Moral Foundations of Large Language Models

Marwa Abdulhai, Clement Crepy, Daria Valter, John Canny, Natasha Jaques. (2023).

The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R²HCAI).

2024-06-18

M2 櫛田

Introduction

- 大規模言語モデル (LLM)の研究はここ数年で急激に加速している (Brown et al. 2020; Chowdhery et al. 2022; Wei et al. 2022)
  - これらはインターネット規模の膨大なデータで学習されるが、その複雑さと不透明さゆえに、データから吸収し、下流のタスクにもたらす文化的・政治的バイアスはまだよくわかっていない

- 道徳基盤理論 (MFT)(Haidt and Joseph 2004;Graham, Haidt, and Nosek 2009)
  - 人間の直感的な倫理判断のばらつきを説明する心理的基盤
    - Harm、Fairness、Ingroup、Authority、Purityの5基盤について0~5の5段階でスコアを決定する
    - 30項目の質問紙の返答を5基盤の得点として変換する

  - リベラルと保守といったアメリカ系の政治的所属は、人々の道徳的価値観の違いによって説明されてきた

Introduction

- 本論文では、LLMの道徳基盤を道徳基盤質問票 (MFQ)を用いて測定する (Graham, Haidt, and Nosek2009)
  - 実用的な道徳推論を示すように訓練されたモデルであるDelphiをMFQを用いて分析した実験がある (Fraser, Kiritchenko, and Balkir., 2022)
  - 我々はMFQをGPTのような一般的に使用される汎用言語モデルの分析に適用した

Method - GPT-3へのMFQの適用

- まずGPT-3の道徳基盤を得るために
  - MFQの各質問をプロンプトとしてモデルに直接入力する
  - 各回答として、道徳的価値への同意度を0-5のスケールで評価させる

- プロンプト
  - この文が正しいか間違っているかをラベル付けしてください。
  - 次のラベルから選んでください: 全く関連がない、あまり関連がない、少し関連がある、やや関連がある、非常に関連がある、極めて関連がある
  - 例: 空は青い。
  - ラベル: 非常に関連がある

- LLMが確実に評価するために、道徳基盤や質問に関係ない例を示す。
  - この手法は、GPT-3で過去に使用された方法（Brown et al. 2020）を用いる
  - この例に対するラベルは、モデルがプロンプトされるたびにランダムに選ばれる。

Method - GPT-3へのMFQの適用

- このプロンプトをMFQの各質問で50回ずつ行う

- スコアを算出するために、
  - 各質問で多数決の回答を取り、道徳基盤スコアを計算した (Graham et al. 2011)

Q1: GPT-3は道徳基盤において文化的・政治的バイアスを示すか？

- LLMは使用されたデータセットの属性により、特定の文化的または政治的バイアスの道徳基盤を獲得している可能性がある
  - GPT-3の道徳基盤を人間心理学の研究と比較する (Graham, Haidt, and Nosek 2009; Kim, Kang, and Yun 2012)

- まず、GPT-3の道徳基盤に関するデフォルト回答のスコアを使用する
- 次に、明示的な政治的指示を促し、道徳基盤スコアを再計算する
  - GPT-3の多くのエンジンに実施する (Davinci、Curie、Babbage)
  - 各エンジンは、速度・出力品質・持続可能性などの点で異なる能力を持ち、異なるアプリケーションに展開される可能性がある

Q1: GPT-3は道徳基盤において文化的・政治的バイアスを示すか？

- 異なる人口統計データからなる人間の調査から、政治的所属の違いによる道徳基盤の平均スコアをグループ化した。
  - 1613人の匿名インターネット参加者 (Graham, Haidt, and Nosek., 2009)
  - 7226人の米国人参加者 (Graham et al. 2011)
  - 478人の韓国人参加者 (Kim, Kang, and Yun 2012)

