資料１－２

AI利用と個人情報の関係の考察

2024年4月3日

日本電信電話株式会社
NTT社会情報研究所
高橋克巳

AI利用と個人情報の関係：検討の目的

- 検討の目的
  - AIを事業に供する側には、AIでデータを利用することと個人情報保護の規律とは相性が悪いとの指摘がある
  - 一般の市民には、AIの活用に関して個人情報やプライバシー保護の漠然とした不安がある
  - 今後のルールづくりのためのAIに関する技術情報の提供を目的とする
- 近年AIの技術革新
  - 近年の技術革新で、AIは複雑な概念をより自由に大量に扱えるようになり、その結果、精度、汎用性、処理速度などが格段に向上した
    - 関連ワード：大規模言語モデル、基盤モデル、生成AI
- AIの個人情報保護の規律
  - AI全般に関して個人情報保護の規律を見直す良いタイミングである
    - 課題に対応し、AIの仕組みにフィットした新たな規律
    - 「近年の技術革新によるAI」を最初のターゲットにすることに一定の合理性がある

AI利用と個人情報の関係について、以下を説明

1. 新たな規律の考え方
2. AIの仕組みとAIで個人情報が扱われるかどうかの評価
3. 新たな規律下におくAIの範囲

AIデータ利用個人情報保護の新たな規律の考え方

- AI活用事業者が個人情報を取り扱う場合、以下に関する義務に留意する必要がある
  - 例）利用目的、 個人情報の取得、個人データの管理、第三者提供の制限、開示・訂正・利用停止、等

- 一般の市民には、AIの活用に関する漠然とした不安がある

- AIには独特の仕組みや運用があり、その一部には従来の個人情報取扱の習慣と異なったものがある

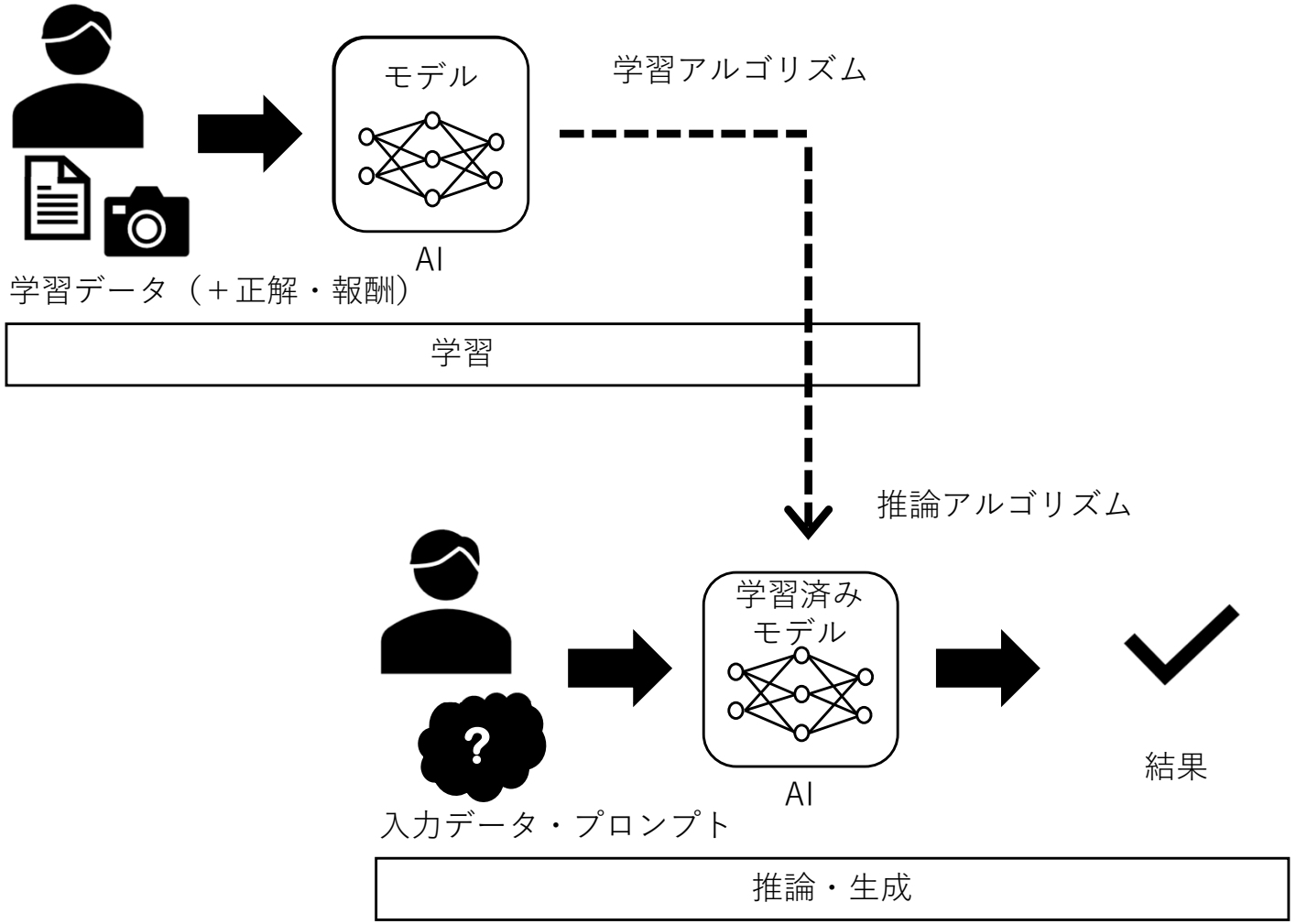
- 課題（事業者義務の困難さ、市民の本質的な不安）に対応し、AIの仕組みにふさわしい形で新たな規律が確立されることが待たれる

