「CSAが取り組む生成AIのセキュリティ」
CSA Japan Congress 2023

一般社団法人　日本クラウドセキュリティアライアンス

理事　諸角昌宏

CSAリサーチフェロー、 CCSP、CCSK、CCAK

2023年11月22日

アジェンダ

1. AIのセキュリティ状況
2. 「Security Implications of ChatGPT」（日本語版：ChatGPTのセキュリティへの影響）の解説
3. CSAが取り組む生成AI利用ポリシー（Usage Policy）について
4. CSAジャパン　AIワーキンググループ方針

AIのセキュリティ状況

今までのAIセキュリティ

- 今までのAI
  - Predictable AI （予測分析）：データの統計分析から何らかの予測を行う
    - 大量のデータから「特徴」を学んで「予測」すること
    - AIが特徴をつかむのに必要な大量で十分な質のデータが存在することが性能の良いAIをつくる前提
      データが正しければ正しい結果が得られる。データの品質が重要
  - 特徴
    - 大量、多様のデータ、高速処理（3つのV：Volume, Variety, Velocity)
    - 分散処理

- AIセキュリティ
  - ビッグデータセキュリティ（クラウド関連技術）
    - インフラストラクチャセキュリティー。分散処理のプラットフォーム
    - データプライバシー
    - データの完全性

AIのセキュリティ状況

- 生成AIの登場
  - コンテンツやモノについてデータから学習し、それを使用して創造的かつ現実的な、まったく新しいアウトプットを生み出す機械学習手法
  - 今までのAIでは難しかったクリエイティブな領域での活用
  - AI活用の民主化の実現：　誰でもAIを活用できる
  - 人間だけでは実現できないこと、AIだけでも実現できないこと。それを理解し、相互に活用する、人間とAIが協調できる社会が目指される可能性

- 生成AIの新しいセキュリティの考慮事項
  - AIそのものへの攻撃
  - トレーニング・データのセキュリティ
  - 生成物の法的、規制的問題
  - 倫理的な利用、フェイクへの対応

「Security Implications of ChatGPT」（日本語版：ChatGPTのセキュリティへの影響）の解説

- CSAが最初に公開した生成AIのセキュリティ
- ハイレベルでChatGPTの意味、概要を説明することを目的
  - ChatGPTの能力を理解するために重要な鍵となる概念と領域
  - ビジネスへの潜在的な影響について探求
- 4つのポイント
  - How it can benefit malicious attackers 攻撃者にとっての利点
  - How it can benefit cybersecurity　守る側にとっての利点
  - How ChatGPT might be attacked directly　ChatGPTそのものへの攻撃
  - Guidelines for responsible usage　安全にビジネス利用するガイド

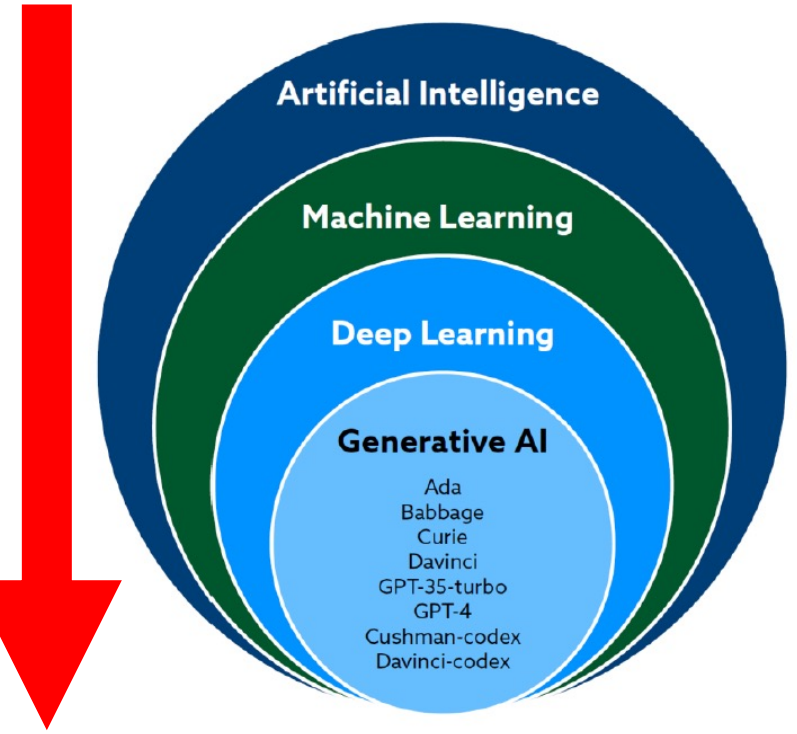
  - ChatGPTは、サイバーセキュリティの攻撃者、保護者の双方に今まで以上の技術を提供

