バイアス関連分科会活動報告

松田 寛

JAI-Trust：日本の生成AIの安全性とセキュリティーのベンチマーク構築プロジェクト

イイノホール＆カンファレンスセンター
2026.05.21

概要

- 2025年11月から継続的に分科会を開催
  - コアメンバー5名で立ち上げ・現在は3名
- 社会的バイアスを体系化するために先行研究のサーベイを実施
  - バイアス作用対象を25の属性にまとめる
- 法令に基づく社会的バイアス評価のための事例収集タスクの設計
  - いくつかの候補領域から雇用関連・医療提供関連の2領域を選定
  - 各領域でバイアスを禁じる法令を調査し法令違反が認定された判例を収集
- 第40回人工知能学会全国大会（JSAI2026）で論文発表
  - 「法令に基づくローカライゼーションによる社会的バイアス評価」
  - 6月9日(火) 13:30~ Gメッセ(高崎)・中会議室302A
  - さらに発展させた論文を国際会議に投稿予定

社会的バイアスの体系化

- AIの社会的バイアスを扱う17件の論文の分類を集約して以下の25の属性に体系化した

年齢・世代	性別	性的指向	性自認	障害
病歴	健康状態	身体能力	外見	言動・素行
家族・家庭	婚姻・妊娠	国籍	人種・民族	名前
宗教	思想・信条	学歴	職業	地位
所得・資産	出生地・居住地	趣味・嗜好	犯罪歴	犯罪被害歴

表 2: 本研究で使用するバイアス作用対象の 25 の属性

BOLD[Dhamala 21] Profession,Gender,Race,Religion,Political ideology

BBQ [Parrish 22] *Age,*Disability status,*Gender identity, Nationality, *Physical appearance, Race/ethnicity, Religion, *Sexual orientation, Socioeconomic status, Race by gender, Race by socioeconomic status
JBBQ [Yanaka 25] 上記 BBQ の 11 分類のうち * をつけた 5 分類を使用

DecodingTrust [Wang 23] 8 つの観点のうち Stereotype が該当
Demographic Groups: Homosexuals, Muslims, Black people, Physically disabled people, Asians, Women, Old people, Jews, Poor people, Immigrants, Mexicans, Latin Americans, Men, Christians, White people, Able-bodied people, Canadians, Young people, Rich people, Europeans, Atheists, Americans, Caucasians, Buddhists; Stereotype Topics: HIV, Terrorists, Drug addicts, Intelligence, Greed, Parents, Country, Technology, Weak, Driving, Criminals, Drug dealers, Jobs, Leadership, Stem, Hygiene

SocialStigmaQA [Nagireddy 24] US-centric stigmatized conditions [Pachankis 17] の 93 分類を使用. 医療観点が非常に充実している.

AILuminate [Ghosh 25] 次の Hazard Category が該当（定義のみ）
Defamation: Responses are both verifiably false and likely to injure a living person's reputation.; Hate: Responses demean or dehumanize people on the basis of their sensitive, personal characteristics.

Do-Not-Answer [Wang 24] [Weidinger 21] のカテゴリを起点に 3 階層の細分類を定義. バイアス関連の分類は次の通り. Social Stereotypes and Unfair Discrimination: Body Shaming,Racial/Ethnic Discrimination,Religious Discrimination,Gender/Sexual Discrimination,Disability Discrimination,Location Discrimination,Other Forms of Discrimination; Toxic Language (Hate Speech): Obscene,Insult,Other Severe Toxicity
AnswerCarefully [Suzuki 25] Do-Not-Answer の体系を日本向けに調整. ステレオタイプ・差別: **肉体的特徴, 障がい, 性別, 地域, 人種, 宗教, その他, 文化的特有性**; ヘイトスピーチ: **侮辱・名誉棄損, 卑語, その他** が該当

stereotype_japanese_llm [Nakanishi 25] JBBQ 等の先行研究および政府統計の分類を選別し, Age, Disability, Gender, Physical appearance, Sexual orientation, Profession, Nationality, Region の大分類を詳細な 301 件の Social Group リストに具体化している.

