参考資料２

米国におけるAIの脅威の認知状況調査

2024年5月31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

©2024 独立行政法人情報処理推進機構（IPA）

米国脅威調査の概要

- AIサービスの普及が加速度的に進む中、AIのセキュリティ脅威がどう認知されているかを把握すべく、米国における**AIに関するセキュリティ脅威とその認知**に関する調査を実施
- 調査方法
  - ８０以上の関連文献調査
    - 重要文献10件を抽出
  - 様々な領域の専門家10名のインタビュー
    - AI・サイバーセキュリティ・行政・法務等
- 分析内容
  - セキュリティ脅威の5つの類型における脅威と認知状況の実態
  - 各分野のリスクの大きさ（時期・影響度）に関するマップ
- 調査期間
  - ２０２４年1月～２月
- 公開情報

IPAテクニカルウォッチ｜情報セキュリティ｜IPA 独立行政法人 情報処理推進機構
https://www.ipa.go.jp/security/reports/technicalwatch/index.html

©2024 独立行政法人情報処理推進機構（IPA）

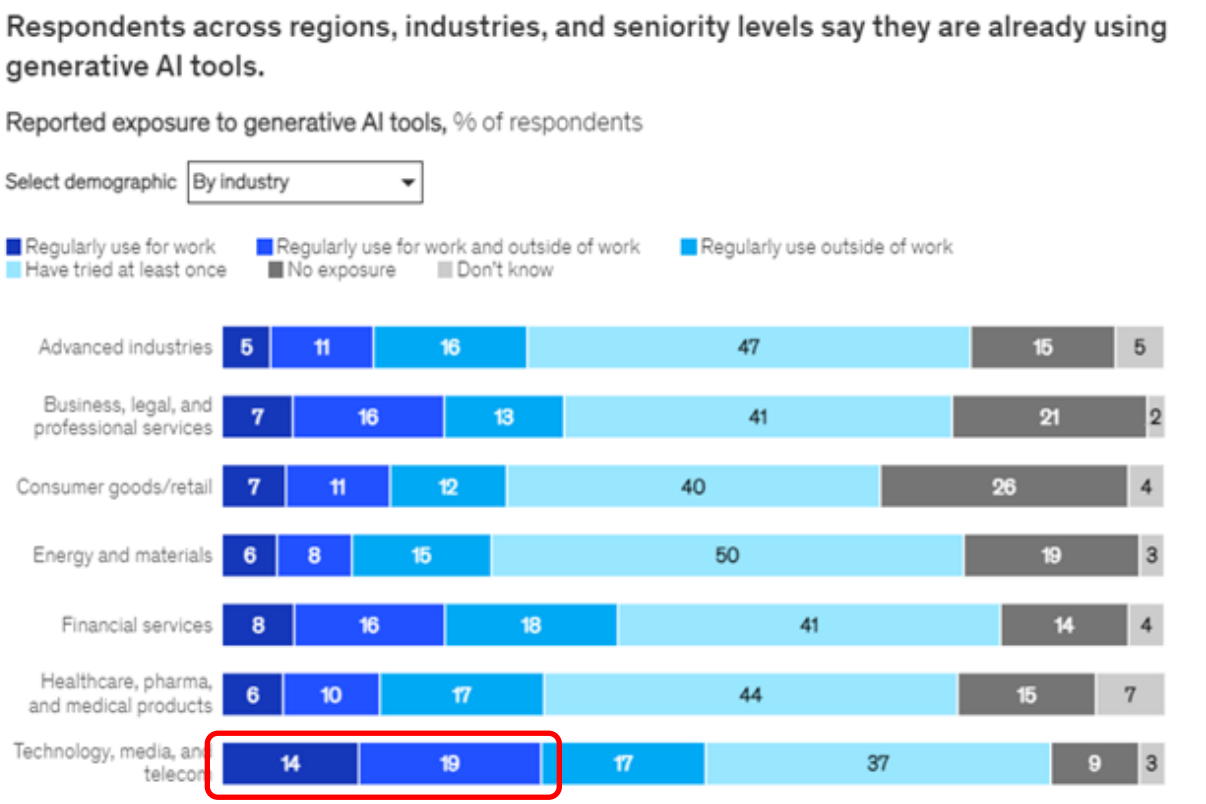
米国脅威調査の詳細

1. AI利用の実態
2. AIで強化された従来のサイバー攻撃
3. AIを利用した虚偽情報
4. AIによるシステム障害とAIシステムへの攻撃
5. AIによる国家安全保障上のリスク
6. 生成AIの誤用によるビジネスリスク

©2024 独立行政法人情報処理推進機構（IPA）

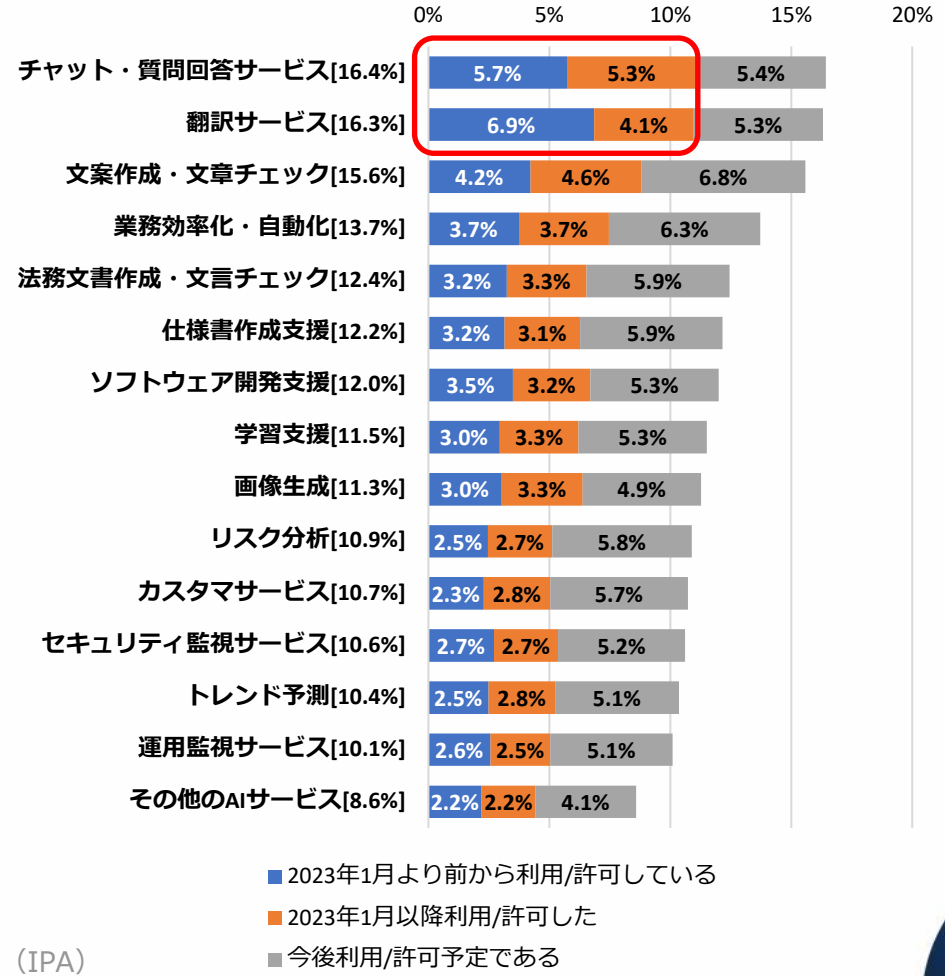
1. AI利用の実態

米国：業種別生成AI利用状況（報告書に記載）
