消費者とAIのリスク

20240910

理化学研究所　革新知能統合研究センター
荒井ひろみ

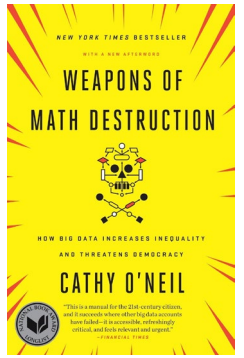
消費者にとってのAIのリスクの例

- AIの出力におけるバイアスや不公平
- AI利用についての理解
- データ提供者、サービスユーザーのプライバシー、安全性

AIのバイアスやブラックボックス化の懸念

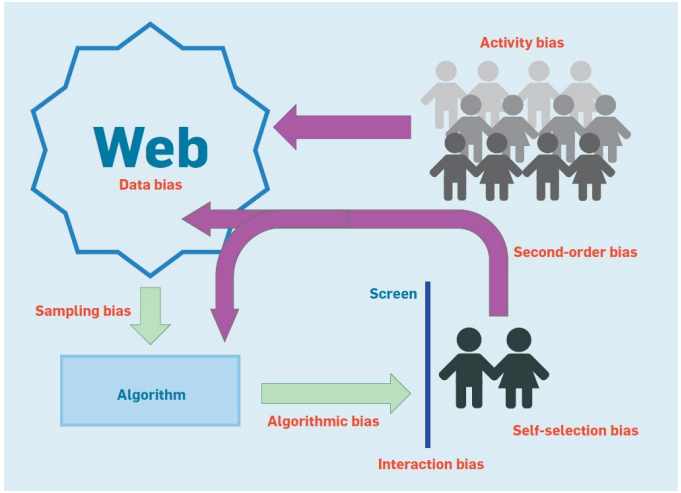
- Weapons of Math Destruction
  - 利潤最大化傾向の強い私企業が純粋に数学的な手法によって効率化を図ると多くの場合に公平性が失われるという指摘
  - 「数学破壊兵器」の特徴
    - 不透明性
    - 規模拡大特性
    - 有害性

O'neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.



- Bias on the web
  - 人間のバイアスとWeb内部のバイアスがWeb上のバイアスとなる

Baeza-Yates, R., & Murgai, L. (2024). Bias and the Web. *Introduction to Digital Humanism: A Textbook; Werthner, H., Ghezzi, C., Kramer, J., Nida-Rümelin, J., Nuseibeh, B., Prem, E., Stanger, A., Eds*, 435-462.



AIのバイアスの課題

- 実データから学習したAIは様々な理由からバイアスを含む場合がある
- AIが人種や性別などの属性に基づいた差別的な振る舞いをする問題
  - 結果の不公平
    - 精度や合格率が属性間で不平等
  - 社会的バイアス
    - ステレオタイプが反映された出力

例：顔識別における性能格差

- 黒人女性の顔識別率の低さ：Gender Shades
  - 訓練データにおける少数グループは不利な結果を得る

顔画像データ

学習モデル構築

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

顔識別API

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

例：画像生成における社会的バイアス

- 検索や生成においてステレオタイプが強化されるような結果や差別的な結果を出力

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., & Kersting, K. (2024). Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*, 1-21.

AIの公平さ、バイアスについての対策

- AIのバイアスの評価
  - バイアスを認識、定義し、測定する
    - 例：ベンチマークデータの作成、評価
- 公平なAIの作成
  - より公平なAIを学習できるように訓練データを調整する
  - モデルの学習過程においてより公平な出力をするような学習方法を適用
  - 学習済モデルのバイアスを調整する

AIの公平性、バイアス評価における課題

- 顔認識におけるベンチマーク利用における課題
  - どのような公平性基準を選択するか
    - 属性や評価指標の選択
  - ベンチマークに利用される顔画像提供者のプライバシー
  - 一般性，代表性の確保

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020, February). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 145-151).

- データ作成時のアノテーションにおける課題
  - 人種差別的バイアスの懸念
    - アノテーターによっては黒人英語をより攻撃的なものと分類するという指摘も

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and philosophy*, *36*, 419-468.

