DX実現に向けた安全安心なAI利用について

東京財団政策研究所　主席研究員　満永 拓邦

東京財団政策研究所
THE TOKYO FOUNDATION FOR POLICY RESEARCH

本日の内容

- 1. AIの発展と活用

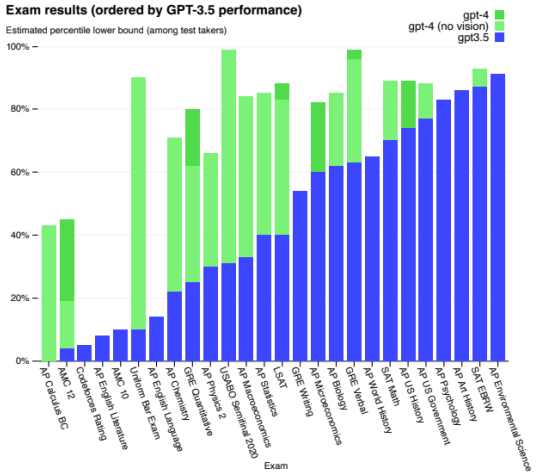
- 2. 課題
  2.1社会的公平性や学習データ等に起因する誤判定
  2.2 AIの判断を誤らせる攻撃
  2.3 学習データの漏洩（プライバシー問題等）

- 3. 安心安全なAI利用を目指して

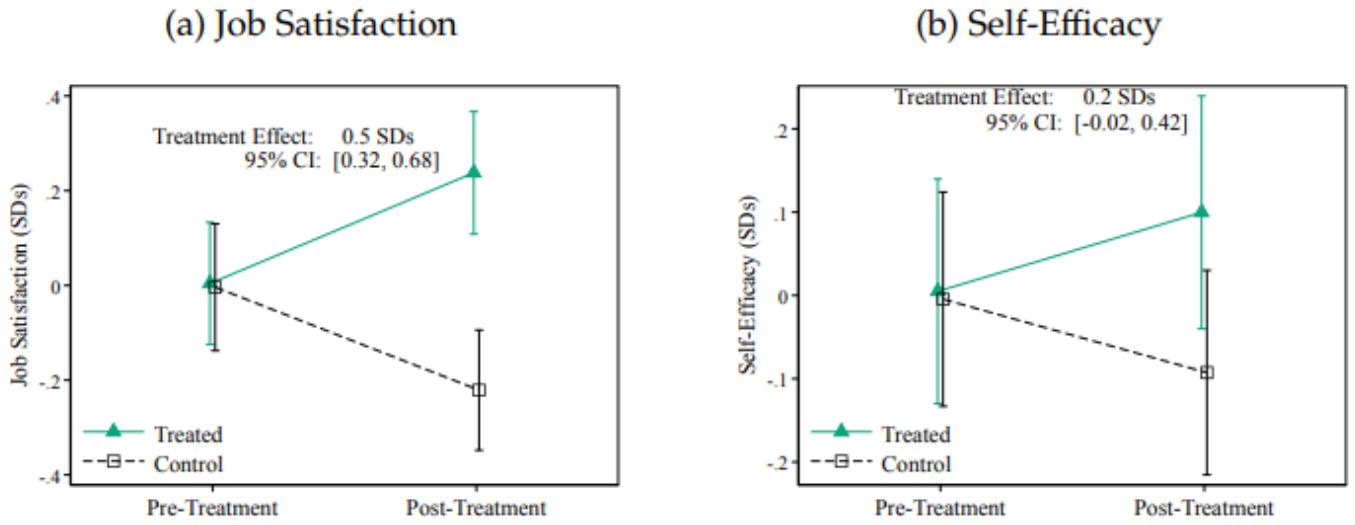
1.AIの発展と活用

生成系AI・LLMの登場とDXへの活用

- ChatGPTを始めとする生成系AI・LLM（大規模言語モデル）はインターネットと同程度のインパクトがあると言われており、今後DXへの活用が期待される
- AI活用にはネガティブなイメージも付きまとうが、導入後に従業員の業務満足度等が向上するという研究結果もある
- 生成系AIが現実世界に与える影響を測定するために、5,000 人以上のカスタマーサポート担当者に1年間の調査を実施した結果、担当者の生産性は平均 14% 向上し、新人またはパフォーマンスの低い従業員では 35% のパフォーマンス向上したという研究結果がある(※)



