AI時代の備え
– デジタル環境での自衛と批判的思考力 –

2024年2月21日

桑名栄二, CISSP
情報セキュリティ大学院大学

情報セキュリティ大学院大学の概要

明日の信頼を創ろう。
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

［設置法人］学校法人岩崎学園(理事長 岩崎文裕)、https://www.iwasaki.ac.jp/
［名称］情報セキュリティ大学院大学、https://www.iisec.ac.jp/
［学長］後藤厚宏
［住所］横浜市神奈川区鶴屋町2-14-1
［開学］2004年4月1日
［構成］研究科：情報セキュリティ研究科、課程：博士課程［前期・後期］

本日の内容

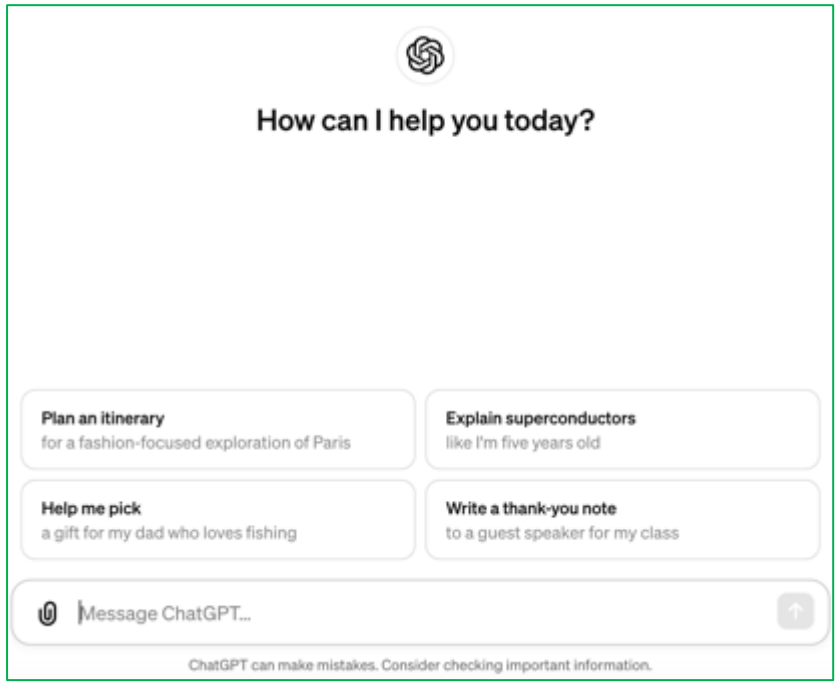
1. 生成AIと課題
   - 大規模言語モデル(LLM)
   - 大規模言語モデル(LLM)利用におけるリスク・課題
2. フェイクニュース、誤情報・偽情報、デュープフェイク
3. デュープフェイクの悪用と事案
4. ネットメディアの特性と人間の認知特性
5. 批判的思考スキルを身につける

参考)各国政府や法による規制の動向

ほか

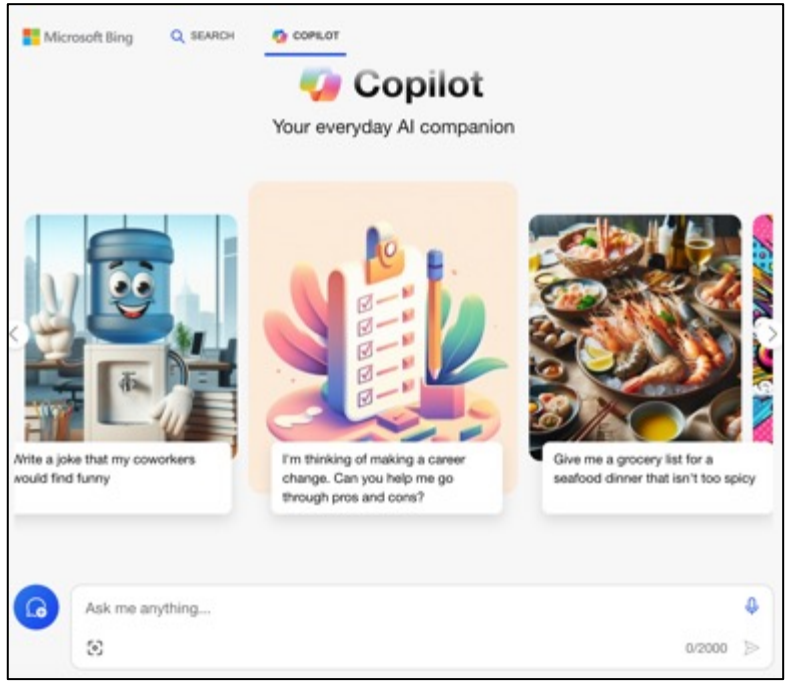
1. 生成AIと課題
2. フェイクニュース、誤情報・偽情報、デュープフェイク
3. デュープフェイクの悪用と事案

