング手法ガイドなど実務ドキュメント作成。	• フロンティアモデルの評価ベンチマークとテストプラットフォーム構築。 • AI安全性・セキュリティの最新研究の白書の発行。 • 米国AISIとの共同テスト、カナダAISIとの協力などAISIネットワークのハブ。	• OpenAI・Anthropic等との間で、フロンティアモデルの事前評価（プレリリース・テスト）協定。 • モデル評価・リスク管理の技術スタンダードを策定。 • Google、Microsoft、Anthropic 等200社超を巻き込んだ共同研究。

背景としては：アンソロピック　脅威インテリジェンス・ブリーフィング
(2025.11「初めて報告されたAI主導型サイバー諜報活動の阻止」)

1. **Claude**（アンソロピック社の生成AI)**を活用したサイバー攻撃が複数報告**。攻撃手法が**驚異的なスピードで大きく進化**。人間の関与が10~20%に留まり、**AIが自律的にサイバー攻撃するフェーズへ**。**技術や資金の少ない攻撃者でも、大規模かつ効率的なサイバー攻撃が可能**に。

2025年3月

- 英国拠点の脅威アクターがClaudeを活用し、**技術力不足を補い、ノーコードでランサムウェアを開発**。
- ダークウェブで高度なマルウェアを流通・販売（$400~1,200)。

2025年5月

- ロシア語を話すサイバー犯罪者がClaude Code（アンソロピック社のAIエージェント型コーディング支援ツール）を使い、国内外の17の標的に対して**大規模な恐喝を実施**。（要求額：ビットコインで＄75,000~500,000)
- Claude Codeが大規模な偵察、認証情報等の収集、ネットワーク侵入を自動化。

2025年9月

- 中国政府支援グループがClaude Codeを使い、**自律型サイバー攻撃エージェントを構築**。Claudeをオーケストレーションシステムとして用い、複雑な多段階攻撃を個別の技術タスク（脆弱性スキャン、認証情報の検証、データ抽出、横展開等）に分解することで、悪用検知が非常に困難に。
- **サイバーキルチェーン全体（脆弱性発見、侵入、自律的分析、横展開、権限昇格、情報流出）を概ねAIが自律的に実行**。

2. **AIは防御にも不可欠。AIを使った高度な侵入検知技術の向上やAIを使った自動診断・自動パッチシステムなどの安全対策の強化**がますます重要に。

出典：Anthropic, “Disrupting the first reported AI-orchestrated cyber espionage campaign”, https://www.anthropic.com/news/disrupting-AI-espior

背景としては：米Anthropic社の新たなAIモデル「Claude Mythos」

◆ 2026年４月７日、米Anthropic社（アンソロピック）は新たなAIモデル「Claude Mythos Preview（クロード ミトス プレビュー）」を発表

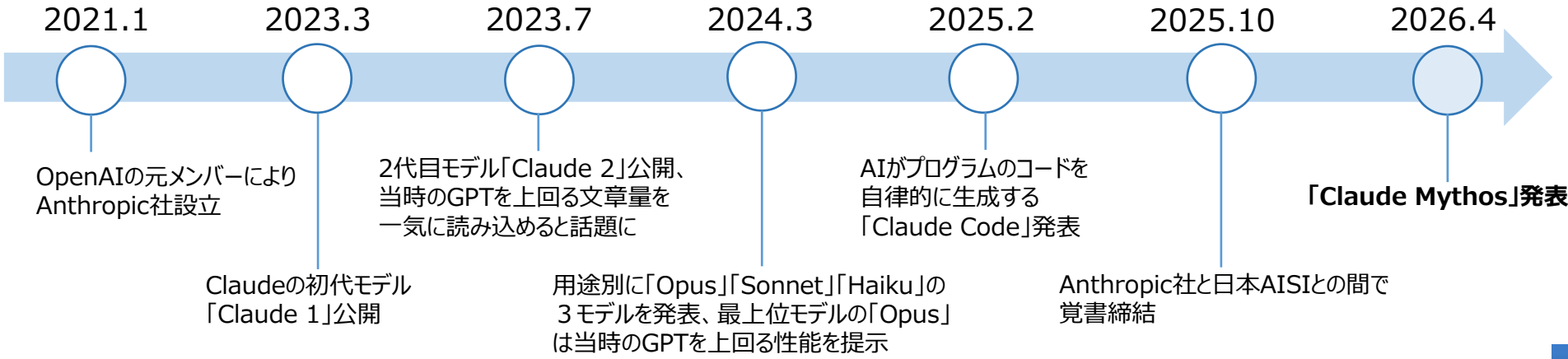
◆ 当該モデルは様々なAIの評価指標で過去最高となる性能を示しており、特に、サイバーセキュリティ上のリスクや脆弱性を見つけ出す能力が高いとのこと