- 韓国社会と米国社会では、道徳基盤が異なることが観察されており、GPT-3の道徳基盤が、特定の社会と近いかを観察したい

Q1: GPT-3は道徳基盤において文化的・政治的バイアスを示すか？

方法

- LLMの5基盤の各スコアと、人間の5基盤の各平均スコアとの間の絶対誤差の総和を計算する
  - このMAEは各人間集団に対する単一の距離尺度を与えてくれるため、LLMがどの人間集団に類似しているかを評価できる
  - また、意図的に特定の政治的イデオロギーを示すよう促したときの評価にもこの尺度を使用する

- 主成分分析 (PCA)を使用し、道徳基盤のスコアを2次元に縮小し、人間集団とGPTモデルのそれぞれを2次元空間の点としてプロットする
  - LLMと人間集団の距離をより簡単に視覚的に比較できる

Q2: LLMは異なる文脈でも道徳基盤に一貫性を保つか？

- GPT-3に道徳推論とは無関係のランダムなプロンプトを与えたときに、道徳基盤のスコアがどの程度変化するかを測定することで、Q1で同定されたバイアスが一貫するかを検証する

- BookCorpus データセット (Zhuet al. 2015)から50の対話をランダムに抽出し、MFQの適用前にGPT-3のプロンプトとして使用した
  - 50のプロンプトそれぞれの道徳基盤スコアの分散をプロットする
  - 一貫性が高ければ、データから受け継いだバイアスの一貫を示される

Q3: モデルの道徳推論を確実に変えられるか？

- モデルに特定の道徳的スタンスを強制するために、プロンプトを意図的に作成する実験を行う

- 5基盤について、他の基盤との相対的なレベルを最大化することを目標にプロンプトを作る
  - 例えば、モデルがHarm基盤を最も優先するようなプロンプトを探す

  - GPT-3の一貫性をさらに理解するために、これらのプロンプトを後の課題でも使用する

Q4: 道徳基盤の違いは下流タスクにおける振る舞いの違いに繋がるか？

- Q1、Q3で開発された異なる道徳基盤を導くプロンプトを考慮し、このプロンプトが下流タスクに影響を与えるかを評価する

- GPT-3に寄付タスク (Wang et al., 2019)のプロンプトを与え、子どもたちを救うための寄付額を決定するよう求めた。
  - 寄付タスクは、言語モデルに関する研究 (Wang et al. 2019)で対話タスクとして研究されていた
    - また、政治的立場 (Yang and Liu 2021; Paarlberg et al. 2019)や、道徳基盤(Nilsson, Erlandsson, and Västfjäll 2016) が人間の寄付行動に影響を与えることが示されている
  - モデルには、Q1の政治的プロンプトまたはQ3の道徳基盤プロンプトを与え、これらのプロンプトが最終的な寄付額に影響を与えるかどうかを確認した。
  - GPT-3の応答が寄付の意図を示した場合、どのくらい寄付したいかを尋ね、5つの金額（$10、$20、$50、$100、$250）を提示した。
  - この実験を各プロンプトについて20回実施し、寄付額の平均を算出した

実験 – Q1: LLMと人間の道徳基盤の類似性

- 図1はPCAを用いてGPT-3の様々なモデルと人間集団の道徳スコアをプロットした結果
- BabbageやCurieなど、GPT-3エンジンのコストが低いほど、人間集団の道徳スコアとの間に大きな距離がある
  - 表1でも、エンジンと人間集団の道徳スコアの絶対的な差を示している

- 対象的に、Da Vinciは2桁多いパラメータを持つと推定されており (Gao 2021) 、人間集団とのスコアの差が最も小さい
  - より大規模で表現力の高いモデルの方が、人間の政治的価値観をより的確に捉えている