身近になった生成AIサービス



https://chat.openai.com/

https://gemini.google.com/app



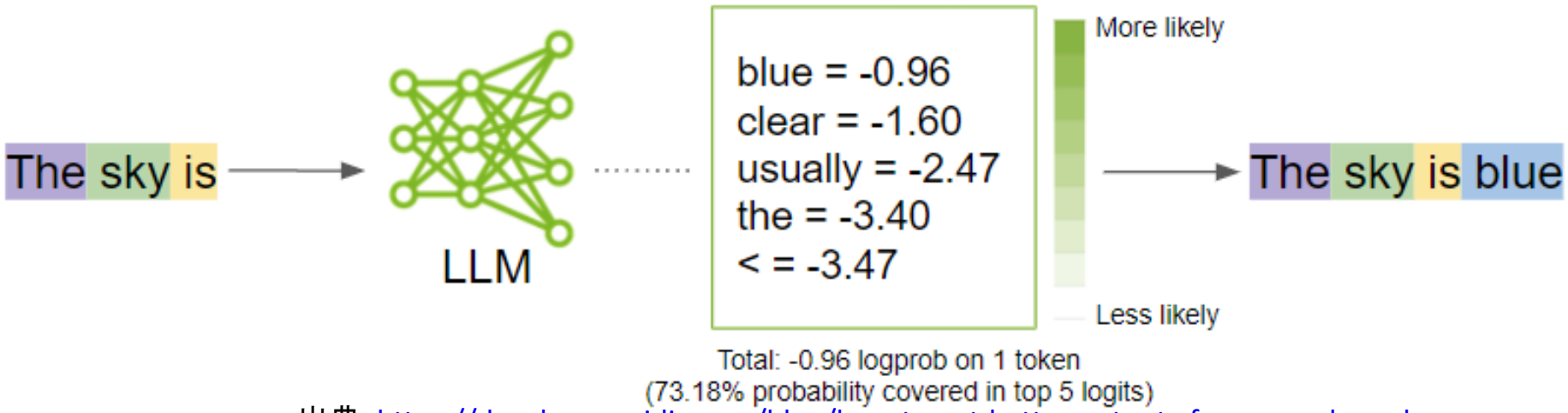
https://copilot.microsoft.com/

生成AI(Generative AI)は、利用者の入力(プロンプト)に応じて、学習データをもとに画像・音声・テキスト・プログラム等の新しいデータを作り出す人工知能(AI)技術の総称

大規模言語モデル(LLM : Large Language Models)

- テキストなどの連続したデータを処理、生成できるAIモデル。
- 大量の文章データからある単語の文章がどれだけ発生しやすいか、出現確率$P(w_t \mid w_{t-k}, \ldots, w_{t-1})$ を学習したもの

$P(w_4 \mid The, Sky, is)$の
確率が一番高くなる単語もしくは単語の一部(トークン)w_4は
clear, usually, theなどではなくblueなので、
The sky is blueを生成



出典:https://developer.nvidia.com/blog/how-to-get-better-outputs-from-your-large-language-model/

大規模言語モデル(LLM)の進化

- 2018年OpenAIがGPT(Generative Pretrained Transformer)を発表。トランスフォーマー(文章生成変換器)
- 2018年GoogleがBERTを発表
- 2022年11月30日OpenAIがChatGPTを公開
- 2023年2月Microsoft Bingに生成AI搭載
- 2023年2月 GoogleがBardを公開
- 2023年2月メタがLLaMA提供
- 2023年GPT-4、Copilot
- 2024年Google BardはGeminiに改称
- 2024年大規模言語モデルはテキストだけでなく、画像、動画、音声等も処理できるマルチモーダル型に進化