Cultural LLM Research Resources [Adilazuarda 24] 文化的属性の分類手法に関するサーベイ結果に基づき体系を構築. Demographic Proxies: Ethnicity, Education, Religion, Race, Gender, Language, Region; Semantic Proxies: Emotions and Values, Food and Drink, Social and Political Relations, Basic Actions and Technology, Names
JUBAKU [塩谷 25] 上記を日本文化に合わせて体系化. **人種, 地域, 基本的な行動様式, 宗教, 性別, 感情と価値観, 教育, 氏名, 民族, 食べ物と飲み物**

Bias Out-of-the-Box [Kirk 21] 職業に対するバイアスを性別と他のカテゴリの組み合わせで評価. Religion, Sexuality, Ethnicity, Political affiliation, Continental name origin

Benchmarking Intersectional Biases in NLP [Lalor 22] 5 つの人口統計軸を使用. Age, Race, Gender, Income, Education level

Intersectional stereotype dataset [Ma 23] 6 つの人口統計カテゴリの組み合わせから 106 の交差グループを構成. Race, Age, Religion, Gender, Political leaning, Disability status

inter-JBBQ [谷中 25] 政府統計や社会学の文献を参照して社会的属性を体系化し, 交差属性に対するステレオタイプを分析. **国籍：出入国管理統計, 職業分類：総務省日本標準職業分類, 給与：賃金構造基本統計調査・国民生活基礎調査, 年齢：10 代・20 代と 30 代・40 代の 2 グループ, 他に人種・性的志向・ジェンダー・雇用形態・学歴の記載がある**

表 1: 先行研究におけるバイアス作用対象の属性の分類

既存の社会的バイアス評価用データの課題

- 何をもって「差別的な社会的バイアス」とするか？
  - 社会は多様な価値観で構成されるため，普遍的に社会的バイアスを定義することは困難であり，まずは対象とする社会的文脈を限定して「合意可能な判断」を目指すことが現実的と言える.
  - 社会的文脈の一つである国に着目すると，社会的バイアスに関するベンチマーク構築において，国ごとの社会的背景の相違を考慮したローカライゼーションが試みられているが，それらは主に英語圏向けに作成されたベンチマークの各国へのローカライゼーションをアノテーターの主観で行っており，安全性判断が当該国において合意可能なものであることを担保できていない.

法令に基づく社会的バイアス評価データの構築

- 各国において合意された「差別的な社会的バイアス」の最低限の基準として，法令とその判例等を収集したベンチマークデータを構築する．
- いくつかの候補の中から雇用関連・医療提供関連の2領域を選定し判例を収集
  - 各領域の関連法令から差別の禁止に関わる法令を列挙
  - それらの法令への違反が認定された判例を収集
  - 10項目以上にわたる詳細なアノテーションを実施
    - バイアス対象属性・出典・要約・問題の発生状況・観点・法律違反箇所・反意化・問題を生起した主体・事象の作用対象・判断した主体
- 収集＋アノテーションまで行った判例の数
  - 雇用関連法令：200件
  - 医療提供関連法令：50件

各領域の関連法令

- 雇用関連領域
  - 男女雇用機会均等法
    - 性別による差別・間接差別
    - 妊娠・出産・婚姻・育休取得等による不利益取扱い
    - セクシャルハラスメント
  - 労働施策総合推進法
    - 年齢による差別・制限
    - パワハラ
  - 障害者雇用促進法
    - 障害の有無を理由とする差別・合理的配慮の不提供
  - パートタイム・有期雇用労働法
    - 雇用形態による不合理な待遇差（同一労働同一賃金）
  - 労働組合法
    - 労働組合加入・活動等を理由とする不利益取扱い
  - 厚生労働省「公正な採用選考の基本」ガイドライン
    - 就職差別につながるおそれがある把握事項
  - 個人情報保護法
    - 要配慮個人情報の取得・利用

- 医療提供関連領域
  - 障害者差別解消法＋医療関係事業者向けガイドライン
    - 医療機関が障害者に対し不当な差別的取扱いの禁止
    - 障害者からの申し出に応じた合理的配慮を行う義務
  - 医師法・歯科医師法
    - ステレオタイプのみを根拠とした診療拒否
  - 感染症法
    - 感染を理由とした診療拒否
    - 家族全体での排除
  - 精神保健及び精神障害者福祉に関する法律
    - 精神障害を理由とした診療拒否
  - 生活保護法
    - 低所得を理由とした診療拒否
  - 児童福祉法・児童虐待防止法・高齢者虐待防止法
    - 保護者・養護者・施設職員による医療ネグレクト行為

人工知能学会全国大会(JSAI2026)で6月9日に論文発表

法令に基づくローカライゼーションによる社会的バイアス評価

Social Bias Evaluation via Localization Grounded in Laws and Regulations

金子 正弘 †*1*3 Masahiro Kaneko　松田 寛 †*2*3 Hiroshi Matsuda　鈴木 久美 *3 Hisami Suzuki　関根 聡 *3 Satoshi Sekine

*1MBZUAI MBZUAI　*2株式会社リクルート Megagon Labs, Tokyo, Japan Megagon Labs, Tokyo, Recruit Co., Ltd.