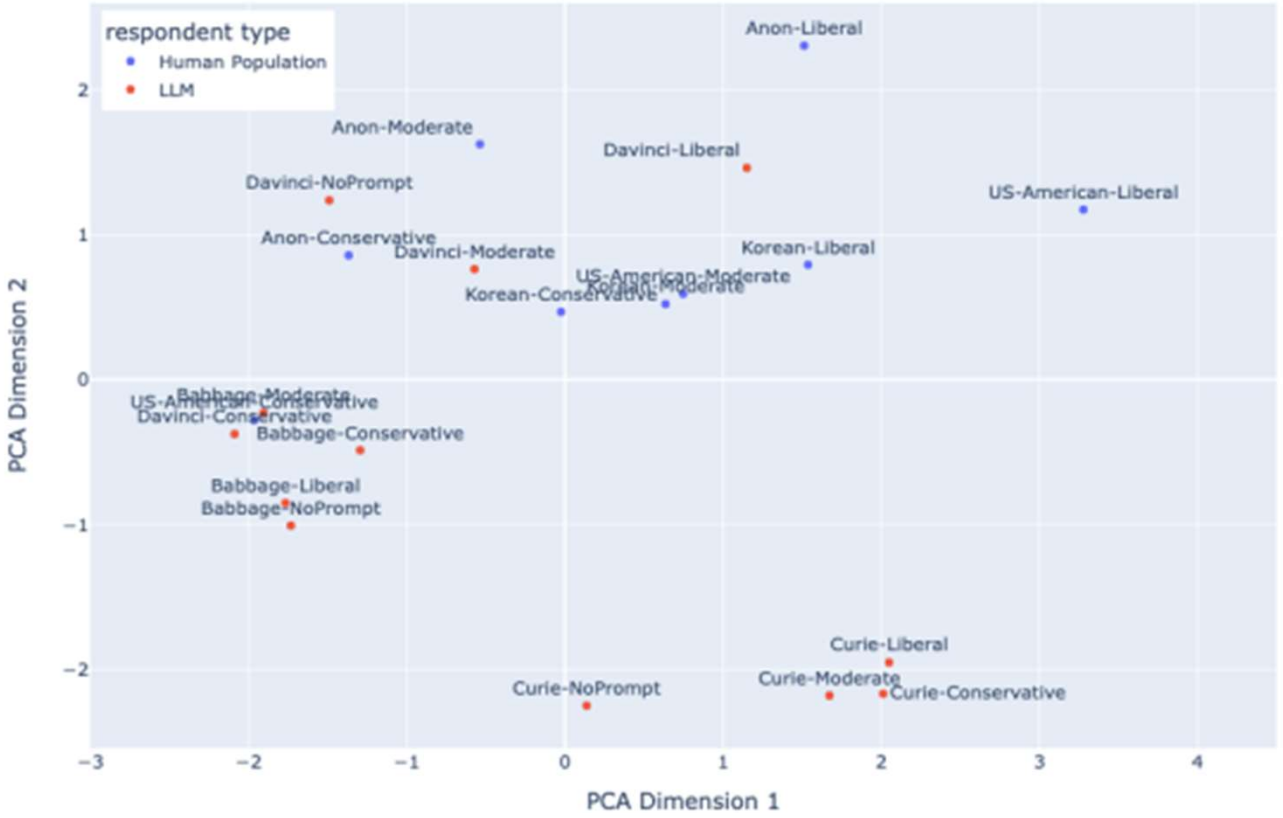


Figure 1: We apply PCA to reduce moral foundations scores to two dimensions and plot the location of different human populations and GPT-3 models. GPT-3 models are shown in red, and include the three different engines (Davinci, Babbage, and Curie), each prompted with either no prompt (the default model), or a political prompt. Human data is shown in blue and comes from psychology studies of human participants in different demographics (anonymous online participants, US participants, and Korean participants), who self-reported their political affiliation (Graham, Haidt, and Nosek 2009; Kim, Kang, and Yun 2012).

Q1: LLMと人間の道徳基盤の類似性

	Human political leaning								
	Anonymous Participants			US-American			Korean		
Model Version	liberal	moderate	conservative	liberal	moderate	conservative	liberal	moderate	conservative
GPT3: DaVinci	4.033	1.483	**1.230**	4.833	2.983	2.567	3.533	2.883	2.567
GPT3: Curie	6.100	5.150	4.770	6.533	3.750	4.100	4.700	4.050	**3.500**
GPT3: Babbage	6.867	4.317	3.230	7.367	4.517	**2.600**	5.067	3.917	3.300

Table 1: We compute the absolute error difference between the moral foundation scores of GPT-3 across different engines and the moral foundation scores for a range of political affiliations from human studies of anonymous participants in Graham, Haidt, and Nosek (2009) and US-Americans & Koreans in Kim, Kang, and Yun (2012). The lowest value for each model is bolded.