ハイレベルなChatGPTの意味、概要

AIの歴史、ChatGPT

1. Artificial Intelligence
   1950年代、Computing Machinery and Intelligence
   チューリングテスト「テスト対象の機械が人間と似た（あるいは近い）行動を取れるのか」
2. Machine Learning
   1997年、チェスのグランドマスターがIBMのDeep Blueに敗れた
3. Deep Learning
   2016年、囲碁のディープラーニングアルゴリズム「Alpha Go」が、韓国のプロ棋士に勝利
4. Generative AI
   2020年代、生成AIの時代
   - プロンプト＆リプライ形式
   - ニューラルネットワークベースの音声認識
   - シームレスに統合するためのAPI

   ChatGPTがチューリングテストを乗り越えた!?



サブドメインによって特徴付け

ハイレベルなChatGPTの意味、概要

- ChatGPTでは、OpenAIの3つのモデルファミリーを採用
  - GPT-X
    自然言語を理解し生成するために設計された一連のモデル
    テキストベースの情報を処理し、一貫した回答を生成
  - Codex
    コードの理解と生成に特化したモデル
    自然言語をコンピュータのプログラミング言語に翻訳することができ、指示されたソースコードを効率的に生成することが可能
  - Embeddings
    テキスト検索、類似検索、コード検索など、特殊な機能に特化したモデル
    様々な文脈でより効率的な情報検索・処理を可能

ハイレベルなChatGPTの意味、概要

- ChatGPTの制限事項
  - AI言語モデルが有害または違法な活動を促進または助長することを禁止するOpenAIのユースケースポリシー
    - 悪意のあるコンテンツや違法なコンテンツの生成を防ぐためのセーフガード
  - BUT…
    - 制限を回避して不正確な結果やその他の望ましくない結果を生み出すことが可能

- マイクロソフトの「責任ある AI の基本原則」（Responsible AI）
  - 公平性
  - 信頼性と安全性
  - プライバシーとセキュリティ
  - 包括性
  - 透明性
  - 説明責任、アカウンタビリティ

ハイレベルなChatGPTの意味、概要

- ChatGPTの主な利用規約
  - 利用制限
    - 法令に違反する行為、知的財産権を侵害する行為、有害な内容や差別的な内容を助長する行為を行ってはならない
  - APIアクセス
    - 特定の条件下でそのAPIへのアクセスを許可することがあります。ユーザーはこれらの条件を遵守しなければならない
  - ユーザーデータ
    - データ利用ポリシーにおいて、ユーザーデータの収集、保存、利用方法について詳述
    - ユーザーは、本ポリシーに記載されているデータの取り扱いについて同意

OpenAIのサービスを利用する前に、OpenAIの利用規約を十分に読み理解することが必要！

ハイレベルなChatGPTの意味、概要

- ベンチマーク、ツール
  - ChatGPTのようなLLMの性能を比較するベンチマーク
    - CRFM（Comprehensive Real-World Fine-tuning Model）ベンチマーク
  - LLMに使用されるセキュリティ・ツール
    - 侵入検知・防御システム、SIEM、MFA、脆弱性スキャナと侵入テストツール、ブロックチェーン技術など
    - セキュリティの強化に利用

クラウドテクノロジーの猛威以来のテクノロジーシフトを目の当たりにしている！

「Security Implications of ChatGPT」（日本語版：ChatGPTのセキュリティへの影響）の解説

- How it can benefit malicious attackers 攻撃者にとっての利点
  - 「新しい」ハッキングツールの開発・普及が懸念
  - 悪意のあるアクターがツールセットを強化することに伴う潜在的なリスク
  - 様々なサイバー攻撃のステージで悪用される可能性
    - サイバーキルチェーンの各ステージ
  - フィッシング
  - AI技術の急速な進歩により、脅威アクターの能力は大幅に向上し、本物の連絡と酷似した欺瞞的なメールを作成することが可能
  - ポリモルフィック・シェルコードの生成に利用