AI利用におけるユーザーの理解に関わる課題

- パーソナルデータ利用についての説明の課題
  - データプラクティスが複雑であるほど長文である
  - 専門用語を多く含む

- AIの振る舞いの理解についての課題
  - どのユーザーに何を説明するべきか
  - 説明が適切に提供されているか

専門用語の難解さや誤った期待

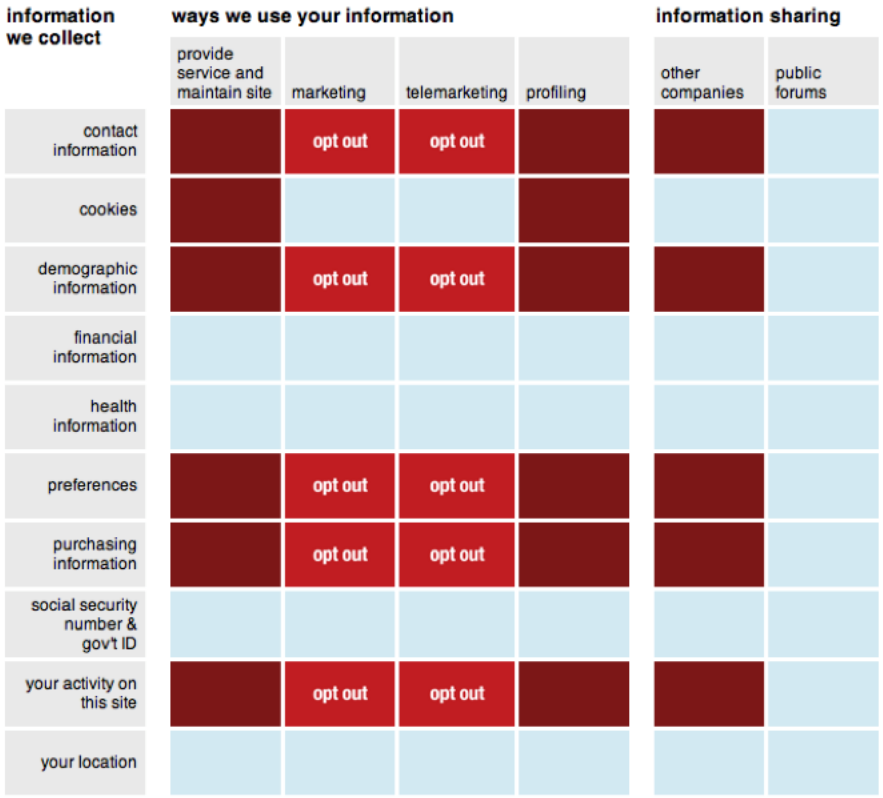
- プライバシーポリシーに使用される技術用語に対する誤解
  - 362名の日本人にアンケート調査、16個の技術用語のうち参加者の正答率が50%を超える用語は存在せず
  - プライバシーポリシーにおける技術用語の使用が日本のユーザの同意率に影響をもたらす
- 個人情報保護法に対しても誤った期待を持つ参加者が多い
- 米国での同様の調査とはやや結果が異なっており、国や文化によって異なる支援を実施する必要性

金森祥子, 池田美穂, 亀石久美子, & 長谷川彩子. (2023). プライバシーポリシーに使用される技術用語および個人情報保護法に対するユーザの理解度の調査. *コンピュータセキュリティシンポジウム 2023 論文集*, 1012-1019.