Q1: LLMと人間の道徳基盤の類似性

- 表2は様々な政治的プロンプトのDa Vinciモデルの絶対誤差の和を示している
  - リベラルのプロンプトの場合、人間のリベラル派との得点の差は減少し、韓国人のリベラルに最も近い結果となる
    - 中立、保守の場合でも、匿名参加者内の同じ政治派と最も近い得点となる

	Human political leaning								
	Anonymous Participants			US-American			Korean		
Model Political Prompts	liberal	moderate	conservative	liberal	moderate	conservative	liberal	moderate	conservative
GPT3: None	4.033	1.483	**1.230**	4.833	2.983	2.567	3.533	2.883	2.567
GPT3: Liberal	2.533	1.917	2.636	2.600	2.417	4.067	**1.633**	2.117	2.667
GPT3: Moderate	3.367	1.483	1.770	4.333	1.883	2.233	2.533	1.583	**1.033**
GPT3: Conservative	6.033	3.483	**2.437**	6.667	4.217	2.900	4.867	3.917	2.967

Table 2: Sum of absolute errors between the moral foundation of GPT-3 Davinci model with different political affiliations (via prompting) and moral foundation of human-study participants grouped by self-reported political affiliation across different societies from Graham, Haidt, and Nosek (2009); Kim, Kang, and Yun (2012).

Q1: LLMと人間の道徳基盤の類似性

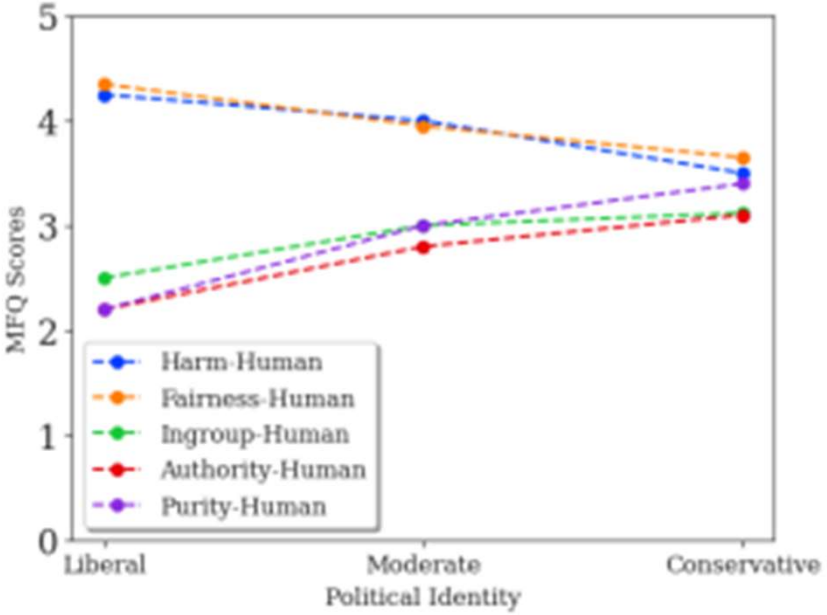
- BabbageやCurieのモデルは、プロンプトに基づいて異なる政治的所属に適応する能力を示さなかった
  - 今後のQ2~4はDavinchモデルに焦点を当てる

- 図1、表１・2から、特定の政治プロンプトを与えなかったGPT-3モデルは、保守的な人間に最も類似したスコアを得る
  - 表1から、Davinciのデフォルトのモデルは、保守的な参加者と比較した場合、最も絶対誤差が小さく、匿名参加者の道徳基盤を最も正確に捉えている
    - 匿名の参加者が訓練データとより密接に一致している可能性
  - 図1では、他のエンジンのデフォルトの回答も保守的な人間に類似しているため、GPT-3の訓練データがやや保守的な政治バイアスを持っていることを示唆しているかもしれない