AI利用と個人情報の関係について、以下を説明

1. 新たな規律の考え方
2. AIの仕組みとAIで個人情報が扱われるかどうかの評価
3. 新たな規律下におくAIの範囲

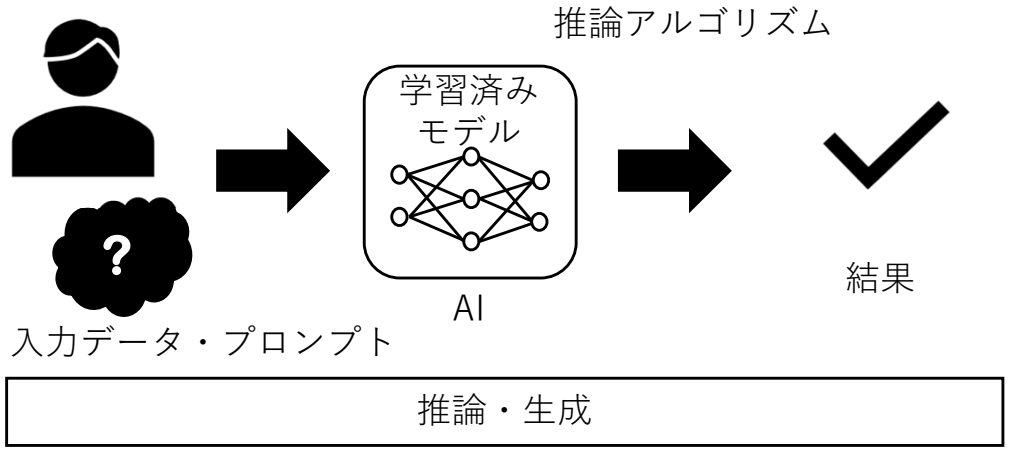
AIの仕組み

- 学習と推論の２つのフェーズがある
- データを学習させた結果を学習済みモデルと呼び、それを用いて推論が行われる



AIの仕組み（モデル）

- 学習済みモデルとは
  - データに学習と呼ばれる処理をして作成したルール（数学的な手続き）の集り
  - ノードとノードが層状に繋がったネットワークで表現
- モデルは何をするのか（ニューラルネットワークで推論の場合）
  - 各ノードはデータを受け取り、次の層のノードにデータを出力する（伝搬）
  - ノードはノードに与えられた手続きを用いてデータを伝搬する
    - 手続きには学習によって得られた重み(w)やバイアス(b)を含む
    - 出力手続きの例：$f(x) = wx + b$
  - 重みとバイアスをまとめてパラメータと呼ぶ