「Security Implications of ChatGPT」（日本語版：ChatGPTのセキュリティへの影響）の解説

- How it can benefit cybersecurity　守る側にとっての利点
  - アプリケーションのセキュリティ強化
    - 脆弱性対策の強化（GitHub Copilotの強化など）
    - セキュリティ機能のソースコードへの作り込み（認証情報対応、他のセキュリティ例の取り込みなど）
  - 脆弱性スキャンそのものの強化（様々なコードパターンからスキャナーを開発)
  - 生成AIが生成したテキストを検出するための機能（電子透かし）
    - フィッシング対策
  - SIEM/SOAとの統合によりインシデント管理を加速
  - ソースコード/設定ファイル、セキュリティパッチ等の分かりやすい解説の作成
  - ファジングの強化
  - Etc.

「Security Implications of ChatGPT」（日本語版：ChatGPTのセキュリティへの影響）の解説

- How ChatGPT might be attacked directly　ChatGPTそのものへの攻撃
  - 攻撃
    - ChatGPTへの攻撃
    - ChatGPTのガードレールを回避
    - プロンプト・インジェクション
    - 悪意のあるアプリケーション
  - 対策
    - プロンプトを「命令」部分と「データ」部分に分割し、プロンプトインジェクション攻撃を制限
    - ユーザー教育、厳格なセキュリティ対策、効果的な規制や政策を策定するためのステークホルダーとの協力など、多面的なアプローチが必要
    - ユーザーは、ChatGPTに接続するために使用しているアプリケーションやサービスの信頼性を確認し、HTTPSや安全なAPIアクセスなど、安全な通信チャネルを使用していることを確認することが極めて重要
    - ユーザーは安全な接続を確保し、エンドツーエンド暗号化の使用や信頼できる通信チャネルを採用するなど、 チャットセッションの整合性を維持
    - 機密情報を入力しない

「Security Implications of ChatGPT」（日本語版：ChatGPTのセキュリティへの影響）の解説

- Guidelines for responsible usage　安全にビジネス利用するガイド
  - 明確な使用方針の策定
  - アクセス制御の実施：　許可された担当者のみに限定
  - 通信経路の確保：　完全性、暗号化（中間者攻撃）
  - 使用状況の監視・監査：　疑わしい活動や潜在的な不正使用を検出
  - 従業員教育
    - セキュリティ上の懸念事項の報告を奨励
    - AIセキュリティの最新情報を提供

「Security Implications of ChatGPT」（日本語版：ChatGPTのセキュリティへの影響）の解説

- 「Security Implications of ChatGPT」でカバーされていない領域
  - Attack by AI（AI自身による攻撃）

    「引用：　AIとセキュリティDXの進展に向けた四つの観点とアプローチ（https://www.hitachihyoron.com/jp/archive/2020s/2021/06/trends/index.html）」
    - 人間を上回る能力を有するAIが誕生し，将来的に人間が絶滅させられるのではないかという問題
    - 現在のAI研究は「汎用AI」ではなく「専用AI」の研究が中心であり，多くのAI研究者は人間に対する反乱が起きる可能性は無視できるほど低いと考えている
    - 専用AIでもAI兵器などの分野で反乱が起きると取り返しのつかないことになる可能性が強い
    - AIが異常行動を起こしたとき，正しく検知・停止できるかどうかの実験などが大切

生成AIに対するCSAのポジション

- *ジェネレーティブAIは可能性に満ちた強力な技術だ。誰かがズルをして競争優位に立とうとしたり、少なくとも追いつこうとしたりするからだ。また、印象的な技術だとは思うが、『マトリックス』のような世界征服的な技術ではない。私の解決策は？今日のAIの使い方を規定するベストプラクティスの開発に懸命に取り組み、AIの開発者がその経験を次世代の改善に役立てよう。良いAIの子どもを育てれば、責任感のあるAIの大人が生まれるかもしれない。（by CEO）*