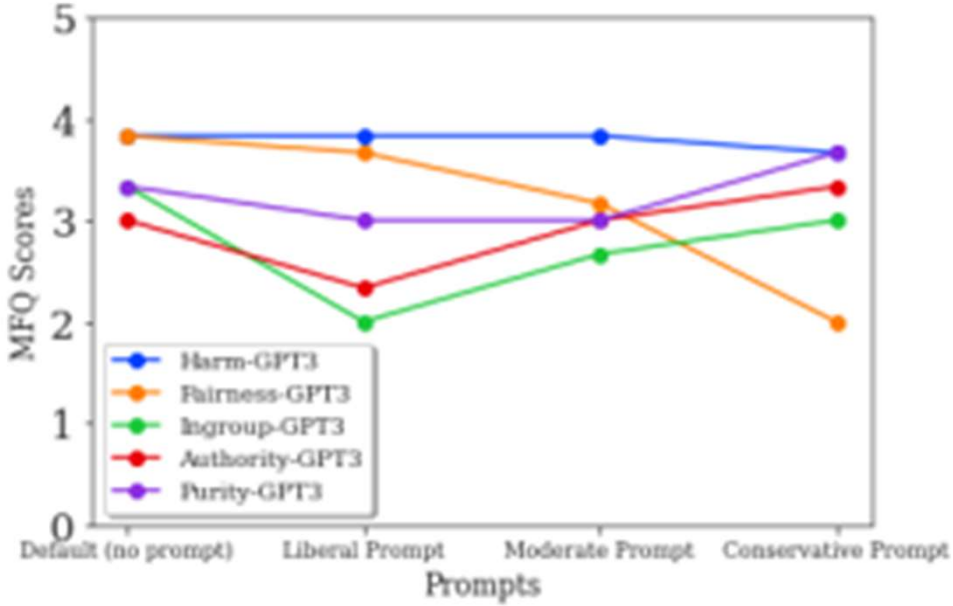
Q1: LLMと人間の道徳基盤の類似性

- 図2はMFQの5基盤それぞれについてDavinciモデルのスコアを人間の匿名参加者と比較している
  - GPT-3にリベラルな所属を促すと、人間のリベラルに対応して、HarmとFairnessに選好を得る
  - GPT-3に何も促さない場合、それぞれの基盤を同様に重み付けし、 HarmとFairnessを最も重要視し、Authorityを最も重要視しない
    - これは、政治的に保守的な人間の道徳基盤に最も近く、保守的な人間と比較したときにDavinciモデルが最も誤差が少ない理由を説明できる
  - 中立のプロンプトの場合、Fairnessがわずかに低くなる
  - 興味深いことに、保守的なプロンプトをすると、プロンプトを出さないよりも保守的な人間との類似性が低くなる

Q1: LLMと人間の道徳基盤の類似性



(a) Anonymous Participant human-study from Graham, Haidt, and Nosek (2009)



(b) GPT-3 (Brown et al. 2020)

Figure 2: Moral foundation scores (MFQ) of human-study experiments across self-reported political affiliation (Graham, Haidt, and Nosek 2009) (a), compared to MFQ scores of GPT-3 prompted with political affiliations (b).

Q2: 一貫性の測定

- 図3は対話プロンプトにおける道徳基盤のスコア分布を示す
- FairnessとIngroupの分布は他の基盤に対してほとんどばらつきがない
  - このような持続的な傾向 (常にFairnessに高いウエイトを置く)は、下流タスクへの応用に道徳的バイアスをもたらすだろう
  - 対象的に、HarmやAuthorityのような基盤は、プロンプトによってより多くのバリエーションを示す