2023 McKinsey Global survey の図にIPA追記



赤枠:業務で定常的に利用

【参考】国内：応用別AI業務利用状況
2024 IPAアンケート調査（2024年度上期公開予定）
大企業・中小企業　IT実務者4941人



©2024 独立行政法人情報処理推進機構（IPA）

2. AIで強化された従来のサイバー攻撃

（〇ページ参照）は
報告書本文の記載ページ

1. 攻撃の傾向

AIにより既存攻撃を素早く、強力かつ効率的に行う傾向がみられる（15ページ参照）

AIを用いた新たな攻撃手法はまだ

⇒ インタビューでは、しろうとがAI利用で高度化、とプロがAI利用でさらに高度化、の脅威について両論あり

喫緊の最大課題はフィッシングへの悪用

2. 事例

大きな事例はないが、AIフィッシングの急増・高度化、組織へのAIによる執拗な攻撃は観測されている

（悪意のフィッシングメールは2022年末と比較して1,265%増加） （19ページ参照）

国家支援アクター等の高度な攻撃者のAI利用もみられる（9ページ参照）

3. リスクの大きさ・認知状況

フィッシングについてはリスク大と認識

高度な攻撃・マルウェア自動生成については脅威大となるのに時間がある、と認識

©2024 独立行政法人情報処理推進機構（IPA）

3.　AIを利用した虚偽情報

（〇ページ参照）は
報告書本文の記載ページ

1.　攻撃の傾向

海外国家支援アクターからの攻撃が懸念大(AIにより言語・文化等の壁がなくなる）　（16ページ以降参照）
⇒　だますのではなく、フェイクを信じる勢力を作って国内でけんかさせる
悪意の生成AIモデルは2021年から出現（WormGPT、FraudGPT・・）　（13ページ以降参照）
民間ではオープンソースAIの普及でフェイクが容易に作れる（23ページ参照）