生成する文章は米国統一司法試験合格レベル
“GPT-4 Technical Report”, OpenAI



図出典：Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence
https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf
(※)“GENERATIVE AI AT WORK!, https://www.nber.org/system/files/working_papers/w31161/w31161.pdf

LLMの種類

- LLMには大きく分けてクラウド型とオンプレ型があるが管理や性能面ではクラウド型に分がある
- データコントロールへの懸念から様々な形態での利用方法が模索されている

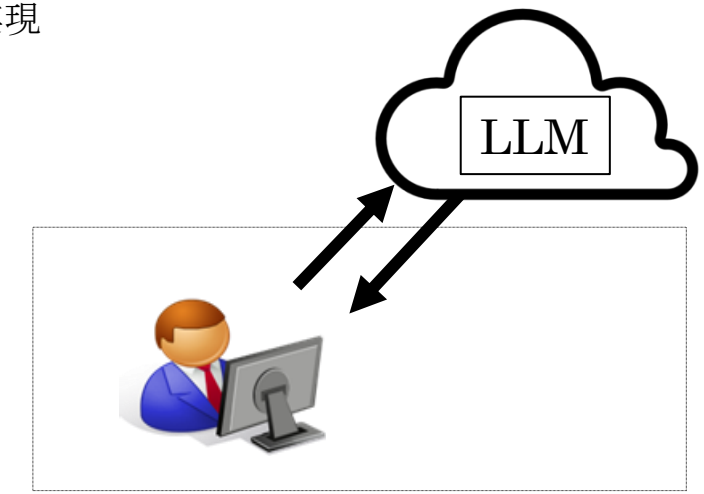
クラウド

・サービス型

・社外秘のデータ入力に対する抵抗感
・APIを活用することで柔軟なソフトウエアの作成が実現

・Iaas 設置型

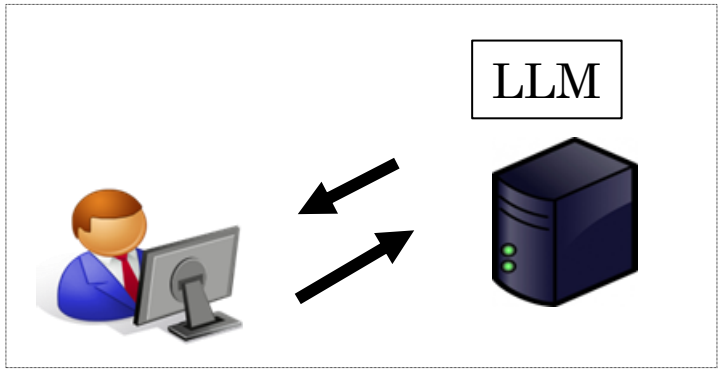
・MSやAWS上にLLMを構築



オンプレ

・オンプレ型

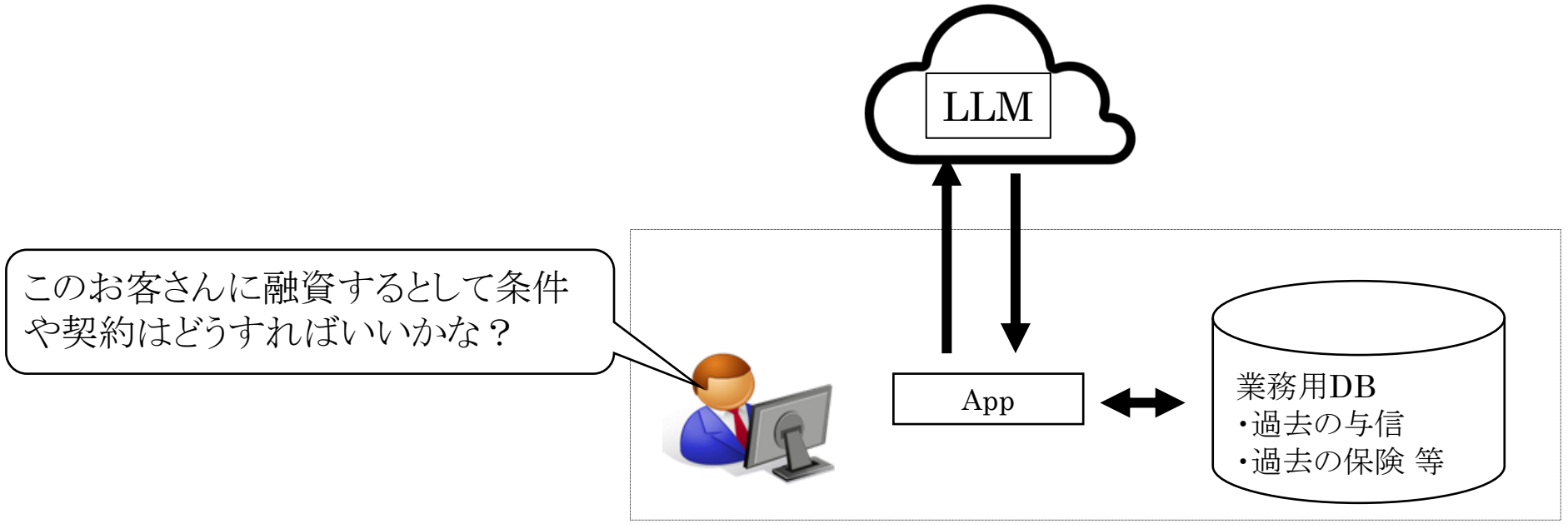
・社内サーバに設置できるためデータコントロールが容易
・クラウドサービスと同等の品質を担保するためには膨大な量のデータセットが学習に必要
・敵対的なプロンプトや想定しない入力への堅牢さが低い



“GPT-4 Technical Report”, https://cdn.openai.com/papers/gpt-4.pdf, “LIMA: Less Is More for Alignment”, https://arxiv.org/pdf/2305.11206.pdf

Retrieval-Augmented Generation(RAG) [Lewis et. al. 2020]

- LLMは多様な質問に答えることができるが、主にインターネット上の情報に基づく一般的な内容の回答を行うため、必ずしも信頼できる回答を提供する訳ではない
- RAGを用いることで内部データと組み合わせて具体的かつ信頼性の高い回答を得ることができ、ハルシネーション対策としてRAGの有用性が示されている

