参考資料

米国におけるAIの脅威・リスクの認知調査サマリ

2024年4月25日

独立行政法人情報処理推進機構
セキュリティセンター　企画部　調査グループ

IPA Better Life with IT

調査の概要・今後の予定

- 近年のAI技術の加速度的に進化する中、現状を把握すべく米国における**AIに関するセキュリティ脅威・リスクとその認知**に関する調査を実施。
- ８０以上の関連文献調査や、工学・法学・サイバーセキュリティリスク等、様々な専門家１０名のインタビューを実施。
- 調査期間は２０２４年1月～２月。
- **５月中旬**に調査結果概要（日本語）及び調査結果全体版（英語）を**公開予定**。

  - AISI関係府省庁向けには調査内容についてご紹介する場を別途計画しております。（P10参照）

【参考】参考文献（一部抜粋）

生成AI以降の最新動向が記された文献多数を調査

- Aspen Digital. ‘Envisioning Cyber Futures With AI’. Aspen Institute, 9 January 2024. https://www.aspendigital.org/report/cyber-futures-with-ai/.
- ‘Artificial Intelligence Index Report 2023’. Stanford University Human-Centered Artificial Intelligence, 2023. https://aiindex.stanford.edu/report/.
- Benaich, Nathan. ‘State of AI Report’. Air Street Capital, 13 October 2023. https://www.stateof.ai/.
- Funk, Allie, Adrian Shahbaz, and Kian Vesteinsson. ‘Freedom on the Net 2023: The Repressive Power of Artificial Intelligence’, 2023. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence.
- Hoffman, Wyat, and Heeu Millie Kim. ‘Reducing the Risks of Artificial Intelligence for Military Decision Advantage’. Center for Security and Emerging Technology, March 2023. https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/.
- McKinsey. ‘The State of AI in 2023: Generative AI’s Breakout Year’, 2023. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.
- Musser, Micah, Jonathan Spring, Christina Liaghati, Daniel Rohrer, Jonathan Elliot, Rumman Chowdhury, Andrew Lohn, et al. ‘Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications’. Center for Security and Emerging Technology& Stanford Geopolitics, Technology and Governance Cyber Policy Center, April 2023. https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/.
- National Cyber Security Centre. ‘The Near-Term Impact of AI on the Cyber Threat’. London, United Kingdom: National Cyber Security Centre, 24 January 2024. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.
- Vassilev, Apostol, Alina Oprea, Alie Fordyce, and Hyrum Anderson. ‘Adversarial Machine Learning A Taxonomy and Terminology of Atacks and Mitigations’. NIST AI 100-2e2023. NIST Trustworthy and Responsible AI. Gaithersburg, MD: National Institute of Standards and Technology, January 2024. https://doi.org/10.6028/NIST.AI.100-2e2023.

【参考】インタビュー実施者

米国のハイレベルな有識者10名からのインタビュー

	略歴
A	民間企業のチーフアーキテクト。AIのプロダクトエンジニア、アドバイザー。画像データ処理パイプラインと機械学習を含むインテリジェンス製品等を開発。
B	官民セクターで 法的な観点からのサイバーセキュリティと AI の分野で長期の実務経験。民間部門では、AI 関連の情報収集とプライバシーの国際戦略の策定、公共部門では、透明性、プライバシー、インテリジェンスの実践に携わる。
C	長年にわたり米軍に勤務し、サイバーセキュリティ、インテリジェンス、テクノロジーの革新に貢献。官民セクターで AI の開発、管理、イノベーションに関連する役職を歴任。 金融業界での勤務経験あり。
D	偽情報に重点を置いたサイバーセキュリティ会社の CxO 経験者。米大統領選挙のロシア介入の調査にも携わる。
E	セキュリティ会社の CxO として顧客のデジタルおよびサイバーリスクマネジメントを支援。 金融機関の最高セキュリティ責任者やグループCISO、また金融機関で技術リスク管理のグローバル責任者など、数多くの上級職を歴任。
F	セキュリティ会社の CEO。Unix セキュリティの研究開発に従事した後、事業会社にて上級副社長兼 CISO。大学においてコンピュータ サイエンスの非常勤教授としての勤務経験あり。
G	民間企業のサイバー対策部門長。過去には米国国家安全保障局 (NSA) の高官として戦略、政策を監督。米国の国家安全保障戦略の実行と支援政策の策定において米国の国防・諜報事業を支援。
H	セキュリティおよび新興技術センターの戦略および基礎研究助成金のディレクター。民間企業の上級アナリストとして、AI の政策と戦略について政策立案者や助成金作成者にアドバイスを行った経験あり。
I	研究機関の研究員として、AI セキュリティ関連プロジェクトに従事。現在、ランサムウェアと強制効果のための暗号化の使用に関する一連のモデル構築や北朝鮮と東アジアの安全保障問題についても研究。
J	数多くの企業の共同創設者や会長であり、多くの公共、非営利、政府機関の顧問も務める。過去にはAIなどに関する国家安全保障局の顧問を務めた。米国の有力圧力団体のメンバーでもある。