2.　事例

有名人のフェイク画像による中傷は当たり前になった（米国：Taylor Swift等）　（23ページ参照）
フェイクの事故画像により株価が乱高下する事態がみられた（米国）　（23ページ参照）
選挙候補の中傷・選挙妨害が生成AIで行われた可能性がある（台湾総統選挙）　（24ページ参照）

3.　リスクの大きさ・認知状況

米国では特に選挙妨害への懸念が大きい
インタビューでは、脅威が非常に大（米国を分断する、等）、とそこまでではないという意見の両論あり

©2024 独立行政法人情報処理推進機構（IPA）

4.　AIによるシステム障害とAIシステムへの攻撃

（〇ページ参照）は
報告書本文の記載ページ

1.　攻撃の傾向

最大懸念はデータポイズニング（26ページ以降参照）

⇒　学習データにノイズや特殊なパターンをいれて性能劣化や意図的な誤作動をおこす

特に懸念される重要インフラセクターは、インタビューで特定されなかった

生成AIへの悪意のプロンプトによる攻撃は新しい課題（66ページ参照）

2.　事例

偏った学習データに基づく採用プロセスで男性へと評価が偏重した(米国)(28ページ参照)

AIシステム攻撃手法に関する研究は進んでいるが、実際の攻撃は未確認

3.　リスクの大きさ・認知状況

データポイズニングの脅威はまだ認知度が小さい模様

大規模な学習データやオープンソースAIが公平か、汚染されていないかの検証は難課題

データポイズニングは内部不正があると大きな脅威になるか、に対しては賛否両論あり

悪意のプロンプトによる情報漏えい・不適切回答のリスクはまだよく見えない

©2024 独立行政法人情報処理推進機構（IPA）

5.　AIによる国家安全保障上のリスク

（〇ページ参照）は
報告書本文の記載ページ

1.　攻撃の傾向

軍用AIの開発・利用は米・中・欧等で進んでいる（29ページ参照）

　米国は軍での生成AI活用・意思決定支援用基盤モデル評価等を開始

米中のAI軍備競争が始まっている（30ページ参照）

米国は　中国が米国国民の諜報・自国民の大規模監視にAIを使っていると懸念

テロリスト集団のAI利用は生成AIによる発信が中心。
ただし、バイオテロへのAI利用は要警戒（34ページ以降参照）

2.　事例

AIベンダーがSNS等で収集した30億人分の顔情報を元に顔識別システムを警察に納入したが
カナダ当局よりプライバシー侵害を指摘された（35ページ参照）

3.　リスクの大きさ・認知状況

AI軍備競争の結果、テストが十分なされないままAIが実戦に投入されるリスクあり

©2024 独立行政法人情報処理推進機構（IPA）

6.　生成AIの誤用によるビジネスリスク

1. 誤用の傾向

営業秘密情報・個人情報をうっかりプロンプト入力し、意図せず学習され漏えいするリスクは周知されてきた
ソフトウェア開発において、学習データによって生成AIが脆弱なコードを出力するリスクがあるが、事例は未確認
米国マッキンゼー調査によれば、生成AI利用社内規則がある、としたのは回答者の21%。
⇒ 個人情報や営業秘密の入力、誤りを含む出力のチェック不備等のリスクあり

2. 事例

開発者がAPIキーがついたソースコードの一部をChatGPTに入力した（米国） （34ページ参照）
調査では、AIを使う開発者はセキュアでないコードをAIがセキュアにすると思いがちだった（米国）
生成AIを用いてコードを開発し、結果的にChatGPTに営業秘密を入力した（韓国）

3. リスクの大きさ・認知状況

調査によれば、生成AI入力の過半数に営業秘密・個人情報が含まれていた（米国）
⇒ 上記の利用規則がある、の回答率と単純比較すると意識がゆるい
生成AIのテストがどの程度できているかが不透明
⇒ このような状況に対して利用者は信用しすぎているかもしれない