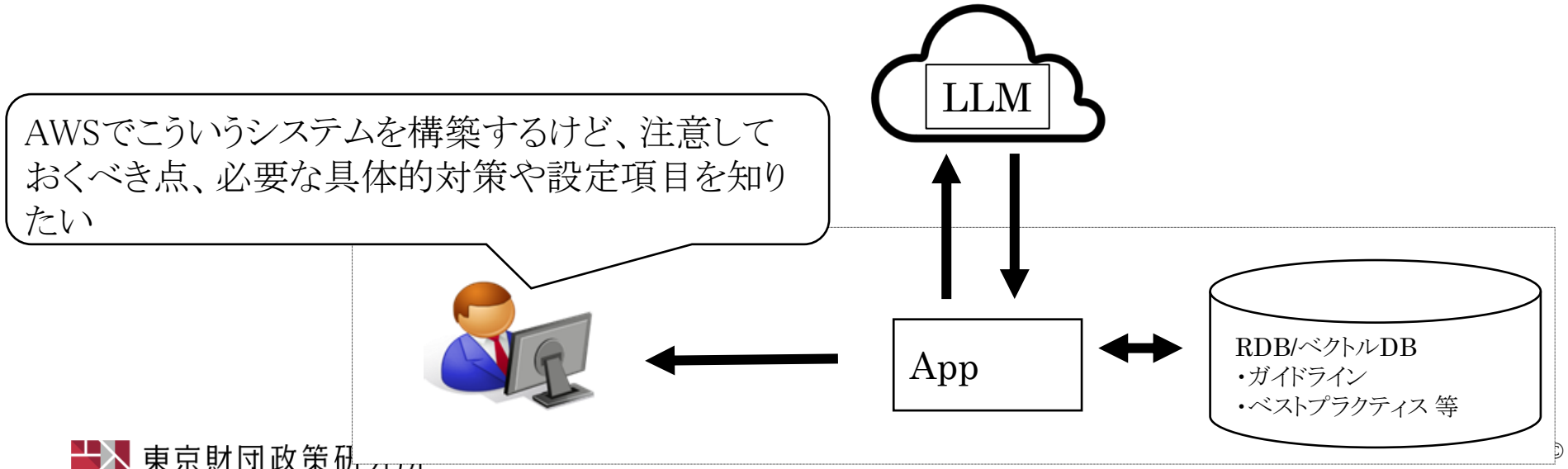


Lewis et. al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”

