

## 実践コラボ演習(AIと社会) 第3回

### AIにまつわるリスク、AIとの正しい付き合い方

理工学部 情報科学科教授 成 凱 せい がい

## AIに対する様々な懸念

- AIが人間への脅威にならないか？
  - AIが暴走してコントロールできなくなることはないか
  - 人間の生死をめぐる判断をAIに委ねて良いか
- AIを信頼して仕事を任せてよいか？
  - AIが人間の意図に反して行動しないか
  - AIに脆弱性がないか、不正利用や攻撃に耐えられるか
- AI判断に公平性、透明性が担保されるか？

実践コラボ演習(AIと社会)(3)

chengk@is.kyusan-u.ac.jp

6

## AIの脆弱性

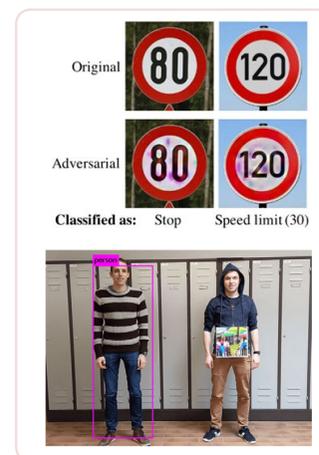
- AIをだます敵対的サンプル(Adversarial Example)
  - AIモデルに誤分類を引き起こさせるために、人間にはわからないようなわずかなノイズ(「摂動」)を加えたサンプル(画像など)
  - AIを応用する際に重大な損害をもたらす深刻な問題
    - ① 自動運転や医療等
    - ② 顔認証、セキュリティ

実践コラボ演習(AIと社会)(3)

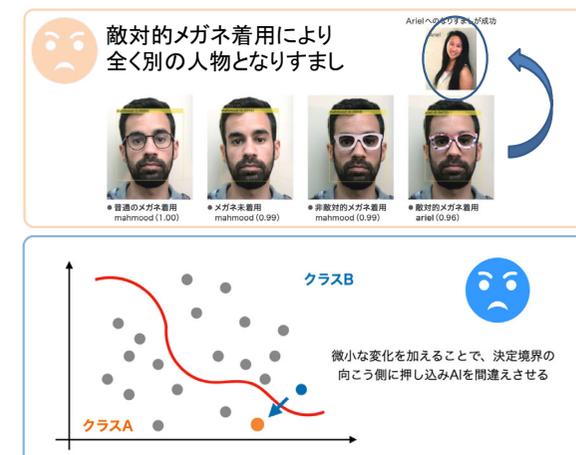
chengk@is.kyusan-u.ac.jp

8

## 敵対的サンプル例



実践コラボ演習(AIと社会)(3)



chengk@is.kyusan-u.ac.jp

9

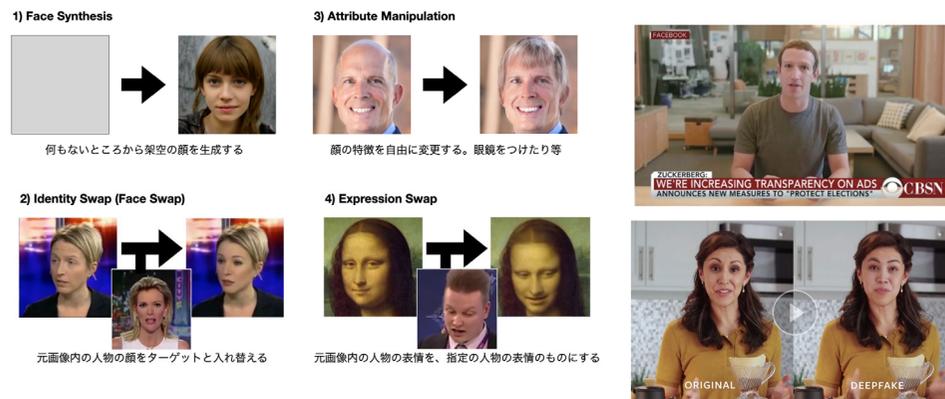
## 従来のプログラムとAIプログラムの違い

- 従来のプログラム
  - 人間のプログラマーがプログラムの動作を直接コントロール
  - 人間がそのプログラムの動作を説明することが可能
- AIプログラム
  - 大量のデータでAIプログラムを学習させ、プログラムの動作がデータに依存する
  - その結果は人間も予測できず、説明することが極めて困難

## AIプログラムの弱点

- AIの機械学習に必要な教師データが常に不足
- 情報に偏りのあるチャンネルからデータ収集
- AIプログラムにその偏りや裏の意味を理解する能力不足
- AIに道徳的・倫理的な善悪の判断・精査を行う能力不足