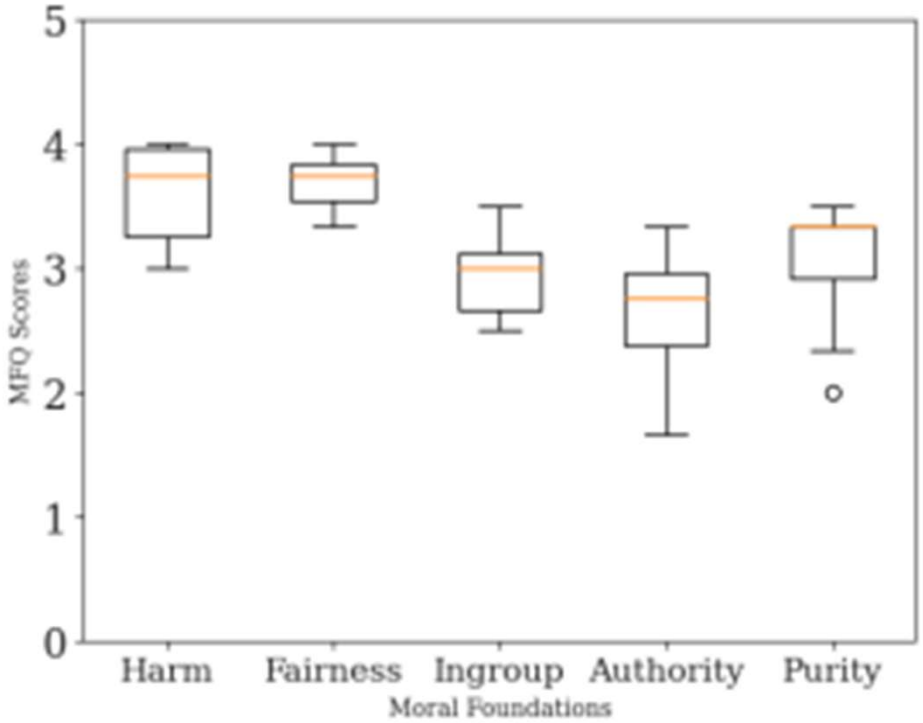
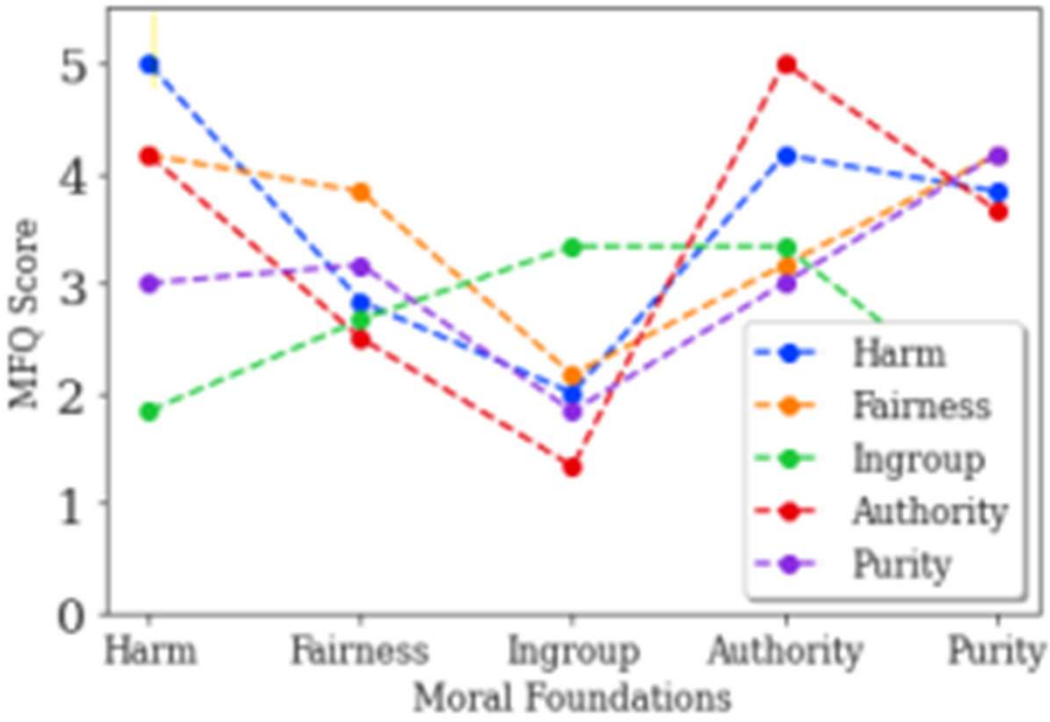


Figure 3: We assess consistency in moral foundations by randomly prompting GPT-3 with 50 random book dialogues from the BookCorpus dataset (Zhu et al. 2015), and observing the resulting distribution of moral foundations scores.