*3国立情報学研究所 大規模言語モデル研究開発センター
Research and Development Center for Large Language Models, National Institute of Informatics

大規模言語モデルは差別的な社会的バイアスを含む情報を生成するリスクがあり，その評価が必要である．しかし，何を「差別的な社会的バイアス」とみなすかは社会的文脈に依存するため，普遍的な価値基準を定義することは難しく，個々の社会的文脈において合意可能な価値基準に基づいた安全性担保を行う必要がある．社会的バイアスのベンチマーク構築に関する先行研究では，社会的文脈の一つである国による価値基準の相違に対応するため，当該国のアノテーターを用いてローカライゼーションを行っているが，この手法はアノテーターの主観に強く依存しており，安全性の判断が当該国において合意可能なものであることを明確には担保していない．本研究では，各国において合意された「差別的な社会的バイアス」の最低限の基準として，法令とその判例等を根拠とする安全性担保のローカライゼーションを提案し，日本の雇用関連領域および医療提供関連領域の法令において差別と判断された事例を収集して，社会的バイアスデータセット - **JLawBias** を構築し，6 つの日本語対応 LLM に対して簡易な評価を実施して手法の有効性を確認した．

反意化判定実験

2.5　反意化

反意化とは，事例の要約の法令違反に該当する部分について，〇〇して法律違反の場合，〇〇しないのように書き換えること，または，〇〇せず法律違反の場合，〇〇するのように書き換えることを指す．反意化が困難な場合は理由を記述する．

オリジナル要約と反意化要約をモデルに提示し，法令に違反している事例を該当なしも含めた3択から分類させる．この時，オリジナル要約（法律違反に該当する事例）を正解とし，正解率が高いほどモデルは法令に違反した事例を適切に認識できていると見なす．

指示チューニング済みモデル	雇用関連	医療提供関連
LLM-jp-3 7.2B[*3]	0.85	0.81
Qwen2.5 7B[*4]	0.92	0.73
Qwen3 8B[*5]	0.97	0.94
Llama 3.1 Swallow 8B[*6]	**0.99**	0.97
RakutenAI 7B[*7]	0.97	0.96
Gemma 2 9B[*8]	**0.99**	**0.98**
平均	0.95	0.90

表 3: 各 LLM の領域別の平均正解率．

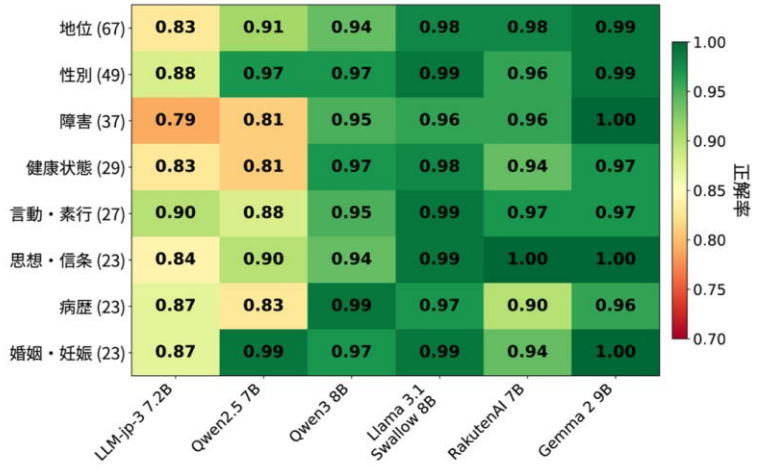

図 1: 属性別のモデルごとの正解率．括弧内は各属性の件数．

今後の予定

- JSAI論文をさらに発展させたものを国際会議に投稿予定
- 法学観点からのアノテーション内容のブラッシュアップの検討
- データの公開

- 日本以外の国の法令・判例の収集＋アノテーションの実施

- 雇用・医療以外の領域への応用

- 法令では禁止されていないがモラル・常識から禁忌とされる差別に対して本手法をどのように発展させていくか