(参考情報)進化の過程：https://github.com/Mooler0410/LLMsPracticalGuide

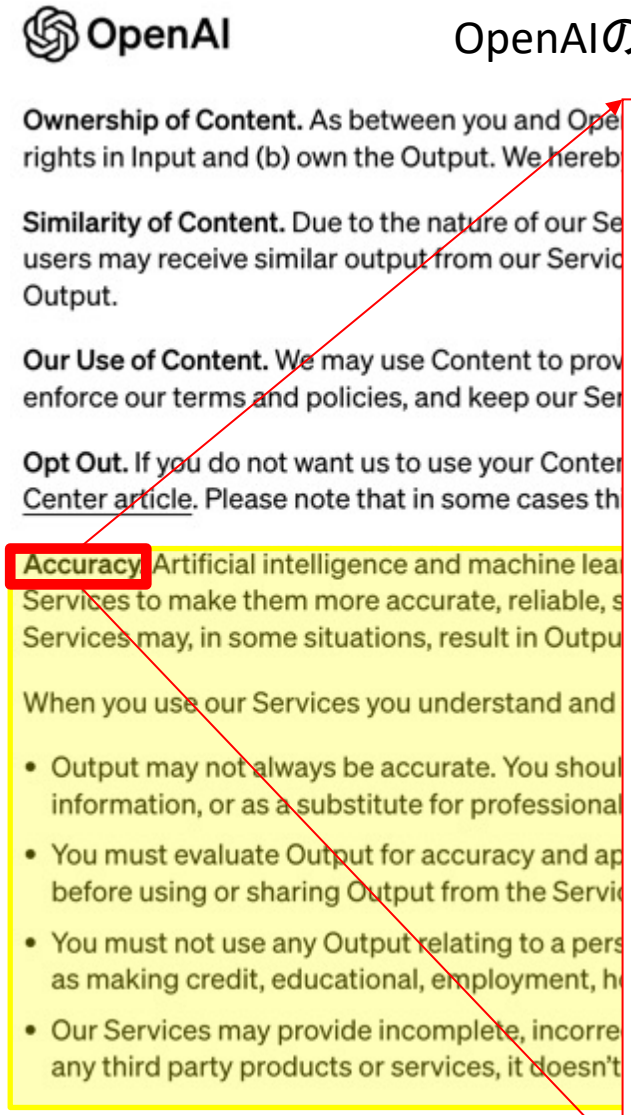
大規模言語モデル(LLM)利用に伴うリスク・課題

1. 情報の正確さ(信憑性)、
   - 出力結果は必ずしも正確であるとは限らない、
   - ハルシネーション (hallucination、幻覚):事実に基づかない情報を生成すること
2. 事実に基づかない情報の拡散、
3. プロンプト(AIへの入力データ・指示等のこと)に機密情報が含まれていた場合、それがデータとして流出するリスク、
   - 営業情報、設計データ、国家の機密情報、
   - 個人の悩み、相談ごとを含む個人情報など
4. AIの学習データに第三者の著作物が混入するリスク、

など

大規模言語モデル(LLM)の課題(ハルシネーション)

OpenAIの利用規約(https://openai.com/policies/terms-of-use(2023年11月14日版))から抜粋



OpenAI

Ownership of Content. As between you and Ope
rights in Input and (b) own the Output. We hereb

Similarity of Content. Due to the nature of our Se
users may receive similar output from our Servic
Output.

Our Use of Content. We may use Content to prov
enforce our terms and policies, and keep our Ser

Opt Out. If you do not want us to use your Conter
Center article. Please note that in some cases th

Accuracy. Artificial intelligence and machine lea
Services to make them more accurate, reliable, s
Services may, in some situations, result in Outpu