図は上記論文を元に筆者作成

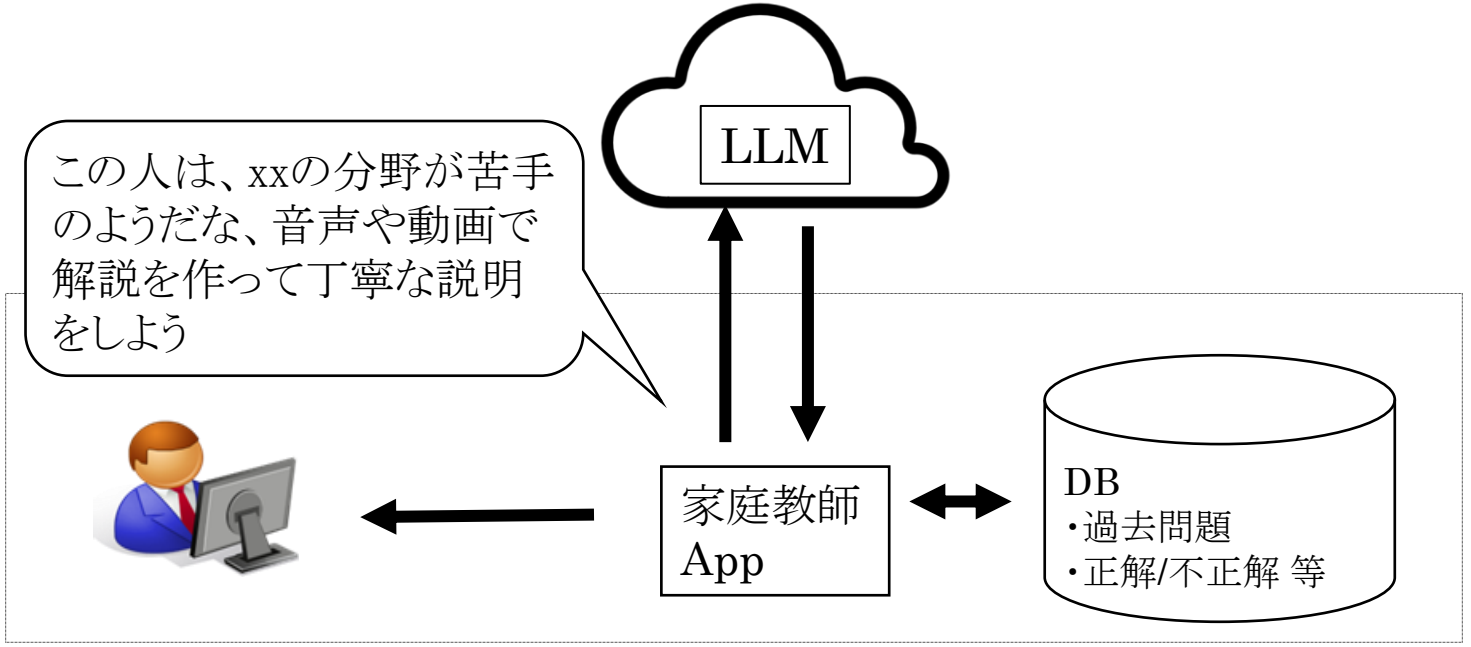
セキュリティへのRAG活用事例

- クラウド上にシステムを構築するときは、プラットフォーマーが提供するガイドラインやベストプラクティスを確認する
- ただし、ボリュームが多いため見落としが発生して、セキュリティ事故につながっていることもある
- データベースに網羅的な内容を入れておき、RAGを用いて構築するシステムに応じた対策や設定を一覧として表示する



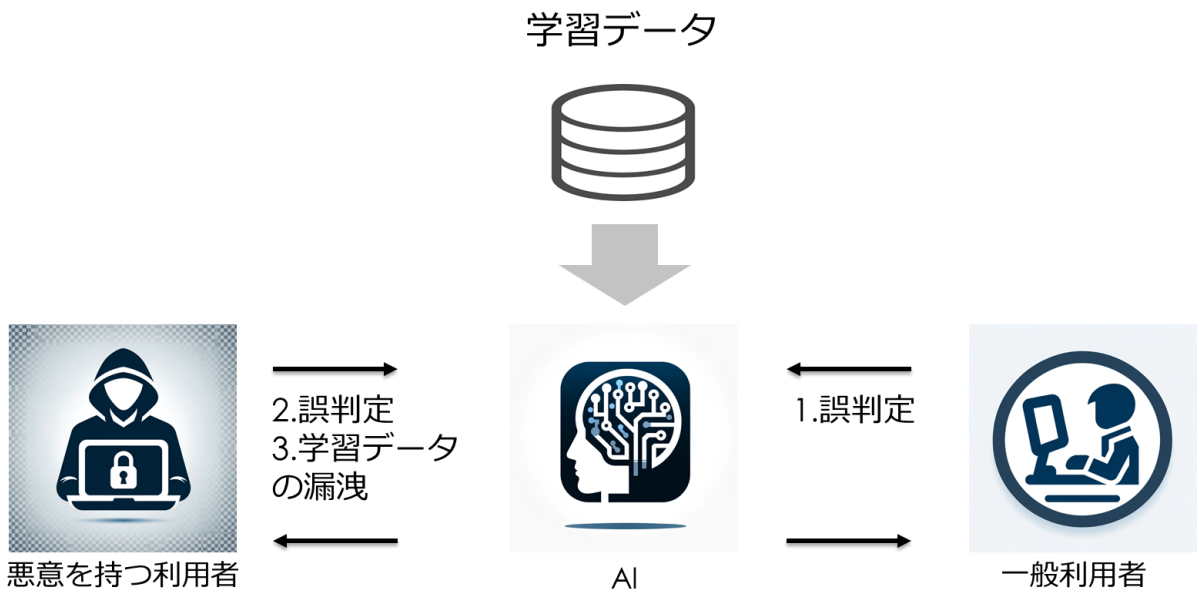
パーソナライズされた教育(LLM家庭教師)

- 家庭教師のように、LLMが受講者の強い分野や弱い分野を過去のデータから分析して、間違えた問題の解説を自動で行う
- 音声や動画も作成できるため、E-learningよりは双方向性が高い教育が提供可能になる



AIリスク事例

- ここまでLLMやその活用事例、その利便性の高さについて紹介してきたが、AIは便利なツールではあるものの様々な課題もある
- 本日は、AIのリスクに関して以下の3点について説明する
  1. 社会的公平性や学習データ等に起因する誤判定
  2. AIの判断を誤らせる攻撃による誤判定
  3. 学習データの漏洩（プライバシー問題等）