©2024 独立行政法人情報処理推進機構（IPA）

米国脅威調査のまとめ

AIのセキュリティ脅威を以下の５分野につき調査

1.　AIで強化された従来のサイバー攻撃

フィッシングは生成AIにより増化・高度化しており、現時点で大きな脅威

特定組織をAIで執拗に攻撃することが観測されている

2.　AIを利用した虚偽情報

偽情報の脅威は非常に大、そう深刻ではない、の両論がある

選挙等では深刻な脅威

3.　AIによるシステム障害とAIシステムへの攻撃

顔認証の誤判断による不当逮捕がシステム障害として懸念された

AIシステムへの攻撃ではデータポイズニングによる性能劣化・誤判定が最大懸念

4.　AIによる国家安全保障上のリスク

軍備競争によりテスト不十分なAIが戦闘で利用されるリスクあり

5.　生成AIの誤用によるビジネスリスク

生成AIのセキュリティを信用しすぎる傾向があり、対応は不十分

©2024 独立行政法人情報処理推進機構（IPA）

米国脅威調査のまとめ：類型別のインパクト

脅威	Risk リスク	影響をうける主体 etc.	時期 1]	影響
AI-enhanced traditional cyberattacks	Force multiplier for disruptive attacks	All sectors but critical infrastructure may be impacted greatly	Medium-term	High
2 AIで強化された従来のサイバー攻撃	Increased capabilities, sophistication, and 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
	L フィッシングの容易化・高速化 social e ency and speed in spear phishing	Individuals, industries, governments, academia, news organizations, critical infrastructure	Immediate	High
AI-enabled disinformation	Domestic Disinformation: increased ce 国内の虚偽情報言動監視 ulnerable gr arian digital norms	Particularly individuals and minorities in authoritarian nations, democracy, freedom of speech	Immediate	Medium
3 AIを利用した虚偽情報	Sta 国家支援虚偽情報キャンペーン ation car 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
	Ele 選挙妨害 ıe censorship, disinformation	Individuals, freedom of speech, democratic nations, electoral process	Immediate	Medium-High

※Immediate: 現時点～2年以内、Medium term：～5年以内、Long term：～10年以内

AI-Enabled disruption or maloperation of s 4 AIによるシステム障害とAIシステムへの攻撃	Data poisoning: false outputs leading to bac nation, dis データポイズニング	Critical infrastructure, social infrastructure, justice system, others	Medium term	High
	Inherent biases and vulnerabilities: rei content ger データバイアスによる意思決定妨害 ng	Individuals, businesses, governments	Immediate	Medium
	Inte failures: ope se out 不正出力による誤判断・誤作動	Critical infrastructure, social infrastructure, justice system, multiple industries	Immediate	Medium-High
AI-enabled national security threats 5 AIによる国家安全保障上のリスク	Military applications: potential aut , military dec 自動兵器・戦闘意思決定支援 nical concerns	Defense sector, governments	Long term	High
	AI ith un テスト不十分な軍用導入 n	Governments, defense sector, industry	Long term	High
	Espionage and Mass Surveillance: hig eous uses by the private sector 諜報・大規模監視	Public and private sector, individuals, privacy	Medium term	Medium
	Ter ans pro テロリズムプロパガンダ	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 6 生成AI誤用によるビジネスリスク	Vul nd dis ins ing): data lea , reg 脆弱なコード流通 非倫理的・不正コンテンツ生成・レピュテーションリスク ıancial losses, operational disruption	Businesses, consumers, employees, privacy	Immediate	Medium
	Lega data leak 法的リスク 営業秘密漏えい non lties	Legal system, privacy, businesses, individuals	Immediate	Medium

©2024 独立行政法人情報処理推進機構（IPA）

【参考】 文献調査・インタビュー結果概要

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。
- 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。**
- 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。**
- **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 Content Authenticity Initiative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。
- **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と連携し、敵対勢力の活動家を想定しAIを活用した攻撃に備えている。

©2024 独立行政法人情報処理推進機構（IPA）

【参考】重要調査文献（一部抜粋）

生成AI普及以降の最新動向を含む文献を調査

- 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.

©2024 独立行政法人情報処理推進機構（IPA）

DisplayOnly

2024年度のAIセキュリティ脅威の調査（案）

本スライドの内容は調査報告書には含まれておりません

- 米国調査２
  - 選挙に対するAIリスクの詳細な分析
    - 選挙インフラとプロセスに対しAIが新たに及ぼすサイバー脅威
    - 選挙のセキュリティ策補におけるAI支援サイバーセキュリティの有効性
  - 米国におけるAIガバナンス・フレームワークの詳細な分析
    - AI大統領令（EO14110）、NIST AI-RMF、NIST SSDF他
    - AI大統領令における各府省庁の役割と相互関係（セキュリティの観点から）
    - 各種取り組みの有効性
- その他
  - 欧州・英国調査（AI法とセキュリティの関係把握・分析等）
  - 国内有識者会合（我が国におけるAIとセキュリティの状況・見通し把握等）

©2024 独立行政法人情報処理推進機構（IPA）

©2024 独立行政法人情報処理推進機構（IPA）