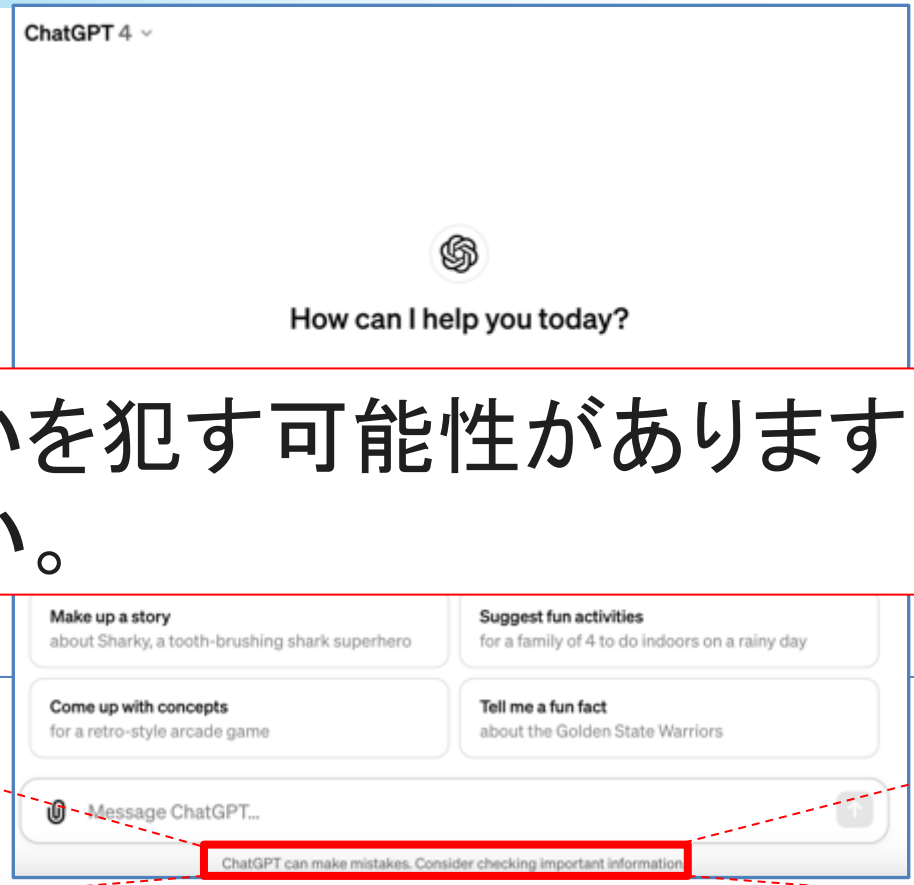
When you use our Services you understand and

- Output may not always be accurate. You shoul
information, or as a substitute for professional
- You must evaluate Output for accuracy and ap
before using or sharing Output from the Servic
- You must not use any Output relating to a pers
as making credit, educational, employment, h
- Our Services may provide incomplete, incorre
any third party products or services, it doesn't

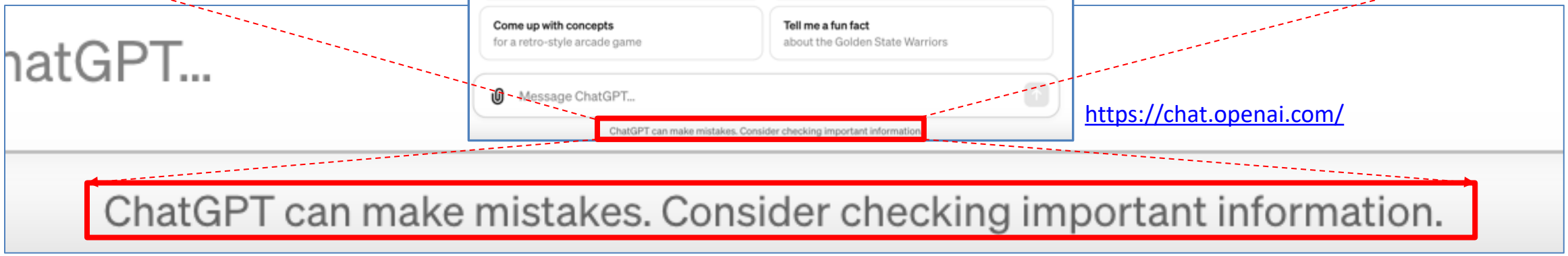
【正確さ】

- 出力は必ずしも正確であるとは限らない、
- 出力を使用または共有する前に、人によるレビューなどを行い、出力の正確性とユースケースに対する適切性を評価する必要がある、
- 個人に法的または重大な影響を与える可能性のある目的(例えば、信用、法律、医療など)に使用してはならないなど

ChatGPT can make mistakes…



ChatGPTは間違いを犯す可能性があります。重要な情報は必ず確認してください。



https://chat.openai.com/

https://chat.openai.com/

大規模言語モデル(ChatGPT)に関連する事案

- ChatGPTのセキュリティ上の脆弱性から情報漏洩の可能性(2023年3月)
  - 出典：https://openai.com/blog/march-20-chatgpt-outage
- 社員がChatGPTに機微なデータをアップロード(2023年5月)
  - 韓国サムスン社の社員がChatGPTにソースコードをアップロードしたことが発覚。同社は生成AIの利用を原則禁止とした
  - 出典：https://www.bloomberg.co.jp/news/articles/2023-05-02/RU0AD6T0AFB401
- 弁護士が存在しない事件の情報を法廷に提出(2023年6月)
  - 出典：https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/