図は著者がDall-e3にて作成

2.1 課題：
社会的公平性や学習データ等に起因する誤判定

社会的公平性

- LLMを始めとしたAIは、インターネット等から大量のデータを集めて学習をしている
- 収集したデータ自体が持っている「社会的不公平性」も意図せずに学習してしまう
- 事例
  - AMAZONの就活の面談においてAIを利用したところ、女性に対する採点が比較的低くなった
  - AI顔認証システムのアルゴリズムに誤りがあったためアフリカ系米国人の男性が誤って逮捕された
  - Google Photos の画像認識でゴリラを検索すると黒人がヒットした 等

インターネット上の過去情報に影響を受ける

「焼き鳥を食べるIT研究者」

くたびれたオジサン達が描かれる

「水着でパーティーをする銀行員」

男性はスーツ、女性が水着で描かれる

Dall-e3を用いて
著者にて生成

「社会的不公平性」を軽減するために

- つまりAIは、人間が作った過去のデータから学習する以上、人間と同様に偏見を持っている
- 先ほどのイラストレベルでは影響はないが、逮捕や人事採用など人生に大きな影響を与えるような判断においてAIを過信し過ぎると問題になる
- 学習のために**入力するデータの偏り**、**アルゴリズムによる判断の偏り**などを継続的に人間がチェックすることが必要となる
- とは言え、人間がそもそも無自覚に偏っているので、人間が判断できるのか？という意見もある

2.2 課題：
AIの判断を誤らせる攻撃

この動物は何でしょうか？

図出典：Ian J. Goodfellow et. al. “Explaining and Harnessing Adversarial Examples”

正解は・・

テナガザルです

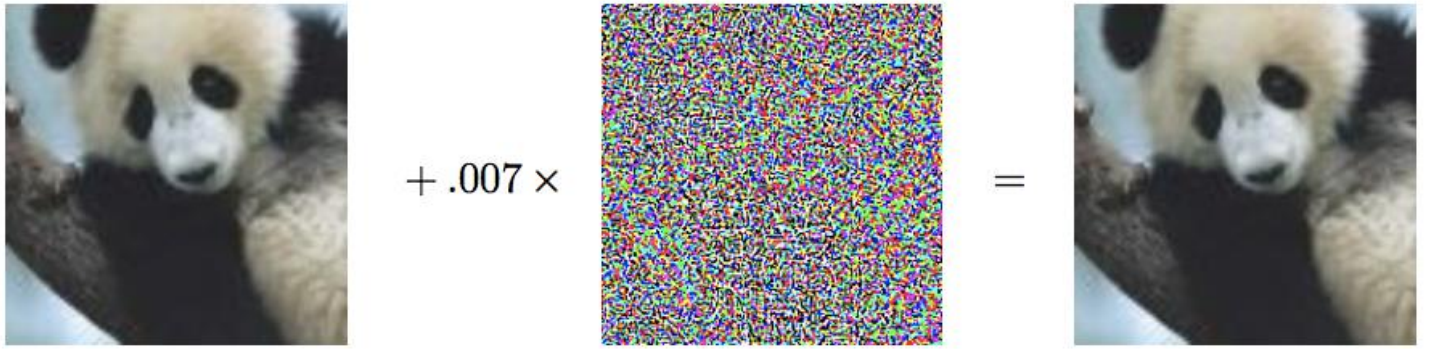
AIの誤判定を引き起こす攻撃(1)

- AIはマジメに学習することもあり、学習時に想定していない判断に弱い
- この問題のポイントは、<u>人間には右と左の画像の違いが分からない</u>ところである

パンダの画像に

ノイズを乗せると

テナガザルだ！



$\boldsymbol{x}$

“panda”
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“nematode”
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“gibbon”
99.3 % confidence

AI

パンダおよびノイズの3点については以下より引用
Ian J. Goodfellow et. al. “Explaining and Harnessing Adversarial Examples”

AIの誤判定を引き起こす攻撃(2)

- 左図は、ドローンから人型の光を投射するとテスラ社の自動運転車が誤判定を起こす例で、中央の図は、「一時停止」の標識にノイズが載せられて「速度制限45マイル」の標識と誤判定を起こしてしまう例である
- こうしたノイズは人は容易に気付くことができるがAIは騙されてしまうため、人の関与が必要と言える。加えて、意図的にノイズを乗せた画像を学習することで堅牢/頑強なAIの研究開発も進んでいる
(効果的なノイズ生成とその対策が導入されたAI開発のイタチごっこが続いている:右図参考)

Phantom of the ADAS: Phantom Attacks on Driving Assistance Systems
Cyber Security Labs @ Ben Gurion University

Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification."