エグゼクティブサマリ抜粋

- 生成型 AI ツールによりサイバー犯罪者の参入障壁が低くなり、より高度で個別化された攻撃の増加につながっている。**AI の民主化(利用の容易化)により、従来のサイバー脅威の加速と増幅が可能になり、防御側の能力を上回る可能性がある。**
- 生成型 AI ツールは、選挙干渉、偽情報、大規模監視のための国家主導のキャンペーン等で、偽情報の状況を一変させている。**ディープフェイクの蔓延は偽情報の常態化を悪化させ、時間の経過とともに制度や民主的なプロセスへの信頼を損なう。** 政府もSNSも、情報の自由な流れを維持しながら虚偽情報に対抗するという課題に直面している。
- 企業への生成 AI 導入が進むにつれ、偏ったコンテンツ・潜在的なデータポイズニング攻撃等のリスクに直面している。これらのリスクは、重要なシステムの信頼性と安全性に影響する。**誤った出力、不注意なエスカレーション、永続的なバイアス等の脅威を軽減するために AI ツールを保護する必要がある。**
- 報告書は、調査とインタビューから得られた一連の重要な発見を記載する。 AI の悪用による脅威、リスク、影響、タイムライン、主要な懸念事項を分析する**チャートを通じて、AI の脅威とリスクの分析も提供**する。

カテゴリ別の脅威・リスクとインパクトの整理

Threat	Risk	Impacted sector/entity/etc.	Timelinei	Impact
AI-enhanced traditional cyberattacks	Force multiplier for disruptive attacks	All sectors but critical infrastructure may be impacted greatly	Medium term	High
	Increased capabilities, sophistication, efficiency of cybercriminals in ransomware and cryptocurrency -related cyberattacks; lowered barrier to entry	Individuals and industries, especially ransomware-prone industries such as health care, financial, and hospitality sectors	Medium term	High
	Lowered barrier to entry for social engineering; increased efficiency and speed in spear phishing	Individuals, industries, governments, academia, news organizations, critical infrastructure	Immediate	High
AI-enabled disinformation	Domestic Disinformation: increased censorship, targeting of vulnerable groups, spread of authoritarian digital norms	Particularly individuals and minorities in authoritarian nations, democracy, freedom of speech	Immediate	Medium
	State-sponsored disinformation campaigns: polarization of societies, erosion of trust in institutions, degrading of democracy	Individuals, democratic governments, electoral process Democratic	Immediate	Medium
	Promotion of crime and discrimination: new class of crime such as deepfake pornography and stock market manipulation	Individuals, finance industry, black market, private sector widely	Medium term	Medium
	Election Obstruction: online censorship, disinformation	Individuals, freedom of speech, democratic nations, electoral process	Immediate	Medium-High
AI-Enabled disruption or maloperation of systems	Data poisoning: false outputs leading to bad decision-making, discrimination, disruption	Critical infrastructure, social infrastructure, justice system, others	Medium term	High
	Inherent biases and vulnerabilities: reinforce stereotypes, biased content generation and decision-making	Individuals, businesses, governments	Immediate	Medium
	Intentional and unintentional failures: operational disruption and false outputs	Critical infrastructure, social infrastructure, justice system, multiple industries	Immediate	Medium-High
AI-enabled national security threats	Military applications: potential autonomous weapon systems, military decision making leading to ethical concerns	Defense sector, governments	Long term	High
	AI race: deployment of AI systems with unproven reliability, risk of escalation	Governments, defense sector, industry	Long term	High
	Espionage and Mass Surveillance: higher scale and speed, erroneous uses by the private sector	Public and private sector, individuals, privacy	Medium term	Medium
	Terrorism: dissemination of propaganda, assist with terrorist plans	Social media companies, individuals, governments	Medium term	Low
	Bioterrorism: development of novel pathogens, efficient information gathering	Individuals, healthcare, and pharmaceutical sectors	Long term	Low
Business risks due to misuse of generative AI	Vulnerable code generation and dissemination (can be due to insufficient oversight and testing): data leakage, reputational damage, regulatory noncompliance, financial losses, operational disruption	Businesses, consumers, employees, privacy	Immediate	Medium
	Legal risks and insider threats: data leakage, trade secret theft, noncompliance, financial penalties	Legal system, privacy, businesses, individuals	Immediate	Medium