## AIの悪用：フェイクニュース・動画



## AIの誤用

- リクナビにおける内定辞退率販売問題
  - 「リクナビDMPフォロー」というサービスで、2018年度に対象企業に応募した学生のリクナビ上の行動履歴などを分析
  - 作成したアルゴリズムで、2019年度の学生の行動履歴と照合することで内定辞退率を予測。予測結果を企業に販売
- BLM運動と顔認証技術
  - アメリカを代表するIT企業が相次いで顔認証システムの開発・販売から撤退、あるいは警察への販売を停止



リクナビ社長らによる謝罪



## 信頼できるAI (Trustworthy AI)

- 頑健性(Robustness)
  - 悪意をもってAIに誤認識・誤判断を発生させようとする攻撃に対する耐性
- 説明可能性(Explainability)
  - AIが「なぜその答えを出したのか」が説明できる
  - 誤っていないかの確認、誤った答えを出したときの追究・修正に重要
- 公平性(Fairness)
  - AIが社会的・倫理的に見て公平な判断を行っている
- 透明性(Transparency)
  - さまざまな役割の利害関係者と情報を共有し、信頼を強化
- プライバシー(Privacy)
  - AIがプライバシーを守り、想定外の情報を引き出せない

## プライバシー保護技術

- 保護すべき情報を特定
  - 識別情報: ID、名前や電話番号
  - 履歴情報: 購買履歴、診療履歴
  - 要配慮情報: 人種、宗教、国籍、病名、犯罪歴
  - 統計情報: 病名別患者数、年代別年収
- 匿名化(仮名化、一般化)
  - 九州産業大学 ⇒ ①(仮名化)A大学 ②(一般化)私立大学
- 差分プライバシー
  - 統計量から個別の情報を推測しづらいようにノイズ追加

## プライバシー保護機械学習

- マルチパーティ計算 (Multi-party Computation)
  - 複数の関係者がそれぞれのデータを共有することなく、共同でデータの分析を行う技術
- 連合学習 (Federated learning)
  - 分散環境にある複数のデバイスでアルゴリズムの学習を実施し、学習した結果をすべての中央サーバーと共有
- 差分プライバシー (Differential privacy)
  - データにノイズを追加してデータ侵害を困難にする
- 準同型暗号 (Homomorphic encryption)
  - 暗号化したデータの状態で計算を行う技術

## バイアスの種類

- データに関するバイアス
  - データ生成時(測定、選択)
  - データ流通時(削除、匿名化)
  - アノテーション時(種類等を付加)
  - 前処理時(解析前のデータ変換)
- アルゴリズムに関するバイアス
  - アルゴリズムが予測・分類に有効な**識別変数**を選択するとき、人種や性別等が関係性のない目的(ローン審査、人事評価など)で使われると、バイアスが生じる可能性がある

## AIと人間との価値観整合

- AIシステムが人間の価値と基準に合わない方法で、自分の目標を達成する可能性がある
- そのため、AIシステムの動作に境界を設け、制約を与える方法が必要
- 価値観整合 (value alignment)
  - AIシステムに要求される価値と原則をモデル化し、それらを確実に守られるように設計

## アシロマの原則 Asilomar AI Principles



2017年、著名なAI研究者・開発者によりAI研究の将来を討議する会議において発表されたガイドライン。AIが人類全体の利益となるよう、倫理的問題、安全管理対策、研究の透明性などについて23の原則としてまとめた

## AIとの正しい付き合い方

- AIシステムが人間の知的活動の一部を、人間以上の精度、効率で担ってもらい、人間の生活をますます便利で快適にしてもらう
- 一方、AIは現段階で、頑健性、信頼性、人間の倫理観・価値観との整合性が保証されるものではない。
- これらを念頭におき、悪用・誤用にも気を付けて、AIと付き合いっていくことが大切
- また、敵対的サンプル等の存在は決して悪いことではなく、AI技術をさらに進化させるきっかけにもなる

## 参考文献

1. 責任あるAI:「AI倫理」戦略ハンドブック, 保科学世, 東洋経済新報社 (2021/8/20)
2. AIの法務と倫理, 古川直裕, 中央経済社 (2021/4/24)
3. データ解析におけるプライバシー保護, 佐久間 淳, 講談社 (2016/8/25)
4. トロッコ問題(Trolley problem)とは? Atmarkit AI・機械学習の用語辞典
5. Philosophy, Ethics and Safety of AI, Artificial Intelligence: A Modern Approach, pp.1032-1062, S. Russell, P. Norvig (著), Pearson Education Limited, 第4版 (2021/5/13)