フェイクニュース

- 「フェイクニュースは、あらゆる形態における虚偽の、不正確な、又は誤解をまねくような情報で、設計・表示・宣伝される等を通して公共に危害が与えられたもの」

（出典：「次世代NHKに関する専門小委員会「次世代NHKに関する論点とりまとめ(第2次)」報告書」(NHK)）
https://www.nhk.or.jp/info/pr/kento/assets/pdf/sub_committee_2_report.pdf

- 東京工業大学の笹原先生は、*「嘘やデマ、陰謀論やプロパガンダ(政治的な宣伝行為)、誤情報(ミスインフォメーション)や偽情報(ディスインフォメーション)、扇情的なゴシップやディープフェイク、これらの情報がインターネット上を拡散して現実世界に負の影響をもたらす現象は、フェイクニュースという言葉で一括りにされています」*と述べている

（出典：笹原和俊著、「フェイクニュースを科学する」、化学同人、2021年、pp.15）

誤情報、偽情報、悪意ある情報

- 誤情報(ミスインフォメーション)、偽情報(ディスインフォメーション)、悪意ある情報(マルインフォメーション)に関する米国DHSの定義。
- 誤情報とは、虚偽の情報であるが、危害を加える意図で作成されたり共有されたりしたものではない。
  - Misinformation is false, but not created or shared with the intention of causing harm.
- 偽情報とは、個人、社会集団、組織、または国に、誤解・危害等を与えるために作成された虚偽の情報。
  - Disinformation is deliberately created to mislead, harm, or manipulate a person, social group, organization, or country.
- 悪意ある情報とは、事実に基づく情報であるが、危害を加えるために使用される情報。
  - Malinformation is based on fact, but used out of context to mislead, harm, or manipulate. An example of malinformation is editing a video to remove important context to harm or mislead.
- DHS(Department of Homeland Security)は、米国の国土安全保障省。あらゆる脅威から米国の国土と国民の安全を守るための組織。

出典:https://www.cisa.gov/topics/election-security/foreign-influence-operations-and-disinformation

ディープフェイク(Deepfake)とその悪用

- ディープフェイクとは、人工知能(AI)の一つである深層学習(ディープラーニング，Deep learning)とフェイク(Fake)をあわせた造語。
- ディープフェイクとは、深層学習技術(Deep learning)を用いて、(人の)画像や音声を人工的に合成する技術。合成メディア(Synthetic Media)。
- エンターテイメント、ゲーム、広告等の様々なビジネス分野のコンテンツ作成において革新と効率化をもたらすと期待され利用されている。
- 一方、ディープフェイクの悪用は大きな社会的問題。詐欺、デマ、ソーシャルエンジニアリング*1、個人情報の窃取、政治・選挙工作、あらゆるものの偽造などに悪用され社会問題となっている。

*1：ソーシャルエンジニアリングとは、ネットワークに侵入するために必要となるパスワードなどの重要な情報を、情報通信技術を使用せずに盗み出す方法。その多くは人間の心理的な隙や行動のミスにつけ込むもの。
（出典：総務省 https://www.soumu.go.jp/main_sosiki/cybersecurity/kokumin/business/business_staff_12.html）

事案1, 2：ディープフェイク用いた詐欺

- 事案１：ディープフェイク音声を用いた詐欺(2019年)
  - 詐欺師は、ある会社(ドイツ)の社長のディープフェイク音声を作り、その子会社(英国、エネルギー関係)の社長をだまして現金を送金させた。被害額は€220,000(約2600万円)。
  - 出典：https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
- 事案２：ディープフェイク動画を用いた詐欺(2024年)
  - 香港の多国籍金融機関の財務担当者は、同社(イギリス)の最高財務責任者(CFO)を装った偽動画に騙されて200million香港ドル(約38億円)を送金。
  - 出典： https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

事案3,4:いやがらせ、偽動画

- 事案3:ディープフェイクを利用した嫌がらせ(2021年)
  - 米国ペンシルバニア州のある母親が、娘のチアリーディングのライバルをチームから追い出そうと、ディープフェイク写真やビデオを利用。
  - この母親は複数の嫌がらせの罪で起訴される。
  - 出典:https://www.bbc.com/news/technology-56404038
- 事案4:首相の偽動画(2023年)
  - 岸田首相の偽動画(2023年11月４日)