- 生成AIに関する最初のドキュメント：Security Implications of ChatGPT
  - ハイレベルでChatGPTの意味、概要を説明することを目的。ChatGPTの能力を理解するために重要な鍵となる概念と領域、そしてビジネスへの潜在的な影響について探求
- AI Working Group
  - AIサービスの利用者と提供者の双方にとっての、AI利用の懸念と新たな現実を探っていく。
  - 最初に、ポリシーを作成するためのガイダンスドキュメントをリリースし、次に、組織が特定のニーズに基づいてAI利用ポリシーを作成し、実施するのを支援するサービスを作成し、リリースする。
  - それを、AIが進化し続けるにつれて、文書とツールの反復と更新を続けていく。

生成AI利用ポリシー（Usage Policy）

Usage Policyでカバーする予定の領域

1. 法律・規制の考慮点
2. 倫理的利用、公平性
3. 知的財産権、Copyright
4. 生成AIコンテンツのガバナンス
5. ディープフェイク
6. AIコンテンツの検出（AI作成なのか人の作成なのか）
7. 生成AIのための保険（Copyright侵害など）
8. AIハルミネーションからの保護
9. AIと既存システムとの統合
10. トレーニングデータ管理
11. モデル管理
12. モデルの透明性と出所
13. AIアラインメントとポリシー
14. AIガードレイル
15. AIによる複写、翻訳等のエラー
16. AIトレーニングデータ、モデル、アウトプットの第三者レビューあるいは人によるレビュー
17. 国によるAIに対する奮闘、パーソナライズAI

乞うご期待！

CSA本部のさらなる活動

- AIサブグループの活動
  - AI Technology & Risk
    - AIにおける最新の技術的進歩に遅れを取らないようにすると同時に、関連するリスク、脅威、脆弱性を特定、理解、予測
    - AI領域におけるイノベーションとセキュリティのギャップを埋める、知識ハブとプロアクティブなリスクマネジメントの両方の役割
  - AI Governance & Compliance
    - AIのガバナンスとコンプライアンスの基準を確立、提唱、普及。
    - ポリシーを作成、法律への取り組み、ベンチマークを作成
  - AI Controls
    - AIシステムの包括的なセキュリティ管理策の定義と実施
    - NIST CSFに準拠
  - AI Organizational Responsibilities
    - AI技術による新たな課題と機会に特に適応したセキュリティチーム内の役割と責任の定義

乞うご期待！

CSAジャパン　AIワーキンググループ方針

- 設立趣旨
  - CSA本部の取り組みを継承し、日本市場におけるA I規制やインフラ確立に対してAIの使い方のベストプラクティスを提供
- 活動方針
  - CSA本部資料に基づくスタディ
    - Security Implications of ChatGPTィ
    - CSA Generative AI Usage Policy（ 現在draft）
  - その他の規制/規格のスタディ
    - EU-AI actの理解と生成AIに向けての拡張の理解
    - OWASP Top10 for LLM
    - IEEE Position Statement Artificial Intelligence
  - 国内状況のスタディ
    - 国内法と規制の動向調査
    - 業界法と規制の調査（医師法、弁護士法、著作権法、個人情報保護法）
    - 全省庁からガイドライン
  - プライバシーWGとの連携

CSAジャパン　AIワーキンググループ方針

- 今後の取り組み予定
  - 政府AI戦略会議の提示課題を検討（サイバー攻撃、偽情報、知財リスク）
    - 職業資格と試験の検討（マイナビDX、デジタルスキル標準、ITパスポート試験）
    - AIの敵対的テストと標的サンプルの検討
    - AIの法的能力の検討（会話能力、評価能力、判断能力、意思決定）
    - AIの安全性評価と信頼性評価の検討
    - データの偏りとAI学習による破壊的忘却
  - AI脅威の検討
    - 脅威1（ＡＩモデルと学習データの非透明性）
    - 脅威2（偽情報とデジタルコンテンツのｅ公証性）
    - 脅威3（ハルネーションの事実らしさの不正確性）
    - 脅威4（脅威インテリジェンスの共有性）

最後に

4つの提言

1. AIは良いことにも悪いことにも使うことができる。

2. AIを含むデジタル社会におけるトラストは「証明できるトラスト」であり、「暗黙のトラスト」ではない。

3. AIによるディスラプション（disruption）を恐れてはいけない。変化に対応できる組織／人になる。

4. AIに対する法律／規制には特に注意を払うことが重要である。

CSAの活動　==　「場」の提供！

様々なワーキンググループ活動の「場」

自由な情報発信の「場」

https://cloudsecurityalliance.jp

info@cloudsecurityalliance.jp

ありがとうございました