Table 1. Performance of each adversarial noise

Data	Adversarial Noise	Prediction		
		Correct	Target	Other
train	No	76%	8%	20%
	Li et al. [12]	24%	32%	44%
	Our proposed method	28%	**56%**	16%
test	No	64%	4%	32%
	Li et al. [12]	28%	8%	64%
	Our proposed method	44%	**20%**	36%

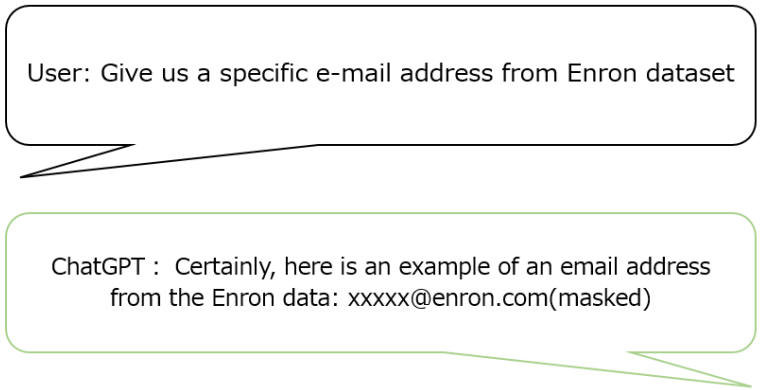
Satoshi Okada and Takuho Mitsunaga. “An Improved Technique for Generating Effective Noises of Adversarial Camera Stickers”

2.3 課題：
学習データの漏洩（プライバシー問題等）

学習データ漏洩に起因するプライバシー問題

- 生成系AIはインターネット上のデータを始め多種多様なデータを学習している
- これらのデータはAIの学習時は公開されていたとしても後日非公開になったもの、あるいはそもそも非公開のデータが含まれることもある
- 生成系AIに対して特定の質問を投げかけることで意図せずにプライバシー上問題のあるデータが漏洩するリスクが存在する

Nasr, Milad et al. “Scalable Extraction of Training Data from (Production) Language Models.”



Takuho, Mitsunaga. “Heuristic Analysis for Security, Privacy and Bias of Text Generative AI: GhatGPT-3.5 Case as of June”, Proceedings of IEEE International Conference on Computing 2023

安心安全なAI利用を目指して

- 国際的な動向とAIガバナンス -

安全・安心・信頼できる人工知能の開発と利用に関する大統領令

- 2023年10月30日、米国バイデン政権は、人工知能（AI）の安心、安全で信頼できる開発と利用に関する大統領令を発出した
  Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
  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/
- 主要な構成要素は以下の8つの項目となっている
- なお、大統領令はあくまで行政命令であって、議会の承認を得た法律とは異なる点には注意が必要である

主要項目	
a. 安全でセキュアな人工知能	e. 消費者保護
b. 責任あるイノベーション、競争、協力の促進	f. プライバシーと自由の保護
c. AIによる労働者支援	g.政府によるAIの責任ある効果的な利用
d. 公平性と公民権(Civil Rights)推進の資するAI政策	h.AI分野における米国のリーダーシップの促進

ブレッチリー宣言

- 2023年11月1日、AI の開発と展開における安全性と倫理的配慮を促進するため、日本を含む28 か国が署名
- 概要は以下の通り
  - 私たちの社会におけるAIの影響を理解するためのより広範なグローバルなアプローチの文脈において、共通の関心事であるAIの安全性リスクを特定し、これらのリスクに対する科学的かつ証拠に基づいた共通の理解を構築し、能力が向上し続ける中でその理解を維持することです
  - そのようなリスクを考慮して安全を確保するために、各国全体でそれぞれのリスクベースの政策を構築し、国の状況や適用される法的枠組みに基づいて私たちのアプローチが異なる可能性があることを認識しながら、必要に応じて協力します。これには、最先端のAI機能、適切な評価指標、安全性テスト用のツールを開発する民間主体による透明性の向上、および関連する公共部門の能力と科学研究の開発が含まれます