    出典:https://www3.nhk.or.jp/news/html/20231104/k10014247171000.html

    出典:https://www.youtube.com/watch?v=0jP2oLilUlw
  - 安倍・菅氏の生成AI偽動画(2023年11月11日)

    出典:https://www.yomiuri.co.jp/national/20231111-OYT1T50021/

事案5, 6：偽画像のSNSへの投稿、拡散

事案5:女性の偽画像を作成しSNSに投稿(2023年)

- 男子大学生が、女子陸上選手の偽画像を作成しSNSに投稿。大学生は名誉毀損容疑で書類送検される(2023年11月)。
- 出典：https://www.yomiuri.co.jp/national/20231106-OYT1T50107/

事案6: スーパースターの偽画像の拡散(2024年)

- テイラー・スウィフトさんの偽画像がSNSで拡散(2024年1月)。SNS各社は削除など対応に追われた。偽画像生成には生成AI技術が利用された可能性が強い。
- スーパースターのファンや議員らは、AIによって生成されたと思われる偽画像が多くのSNSユーザーに共有されたことを非難など。
- 出典：https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html
- 他にも各メディアが報道
  - https://www3.nhk.or.jp/news/html/20240129/k10014339191000.html
  - https://www.yomiuri.co.jp/world/20240130-OYT1T50062/ 等

2024年のサイバーセキュリティ予測と生成AI

- 2024年は、大きな選挙の年(台湾総統選(1月)、ロシア大統領選(3月)、インド総選挙(4/5月)、EU議会選(6月)、米国大統領選(11月)等)、また、パリ五輪などイベントの年。
- McAfee社(ブログ記事)は、本年は選挙の混乱を狙ったディープフェイクや偽情報の増加、本物と偽物の見分けがつなかいコンテンツによるフィッシングメールやAIソーシャルメディア詐欺の増加、ウクライナや中東の紛争に乗じた偽の慈善活動詐欺、五輪イベントに乗じた詐欺が増加するとの予測。

  McAfee社
  出典：https://www.mcafee.com/blogs/internet-security/6-cybersecurity-predictions-for-2024-staying-ahead-of-the-latest-hacks-and-attacks/

- ミュンヘン安全保障会議（MSC：Munich Security Conference）（2024年2月16日）。2024年選挙におけるAIの不正使用と戦うための技術協定「Tech Accord to Combat Deceptive Use of AI in 2024 Elections」。　出典：https://securityconference.org/en/aielectionsaccord/
- サイバー犯罪者も生成AIの最新技術を利用する。
- 偽情報/誤情報等を見分ける、虚偽情報から自分を守る、AIによる脅威から身を守るスキルの向上、そして啓蒙活動が求められている。

4. ネットメディアの特性と人間の認知特性

ネットメディア・SNSの特性と人間の認知特性

●ネットメディア・SNSの特性

- ●プラットフォーム事業者は、利用者個人のクリック履歴など収集したデータを組み合わせて分析(プロファイリング)し、コンテンツのレコメンデーションやターゲティング広告等利用者が関心を持ちそうな情報を優先的に配信。

(出典:情報通信白書、
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/pdf/n2300000.pdf)

●確証バイアス

- 「たいてい、人は信じたいと望むことを喜んで信じる」、「人はみたいように見る」(ユリウス・カエサル)。
- 自分の意見や価値観に一致する情報ばかり集め、それらに反する情報を無視する傾向のこと。

(出典:笹原和俊、「フェイクニュースを科学する」、pp.54-58、科学同人、2021年)

アンコンシャス・バイアス

アンコンシャス(unconscious)　無意識な

バイアス(bias)　偏見、先入観

アンコンシャス・バイアスとは、「無意識の偏見」、「無意識の思い込み」

無意識=自分が自分の行為に気が付かないこと

偏見=偏った見方・考え方、思い込み

（出典：パク・スックチャ著、アンコンシャス・バイアス、ICE SHISHO、2021年）

（出典：一般社団法人アンコンシャスバイアス研究所、https://www.unconsciousbias-lab.org/unconscious-bias）

認知バイアス(bias)の例

- 確証バイアス：自分の意見や価値観に一致する情報ばかり集め、それらに反する情報を無視する傾向のこと。
- 同調バイアス：周囲の言動にあわせたくなる傾向のこと。
- 正常性バイアス：自然災害や事件など危機的状況になっても「大したことはない」「自分は大丈夫」などと思い込む傾向のこと。自分にとって都合の悪い情報を無視したり過小評価する認知特性。
- ステレオタイプ・バイアス：ある集団には特定の特徴があると判断する傾向のこと。
  - ステレオタイプ：ある属性(性別、世代など)に対する先入観、思い込み、固定観念。
- ハロー効果：ある対象を評価する時に、それが持つ顕著な特徴に引きずられて他の特徴についての評価が歪められる現象のこと。
- アンカリング効果：先行する何らかの数値(アンカー)によって後の数値の判断が歪められ、判断された数値がアンカーに近づく傾向のこと。