◆ 他方、このリスクや脆弱性を見つけ出す能力をサイバー攻撃に悪用された場合、甚大な被害が発生するおそれがあることから、同社は当該モデルを一般公開せず、大手テクノロジー企業のみに提供し、防御目的に限って使用を認める方針を明らかにしたところ

※新イニシアチブ「Project Glasswing（グラスウィング）」を、Amazon Web Services（AWS）、Apple、Google、Microsoftなど主要企業11社と共同で立ち上げ。当該イニシアチブ参画企業及び重要なソフトウェア企業40社に対してモデルへのアクセス権を提供

※米国財務省は、大手銀行と緊急会合を開催し、金融システムへのサイバーセキュリティリスクについて警告し、防御体制の強化を要請

（参考）Anthropic社に係る年表



AI基本計画（R7.12.23）におけるAISIの機能強化について

ＡＩイノベーションの好循環を実現し、信頼できるＡＩエコシステムを構築するため、技術開発・実証・評価・運用の各段階において、適正性の確保につながるＰＤＣＡサイクルを 構築する。

これを実現するため、国民や事業者等の自主的かつ能動的な取組を促すよう、国としての基本的な考え方を提示する。当該考え方等を踏まえ、**ＡＩセーフティ・インスティテュート （ＡＩＳＩ）を抜本的に強化することで、ＡＩモデルの技術的評価を適切に行い、当該評価も踏まえ、ＡＩがもたらすリスクに係る実態把握を行うとともに必要な措置を講ずる。ＡＩＳＩの機能強化にあっては、世界屈指の英国 AI Security Institute の規模 をベンチマークとしつつ、人員を直ちに現行の２倍程度に拡充する**。

ＡＩの安全性確保やＡＩを利用した攻撃への対応が、新たなサイバーセキュリティ上の課題として認識されつつあることを踏まえ、体制整備を含めた適切な措置を講ずる。

（１） 信頼できるＡＩエコシステムの構築

ＡＩモデルの安全性にとどまらず、より広範な適正性に係る評価やセキュリティ面での対策を実行できる体制を構築し、技術的・制度的なガバナンスの強化を図る。 その中核として、**ＡＩＳＩの機能を、政府を挙げて抜本的に強化する。**