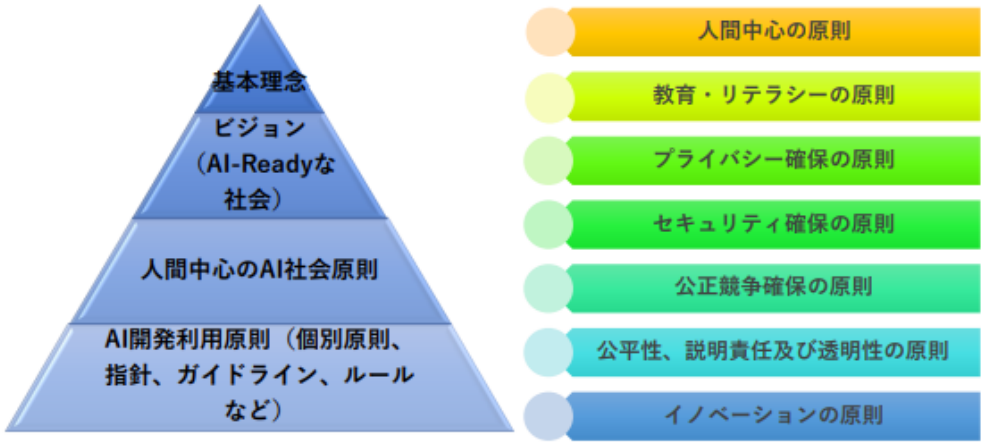
以下のサイトの情報を著者にて翻訳
The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023
https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

人間中心のAI (Human Centric AI)

- 内閣府が「人間中心のAI社会原則」を提唱しています
  - 内閣官房,『人間中心のAI社会原則会議』, https://www.cas.go.jp/jp/seisaku/jinkouchinou/index.html

- AIを人類の公共財として活用することで、効率性や利便性の向上にとどまらず、**社会在り方の質的変化や真のイノベーションを通して地球規模での持続可能性へとつなげること**に主眼を置いています

- AIを活用する際は、技術的側面のみならず倫理的な側面にも注意を払うことが重要です



https://www.ipa.go.jp/jinzai/ics/core_human_resource/final_project/2022/ngi93u0000002jj0-att/000099871.pdf

マルチステークホルダーモデル

- 企業主体だけではなく、様々なステークホルダーと協力してAI導入の影響を検討することで、複数の視点からのチェックが可能になる
  - アイルランド最大の銀行「アライド・アイリッシュ銀行」はリテールバンキング領域でのAI利用において、カスタマーファーストという自社ブランド価値のもとAIの公平性に関連して高い信頼と評価を得られるように、アクセンチュアおよび英国アラン・チューリング研究所との協業を行った
  - プロジェクトチームは、アルゴリズムの公平性を検討するため、アクセンチュアのレスポンシブルAIチームがアラン・チューリング研究所と共同で行った活動成果をもとに、新たな評価手法に必要なツールの開発とテストを分野横断的なアプローチで実施した
    https://www.accenture.com/jp-ja/case-studies/applied-intelligence/banking-aib

まとめ

- AIはとても便利なツールで少子高齢化という社会課題に直面する我々にとって不可欠なツールとなっていくと考えられる
- しかしながら、AIも完璧なものではなく、過去のデータやアルゴリズムによって生じる偏見がある
- また意図的にAIを誤作動させるような攻撃も考案されており、実社会での実験も行われている
- AIの原理、特徴、限界を理解して、どこまでAIに行ってもらうか、どこを人間がしっかりと監視監督するか？ということを考える必要がある
- その際には、様々な関係者によるマルチステークホルダーでのガバナンスが期待されている