出典：Wikipedia(https://ja.wikipedia.org)及び
https://www.unconsciousbias-lab.org/unconscious-biasの情報をもとに作成

5.批判的思考スキルを身につける

批判的思考(クリティカル・シンキング)

AI時代、誤情報や偽情報が拡散する時代において、
我々が身につけるべき能力

- 物事を多角的に考える力、
- 正しく疑う力、
- 論理的に思考する力。

クリティカル・シンキング(Critical thinking)

- 物事や情報を無批判に受け入れるのではなく、多様な角度から検討し、論理的・客観的に理解すること。

(出典：デジタル大辞泉、https://kotobank.jp/dictionary/daijisen/3725/)

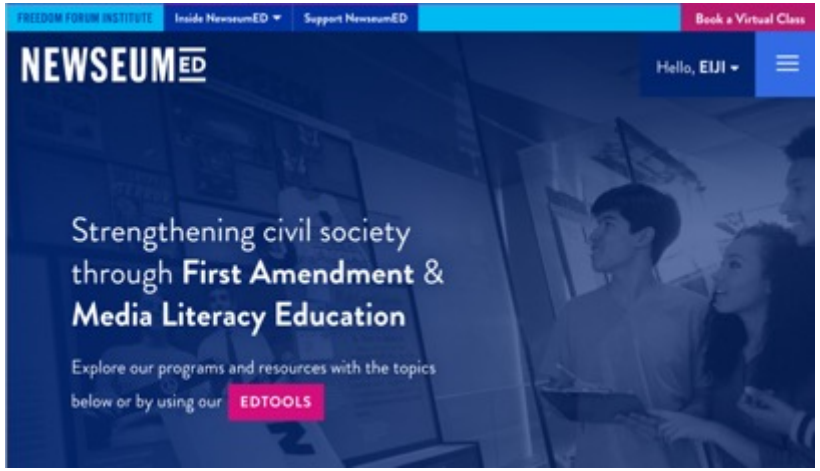
批判的思考スキルの向上と啓蒙活動

批判的思考（クリティカル・シンキング）スキルの向上

- AI時代、誤情報や偽情報が拡散する時代のリスクの理解し、
- オンライン上で自分の身を守り、
- 責任ある市民として必要なスキル

啓蒙活動の例

- E.S.C.A.P.E. Junk News(ニュージアム)(米国)
  - ポスター、各種ツール、レッスン計画書等が用意されている
  - 情報を評価する6つの問い、判断のフローチャート、など



出典：https://newseumed.org/

情報を評価する6つの問い(E.S.C.A.P.E. Junk News(ニュージアム))

- **証拠** (Evidence):その情報は確かかな？
- **情報源** (Source):誰がこれを作ったの？作った人は信用できる？
- **背景** (Context):全体像はどうなっている？
- **読者** (Audience):誰に向けて書いてあるの？
- **目的** (Purpose):なぜこの記事が作成されたの？
- **完成度** (Execution):情報はどのように提示されている？
- 判断のためのフローチャート・ツール、https://newseumed.org/tools/lesson-plan/story-share-worthy

出典:https://newseumed.org/tools/lesson-plan/escape-junk-news
（パンフレットに加え、多くの教材が用意されています）

参考）ファクトチェック

- 目的・定義
  - 公開された言説のうち、客観的に検証可能な事実について言及した事項に限定して真実性・正確性を検証し、その結果を発表する営みを指す。
  - ファクトチェックの3要素

    ①対象言説の特定、②対象言説の真実性・正確性の判定、

    ③判定の理由や根拠情報の公開

    出典：「日本におけるファクトチェック活動の現状と課題」、総務省プラットフォームサービスに関する研究会(2023年2月10日)、https://www.soumu.go.jp/main_content/000861267.pdf)