i Immediate refers to happening currently or in the next couple of years. Medium-term refers to the next 3-5 years. Long-term refers to the next 5-10 years.

‘Report for the Information Promotion Agency (IPA) – AI Risks and Threats - ’ , Prepared by Next Peakより抜粋

文献調査・インタビュー結果概要（１）

関連文献調査、インタビューから以下の事項が判明。

◘ 関連文献調査

- **生成AIはサイバー攻撃の戦術・技術・手順（TTP）を強化**する。今後**２年で既存TTPの進化や拡張による脅威**が生じる。具体的には文献※において、2025年に向けた脅威、とされる。

※National Cyber Security Centre, ‘The Near-Term Impact of AI on the Cyber Threat’.

- 将来、データ汚染攻撃は軍事・医療・自動運転等の重要インフラに用いられるAIに脅威を与える。

◘ インタビュー

- 現状で、**AIが引き起こす明らかな脅威・リスクは　ソーシャルエンジニアリングやフィッシング、偽情報である点は、専門家の共通認識である。**インタビューにおいて、**特定人物を模倣することは新たな大きなリスクであり、生成AIが音声や動画を用いたソーシャルエンジニアリングを拡張する**、とのコメントがあった。

文献調査・インタビュー結果概要（２）

関連文献調査、インタビューから以下の事項が判明。

□ インタビュー

- 大規模言語モデルによる最大の脅威として、安全でないコードの生成が商業的なリスクであるとする専門家がいる一方、把握できないマルウエアが生成される可能性であるとする専門家がおり、**見解は一致していない**。Amazonのような大企業では独自の強固なコード確認システムが機能している一方、中小やスタートアップにおいてそのような機能はなく、**安全でないコードがオープンソースのレポジトリに格納されAIにより学習された場合、AIが生成するコードに脅威が生じうる**、とのコメントがある。
- **米国の大統領令（EO14110）やEUのAI法といった国家レベルの取り組み**のみならず、**the Coalition for Content Provenance and Authenticity (C2PA) といった民間レベルにおいても自主的な取り組みが進められている**点は注目に値する。インタビューにおいて、**オーストラリアやインドにおいてもAI規制が策定されつつあり、ニュージーランドにおいてもワーキンググループにおいて検討が進められている**との意見や、**C2PAにおいて、Adobe社主体となり技術標準を策定することにより、偽情報や誤情報を特定するための取り組み**を行っている、との意見があった。
- **国家の支援を受けた機関がすでにAIを活用**している。

  オーストラリアは現在米国からの支援を受け、中国の活動家を想定しAIを活用した攻撃に向けた準備を進める。

今後の予定

- 本調査結果を踏まえ、今年度調査を行う（６月～１０月）予定。
  - AIセキュリティ脅威・リスクの深掘り
    - 重要インフラ（例えば、大統領選挙等を想定）のケーススタディ
  - AIガバナンスに関する米国の取組みと脅威・リスクに対する有効性確認
    - 大統領令14110（特にSec.4に記載されるもののうち、セキュリティに関連する事項）に対する履行状況とそれに対する業界・組織の反応

AISI関係機関向け勉強会

- **概要**　IPAセキュリティセンターが2023年度に実施した「米国におけるAIの脅威・リスクの認知調査」に基づき、文献調査や有識者インタビューにより示された、現状のAI利用における脅威やリスクについて整理し、得られた示唆を共有する。
- **対象**　AISI 関係機関の特にAIの悪用や誤用による脅威やリスクに関心がある方
- **実施方法**　Teamsによるオンライン。
  質疑応答はチャットと音声を用いた双方向で実施。
- **実施時期**　5月下旬