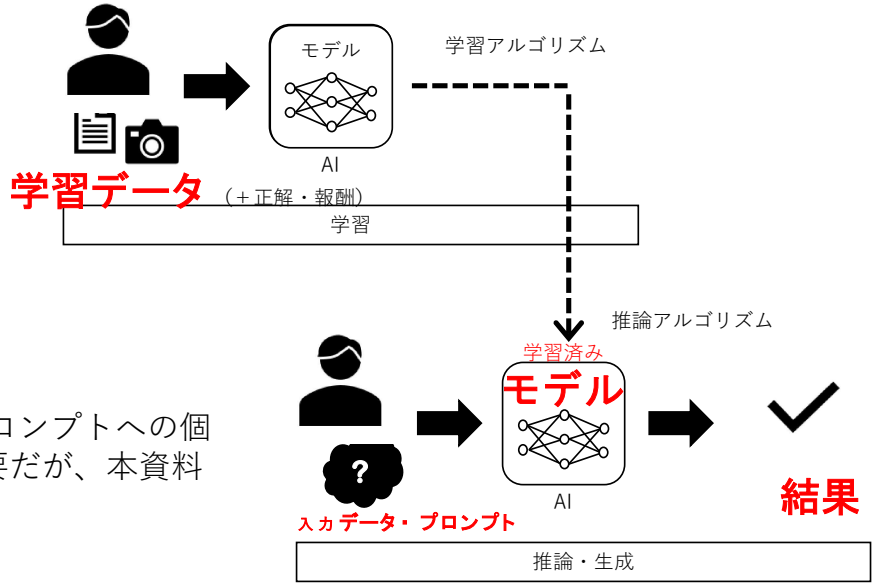


AIで個人情報が扱われるかどうかの評価

- AIで個人情報が扱われるかどうかの基本的な考え方

A)	AIが個人単位のデータ処理を意図して設計・運用される場合	扱われる
B)	AIが個人と関係のないデータ処理のために設計・運用される場合	扱われない
C)	上記以外	どちらの場合もある
※ これらはいわゆる「基盤モデル」前提の考察で、AIが追加してチューニングされた場合や、サブシステムと組み合わせて実装された場合は、その追加要素の振る舞いに依存する		

- Cの一類型として以下を考察
  - C1) AIが様々な情報を一般的な知識や概念として扱う場合
    - 「近年の技術革新によるAI」にこれに該当するものがある
    - 学習データの観点、学習済みモデルの観点、推論結果の観点から個人情報との関係を考察する



※ なお、入力データ・プロンプトへの個人情報入力の留意も必要だが、本資料では論じない

AIで個人情報が扱われるかどうかの評価（続き）

C1「様々な情報を一般的に扱うAI」の場合の考え方

観点		評価
学習データの観点	インターネット上のオープンなデータを用いる場合	• 個人情報は含まれる • 個人情報の完全な排除は技術的に困難とされる
	クローズドなデータ（独自データ）を用いる場合	• 従来の規律に則った取り扱いが可能
学習済みモデルの観点	個人単位のデータ処理を意図して設計・運用される場合	• 従来の規律に則った取り扱いが可能
	一般的なニューラルネットワーク(含む、LLM)の場合	• 個人情報を扱うとはみなせないのではないか（※条件あり）*1
推論結果の観点	• 個人情報相当をAIが一般的な知識の一環として出力することは起こりうる • 出力の抑制が対策の本命として業界で取り組まれている（ガードレール）*2 • なお、完璧な出力の抑制は技術的に困難とされ、リスクベースアプローチに合理性があると考えられている	