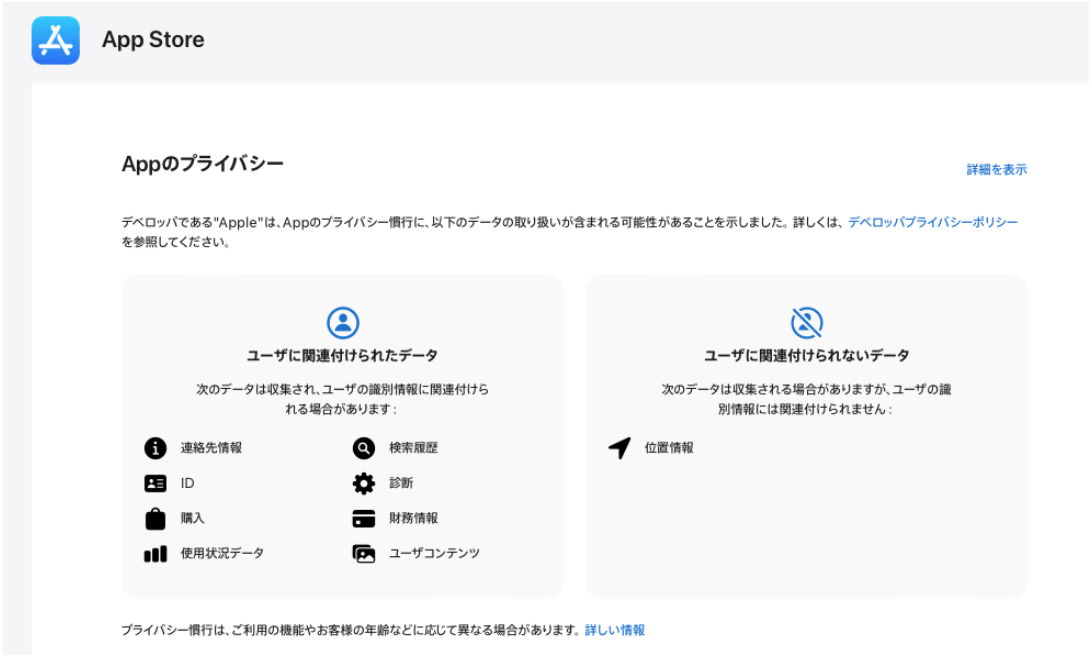
情報の標準化、ラベルやアイコンの利用

- 長文テキストよりわかりやすい、安全性の評価、比較がしやすい

P.G. Kelley, L.J. Cesca, J. Bresee, and L.F. Cranor. Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach. CHI 2010.



- Appleのアプリストアなどで利用



ユーザーフレンドリーにしても読むとは限らない

- 米国の大手医療保険会社が開発したチャットボットの同意フローについてのケーススタディー

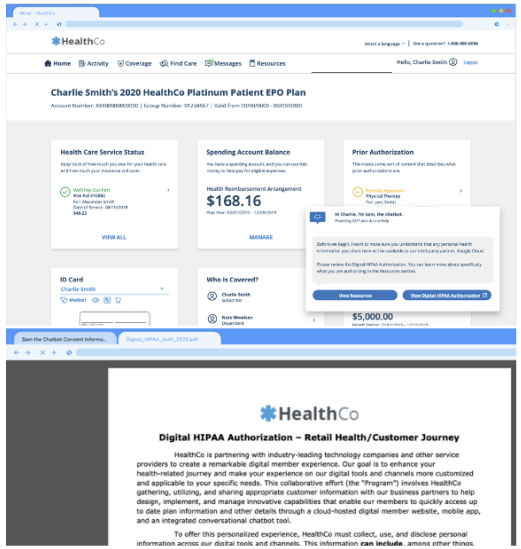


Fig. 1. Screenshots from Prototype 1 including the HIPAA authorization shown as a PDF.

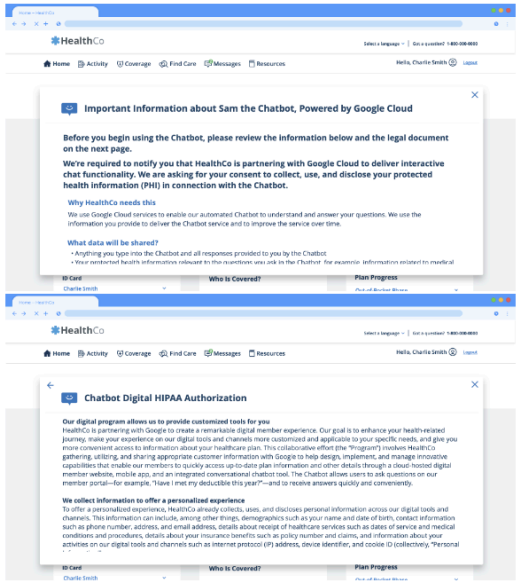


Fig. 2. Screenshots from Prototype 2: "Resources" summary and HIPAA Authorization documents, shown in modal window.