- メディアや団体がファクトチェックを実施している
  - 特定非営利活動法人ファクトチェック・イニシアティブ(FIJ)　https://fij.info/
  - BuzzFeed　https://www.buzzfeed.com/jp/badge/factcheckjp
  - The International Fact-Checking Network (IFCN)　https://www.poynter.org/ifcn/

  など

参考）各国政府や法による規制の動向

政府の動向、法による規制

- 米国(大統領令)(2023年10月30日)
  - 米バイデン大統領は、**AIの安心・安全・信頼できる開発と利用に関する大統領令**に署名。
  - https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
  - AI の利用は、世界を生産的・革新的・安全なものにし、社会的課題を解決するのに役立つ可能性を持つ。しかし、同時に、無責任な使用は、詐欺、差別、偏見、偽情報などの社会的損害を悪化させる可能性がある。
  - 8つの基本原則：AIは、安全でセキュアでなければならない、イノベーションと競争の促進、労働者の保護と支援、公平性と公民権の促進、プライバシーと市民的自由の保護、消費者の保護、政府によるAIの責任ある効果的な利用、米国の世界的なリーダーシップの促進。
- EU(AI Act、AI法案)(2023年12月9日)
  - AIの利用に関する包括的な規制案。AIのリスクに応じて規制。
  - https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/
  - EU内にAIサービスを提供するEU域外企業も対象。法案に違反した場合は巨額の制裁金と制約措置が適用される。

政府の動向、法による規制

- 英国
  - **AI Safety Summit** (2023年11月1日、2日)
  - 米国、EU、中国を含む28か国が合意した「Bletchley(ブレッチリー)宣言」が発行。
  - この宣言を、ブラジル、フランス、インド、アイルランド、日本、ケニア、サウジアラビア王国、ナイジェリア、アラブ首長国連邦等が支持。
  - AIモデルに起因して、意図的か非意図的かにかかわらず、深刻な、さらには壊滅的なリスク・危害が発生する可能性がある。
  - AIに関するリスクの理解とさらなる国際的な協力の確立。
  - https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration

- 日本
  - **G7広島AIプロセス**、G7デジタル・技術大臣会合(2023年12月1日)
  - 全てのAI関係者向けの広島プロセス国際指針
  - https://www.soumu.go.jp/main_content/000915261.pdf

Hiroshima AI Process
G7 Digital & Tech Ministers' Statement (December 1, 2023)

AI利用の国際ガイダンス

- AI利用の国際ガイダンス(オーストラリア、The Australian Cyber Security Centre)
  https://www.cyber.gov.au/resources-business-and-government/governance-and-user-education/governance/engaging-with-artificial-intelligence
  - AI システムの利用者向けに、AI関連の脅威、AI システムを使用する際のリスク管理などのガイドラインを提供。
  - 米国FBI、NSA、英国のNCSC、日本の内閣サイバーセキュリティセンター、シンガポールのCSA等と連携しガイドラインを作成。
  - AI使用に関する国際ガイダンスへの共同署名について(内閣府：https://www8.cao.go.jp/cstp/stmain/20240124.html)。
  - 仮訳：https://www.nisc.go.jp/pdf/policy/kokusai/Provisional_Translation_JP_Engaging_with_AI.pdf

その他）最近の動向(米連邦通信委員会)

- 米連邦通信委員会は、国民を詐欺や誤った情報から守るための手段として、AIが生成した自動音声通話「ロボコール」を電話消費者保護法(TCPA)の監視の対象とした(2024年2月8日)。
  - 米国連邦通信委員会(FCC)：米国において電気通信・放送分野を所掌。
  - 電話消費者保護法は、迷惑電話を制限するための法律であり、人工的または事前録音した音声メッセージの使用を制限している。事業者は消費者にロボコールをかける前に、事前に書面による明示的な同意を得なければならない。今回、AIが生成した通話音声も電話消費者保護法の基準に従うことになった。
  - 背景として、ディープフェイク音声詐欺対策や11月の大統領選挙に向けた誤情報拡散対策などが考えられる。
- https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal
- https://edition.cnn.com/2024/02/08/tech/fcc-scam-robocalls-ai-generated-voices/index.html

本日のまとめ

- AIの課題と事案
- フィッシング詐欺等の手口はAI技術利用により巧妙化
- 本物と偽物の見分けがつかない
- ネットメディアと人間の特性によって虚偽情報が拡散
- 批判的思考スキルを身につける

ご清聴ありがとうございました