*1 モデルと個人情報（一般的なニューラルネットワークに個人情報を含むデータを学習させた場合）
- モデルはノードの集まりで、ノードは手続きとパラメータから構成される（例、 $f(x) = wx + b$ の $f,\ w,\ b$ ）
- 手続きとパラメータは個人情報に相当せず、これらから元のデータの復元も通常できないと考えられる
  - 恣意的にパラメータから元のデータ等が復元できるようにされたモデルが作られた場合はこの限りではない
  - モデルから元データの推定を試みる研究は存在している
- モデルの入り口と出口で、データは単語や文字等を数値表現したトークンという単位で扱われる
  - 例えばトークンに氏名の一部相当が含まれることがあるが、一般的な概念の一環として扱われる限りにおいて個人データとみなせないのではないか

*2 出力抑制の動向
- どのように実装されていくか、どの程度のガードができるかは今後の動向による
- （個人を指定した削除というよりは）例えば回答の内容が不適切であることを見分けて抑制する方向性

（参考）AIと個人情報の関係の評価に標準が使えるか

- 国際標準
  - AIの仕組み（アルゴリズム・モデル）に関する標準はない
  - AIマネジメントに関するものあり（NIST, ISO等）
    - リスク管理のフレームワーク

- 研究者らが提唱する評価指標（透明性指標）
  - 学習データの透明性
  - モデルの透明性
  - 運用の透明性

- これらのフレームワークは有用であるが、以下の理由でAIの新しい個人情報保護の規律の主軸とはならない
  - AIで個人情報が扱われるかに関する直接的な標準がない
  - AI技術の変化が激しく、標準によりAIの善し悪しを判断できる状況にない

AI利用と個人情報の関係について、以下を説明

1. 新たな規律の考え方
2. AIの仕組みとAIで個人情報が扱われるかどうかの評価
3. 新たな規律下におくAIの範囲

新たな規律下におくAIの範囲

- AIに関する個人情報保護の新たな規律を考えた場合、その対象範囲を適切に設計しておく必要がある（どんなAIを当てはめるのか？）

- 「近年の技術革新によるAI」*3で、情報を広く・機械的に扱うもの
- これらのAIを新たな規律下に置く最初のターゲットにすることに一定の合理性があるのではないか
  - *3（近年の技術革新によるAI）近年の技術革新で、AIは複雑な概念をより自由に大量に扱えるようになり、その結果、精度、汎用性、処理速度などが格段に向上した
  - 当該対象AIが、さらに追加してチューニングされた場合や、サブシステムと組み合わせて実装された場合は、その追加要素の振る舞いに応じて規律が定められるべきである

- （参考）既存のAIの分類の考え方を次ページに示す
  - AIの分類には絶対的なものはなく、また今後の変化も予想される
  - 対象範囲は固定せず、柔軟に拡充等対応することが望ましい

（参考）AIの分類例

生成・対話	チャット、画像生成
予測・制御	中古車価格予測、プラント制御、人員計画
認識・推定	音声認識、生体認証、異常検知、劣化推定
分析・要約	統計、疫学、ニュース要約
設計・デザイン	生産計画、パーソナライズ広告
協働・信頼形成	スクリーニング、合意形成

- 解決する課題（用途）
  - 生成・対話、予測、認識、・・・・→
  - 特化型、汎用型
- 扱うデータ
  - 画像、音声、テキスト、時系列データ、マルチモーダル
- 手法（学習）
  - 教師あり学習（分類、回帰）、教師なし学習（クラスタリング、次元圧縮）、強化学習（例、将棋）
- 手法（モデル・推論）
  - 機械学習、知識ベース、統計ベース
    - EU AI法におけるAIの定義はこれ
- 基盤・応用
  - 基盤モデル、(基盤モデル)応用AI、サブシステムと組み合わせて構築されたAI
- 参考
  - 人工知能学会 AIマップβ2.0 (2023年5月版)
    - https://www.ai-gakkai.or.jp/aimap/

まとめ

- AIの浸透に対応した、個人情報保護の新しい規律の確立が必要である

- 課題（事業者義務の困難さ、市民の本質的な不安）に対応し、AIの仕組みにふさわしい形で新たな規律が確立されることが待たれる

- 新たな規律の対象は、まずは情報を広く・機械的に扱う「近年の技術革新によるAI」をターゲットに考えることに一定の合理性がある