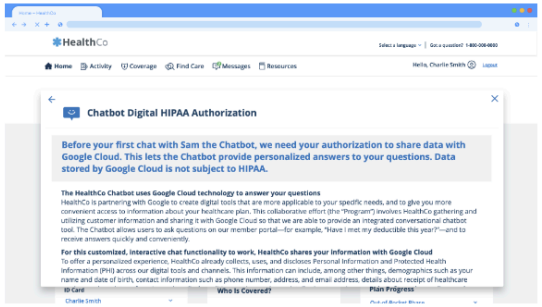


Fig. 3. Screenshot from Prototype 3: HIPAA Authorization document in modal window.

Pearman, S., Young, E., & Cranor, L. F. (2022). User-friendly yet rarely read: A case study on the redesign of an online HIPAA authorization. *Proceedings on Privacy Enhancing Technologies*, *2022*(3).

- ユーザーは説明PDFのリンクへ飛ばない、要約を置くと使いやすいと感じるが内容の理解への効果は限定的、誤った理解をするユーザーも
- 技術的知識の欠如、HIPPAへの過剰なプライバシー保護期待などが原因と考察される
  - Chatbotの中身が人だと思っている人がいた(8.5%)
  - 細かい文字が多いのでなにか悪いことがおきていると思う人も

AIモデルの説明の必要性

- 説明可能AI
  - 説明可能AIのニーズ
    - 認識・倫理上の正当化、AIを改善、未知の脆弱性や欠陥を発見し修正・制御するため、未発見の有効な戦略や法則を発見するため

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

- モデルの適切な選定、運用が重要
  - 多くの機械学習モデルが公開されている
  - 機械学習モデルの性質、性能、振る舞い、訓練データのドキュメント化は特にユーザーへの影響が大きい領域で重要

説明可能AIの利点と課題

- 利点
  - ユーザーに理解可能な形で複雑なモデルの概要や、モデルの判断根拠を提示することで、ニーズを満たすことができる

- 課題
  - 都合の良い説明がなされる可能性
    - モデルの概要を説明する際に、実際より公平なモデルであるように説明をする

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. 2019. Fairwashing: the risk of rationalization. In International Conference on Machine Learning (pp. 161-170). PMLR.

AIモデルに関するドキュメント化

- データステートメント

Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." ACL 2018.

  – 言語データにはサポートしていない集団の存在，収集データの偏りの問題がある
  – 特定の母集団で機能しない，偏りの再現と強化　などの科学的，倫理的な問題
  – データセットの代表集団がわかればその利用の際に役立つ

- モデルカード

Mitchell, Margaret, et al. "Model cards for model reporting." *Proceedings of the conference on fairness, accountability, and transparency*. 2019.

  – 想定されるユースケースでの利用のため、パフォーマンス特性の共有が重要
  – 機械学習モデルの公平性、包括性などの課題
  – モデルに関する記載の標準的なフレームワークの提案

モデルカード

- モデルリポジトリでの実践

https://huggingface.co/learn/nlp-course/ja/chapter4/4

モデルリポジトリの課題

- モデルカード、データカードの普及
  - データセットやバイアスへの言及は限定的という調査結果も

Pepe, F., Nardone, V., Mastropaolo, A., Bavota, G., Canfora, G., & Di Penta, M. (2024, April). How do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension* (pp. 370-381).

- 透明性向上とセキュリティ/プライバシーリスクのトレードオフ
- プライバシー、セキュリティー上のリスク
  - Hugging faceプラットフォーム上のAIモデルのうち脆弱性のあるものが一定割合存在する指摘

Kathikar, A., Nair, A., Lazarine, B., Sachdeva, A., & Samtani, S. (2023, October). Assessing the vulnerabilities of the open-source artificial intelligence (AI) landscape: A large-scale analysis of the Hugging Face platform. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 1-6). IEEE.

まとめ

- AIの公平性、バイアスは一般ユーザーにも影響を与えうるAIの重要課題であり、様々なケースでの課題や対策が議論されている。公平性の評価方法にはまだ課題もあり、適切な指標や方法を取ることが重要。
- プライバシーポリシーにおける説明は改善がなされているが、課題は残る。
- AIのブラックボックス化への対策として説明可能AIのアプローチがある。
- AIの適切な利用のためにモデルカード、データカードなどのドキュメント化が提案されており、普及しつつある。
- 公開されたモデルのプライバシー、セキュリティー上のリスクが指摘されており、利用時には配慮する必要がある。