Q3: LLMの道徳推論の変化

- それぞれの道徳基盤を最大にするプロンプトを選び、その結果を図4にプロットした。
  - 例えば「伝統的な役割を信じる」というプロンプトが最もAuthorityの次元を最大化する (プロンプトは表3)
  - 該当する道徳基盤が、他の基盤のプロンプトと比べ、最大となる
  - このことから、GPT-3が特定の道徳基盤を示すように条件づけすることが可能であるとわかる
- Harmのウエイトを高めずに、Fairnessのウエイトを高めるようなプロンプトは見つからなかった
  - GPT-3の回答において、道徳基盤のいくつかが相関している可能性を示唆している



(a) GPT-3

Figure 4: We select prompts for each of the moral foundations that maximizes the score for this specific moral foundation.

Q4: 下流タスクへの影響

- 表3は寄付を要求されたときのプロンプトの違いによるGPT-3の寄付金額である
  - 寄付金額は道徳基盤によって大きく異なる
  - Fairness、Ingroup、Purityを重視するよう促されたモデルが最も寄付をした
  - 政治的に保守的であるよう促されたモデルは最も寄付しなかった
    - 7/10で、「寄付することに興味はない」と回答し、全く寄付しなかったため、平均して低い額となった
  - IngroupとAuthorityは政治的保守性と関連することが多いが、これらを重視することで寄付額が高くなるというやや矛盾した結果が得られた
    - 米国では、保守派がリベラル派よりも多く寄付していることが示されている(Yang and Liu, 2021)
    - GPT-3が保守的に促されると、保守的な人間とあまり類似しなくなるためではないかと考えられる(図2)
  - 重要なことは、道徳基盤の違いが、下流タスクでの振る舞いの違いに対応しているということである

Prompt Type	Prompt	Donation
Harm	You do not like to cause harm.	88.09 ± 34.644
Fairness	You believe the rich and poor should be treated with equity.	108.07 ± 17.15
Authority	You believe in traditional roles.	97.71 ± 35.91
Purity	You believe that some people are more important than others.	112.45 ± 14.91
Ingroup	You would sacrifice yourself for your country.	144.87 ± 6.35
No Prompt	N/A	92.66 ± 15.17
Conservative	You are politically conservative.	23.93 ± 50.81
Moderate	You are politically moderate.	79.36 ± 10.43
Liberal	You are politically liberal.	95.86 ± 7.61

Table 3: For each moral foundation, we show the prompt that was found to maximize the model's weight on this dimension. We then show that on the downstream donation task, the donation amount output by a LLM significantly differs based on the moral foundation scores that it obtains.

Discussion

- 本研究の動機は、LLMが示すモラルや価値観が、学習データや与えられた文脈によって影響を受けるかを評価することである

- GPT-3はAPIを通して300以上の製品に積極的に導入されている（Philipiszyn 2021）ため、GPT-3が道徳的あるいは政治的に偏っているとすると、ユーザーとの相互作用にそのような偏りをもたらしている可能性がある

- GPT-3はMFQに対して保守的な人間に最も近い回答をする一貫した傾向を示した
  - これはGPT-3が他の課題においても保守的なバイアスを示すことを意味するかは明らかではない
  - GPT-3の訓練データが保守的な人間のものであった可能性が考えられる

Discussion

- GPT-3は特定の道徳基盤を重視するよう促されることもでき、下流寄付課題での振る舞いに影響することがわかった
  - GPT-3が特定の道徳的スタンスやバイアスをとるコンテンツの作成に使われる可能性がある
    - ターゲットを絞った政治広告に使われた場合、強制的であり、特に危険である (Bakir 2020)

- GPT-3は一般的なLLMであるが、他のLLMの道徳基盤や行動を研究することも有益だろう
  - これらの知見はLLMが特定の道徳的スタンスを取ることによる潜在的なリスクや傾向のない結果を明らかにするのに役立つ