（２） ＡＳＥＡＮ等グローバルサウス諸国を含めた国際協調

広島ＡＩプロセスの推進や、**ＡＩＳＩネットワーク等の国際的な枠組みの活用により、ＡＩガバナンスの構築を主導する。**

基本計画推進WG中間とりまとめ（抄）

■ AISIの目的・ミッション

「信頼できるAI」の提供に必要な情報や技術を有し、 日本を世界で最もAIを開発・活用しやすい国にする

① ＡＩ安全性に関連する情報のハブとして、**信頼できる優れたＡＩを活用した製品やサービスの普及**

② ＡＩ技術やＡＩ安全性等に係る国際標準化等により、**我が国のＡＩ関連分野のイノベーションの加速・競争力強化**

③ 情報収集・分析・共有・対策の早期実施により、**ＡＩの開発・普及に伴う国民の生命・財産に危害が及ぶような事象など、国家の安全を脅かす事象の抑止**

上記の①～③を実現するために、AISIが中核的な役割を担いつつ、関係機関とも協力をしていく。

■ ミッション達成のための４つの活動

① ＡＩの物差しを作る

② 自ら評価する能力を持つ

③ ＡＩ関連情報の収集・分析・提供

④ 国際協調の主導

基本計画推進WG中間とりまとめ（抄）

② 自ら評価する能力を持つ

1．ＡＩによるサイバー攻撃・防御への対応（技術動向の調査・共有等）

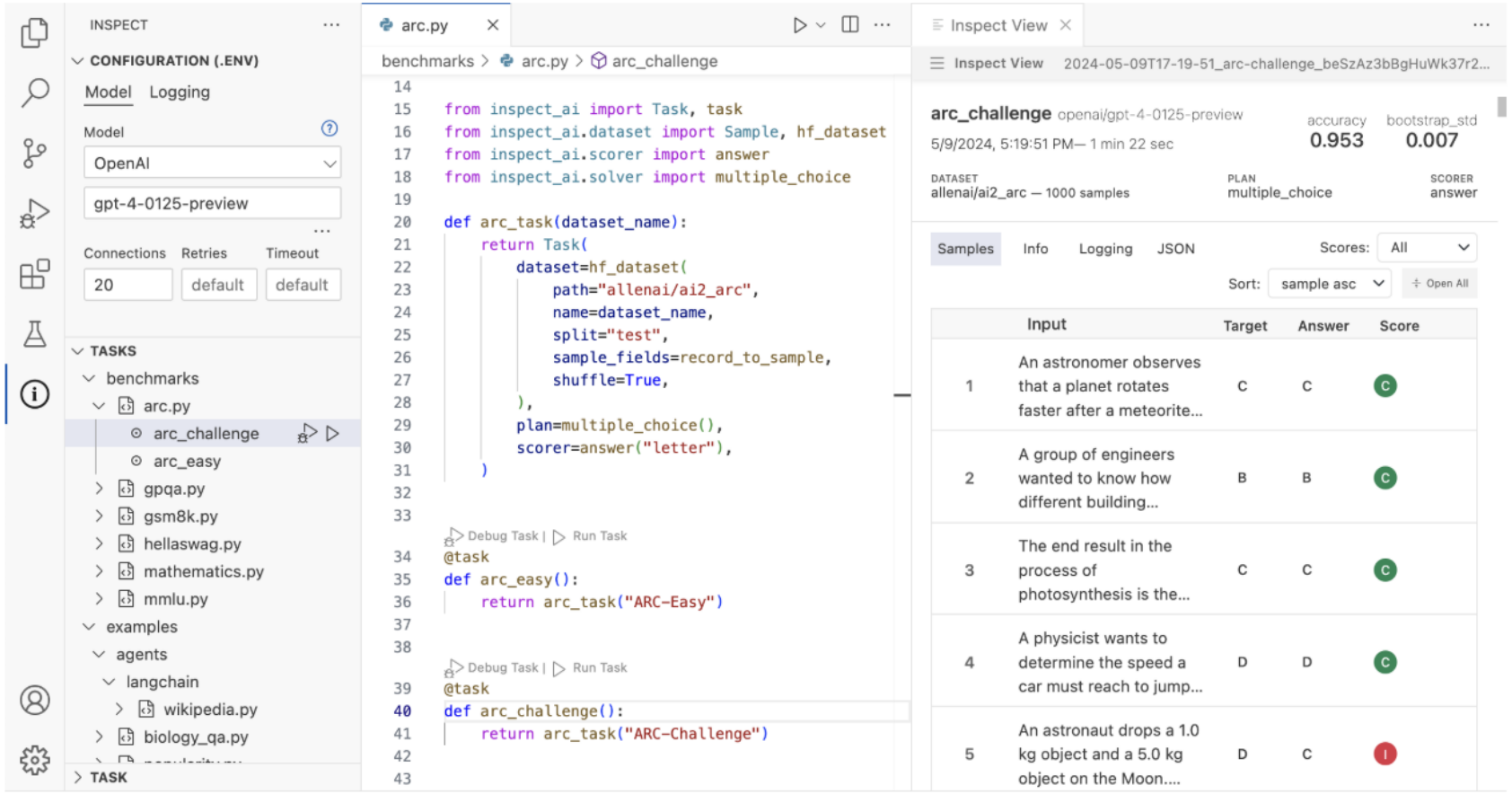
- ＡＩインシデントに係る情報※の収集及びその評価、政府関係機関（NCO、NSS、警察庁、防衛省、経産省等）と共有、対応策について官民が連携して取り組む体制の検討

2．新たなＡＩについて、影響の大きな悪用ができるかどうかを評価

- ＡＩの第三者評価が可能となる評価環境の本格的構築（IPAやNICT等のパートナーシップ機関との連携を図りながら推進）
- 構築したＡＩ評価環境を活用したフロンティアＡＩの評価（MoC締結企業の新製品評価、LLMの評価からＡＩエージェント、フィジカルＡＩの評価へと拡張）

ベンチマーク構築プロジェクトへの期待

- ベンチマーク構築プロジェクトで**「日本発」のLLMの安全性ベンチマーク**が出来ること大いに期待！
- 日本でも、**UK AISIのAI評価基盤「Inspect」ベースの評価ツールをOSSで公開していると思いますが、ベンチマーク構築プロジェクトで得られた成果も反映し、国内外の関係者が利用できる、日本発のＡＩ評価基盤が出来るようになることを期待**しています。



（英国AISIのAI評価基盤「Inspect」の構成）