Title	単語埋め込みを用いた分配判断の予測 : 分析結果の解釈に正則化パラメータが与える影響の検討
Sub Title	Predicting allocation judgments using word embeddings : an investigation of the impact of a regularization parameter on the interpretation of analysis
Author	上島, 淳史(Ueshima, Atsushi)
Publisher	三田哲學會
Publication year	2024
Jtitle	哲學 (Philosophy). No.152 (2024. 3) ,p.33- 58
JaLC DOI	
Abstract	Understanding people's preferences toward the fair distribution of limited resources, such as vaccines for preventing infectious diseases during a pandemic, is of fundamental importance. While there are extant distributive principles that can be used to explain people's distributive preferences, predicting actual human distributive judgments is often challenging because people's judgments are influenced by the rich semantic information of judgment targets, which often cannot be captured by the existing distributive principles. The current paper focuses on the prediction and understanding of vaccine allocation judgments during a pandemic, investigating people's decisions on prioritizing certain occupations for vaccine distribution. To quantify rich semantic meanings associated with more than 130 occupations, we used word embeddings obtained from large-scale text data. The results showed that using word embeddings as regression predictors enabled reasonable predictions over people's vaccination priority judgments for various occupations. It was also indicated that the interpretation method introduced by previous studies, which allows for exploring semantic dimensions involved in people's judgments, did not substantially depend on the choice of regularization parameters of the regression analysis. Taken together, the current paper sheds light on the effectiveness of prediction-oriented judgment research while taking a look at its interpretability.
Notes	特集 : 人間科学専攻特集号 寄稿論文
Genre	Journal Article

単語埋め込みを用いた分配判断の予測

――分析結果の解釈に正則化パラメータが与える影響の検討――

上　島　淳　史*

Predicting Allocation Judgments Using Word Embeddings: An Investigation of the Impact of a Regularization Parameter on the Interpretation of Analysis

Atsushi Ueshima

Understanding people's preferences toward the fair distribution of limited resources, such as vaccines for preventing infectious diseases during a pandemic, is of fundamental importance. While there are extant distributive principles that can be used to explain people's distributive preferences, predicting actual human distributive judgments is often challenging because people's judgments are influenced by the rich semantic information of judgment targets, which often cannot be captured by the existing distributive principles. The current paper focuses on the prediction and understanding of vaccine allocation judgments during a pandemic, investigating people's decisions on prioritizing certain occupations for vaccine distribution. To quantify rich semantic meanings associated with more than 130 occupations, we used word embeddings obtained from large-scale text data. The results showed that using word embeddings as regression predictors enabled reasonable predictions over people's vaccination priority judgments for various occupations. It was also indicated that the interpretation method introduced

* 東北大学大学院文学研究科・日本学術振興会

by previous studies, which allows for exploring semantic dimensions involved in people's judgments, did not substantially depend on the choice of regularization parameters of the regression analysis. Taken together, the current paper sheds light on the effectiveness of prediction-oriented judgment research while taking a look at its interpretability.

導入

限られた資源をどのように社会の成員に分配するかは社会運営の根本に関わる（Rawls, 1971）．例えば，パンデミックの際にどのように感染症予防ワクチンを社会の成員に分配するかの判断は，社会の未来を決定づける要素となる．資源分配に関する規範的研究（「べき」論）は哲学をはじめとする人文学の領域でなされてきた一方で，人々が実際にどのような分配を好むかについては，心理学や行動経済学などの行動科学の領域で研究されてきた（Konow, 2003）．この行動科学による実際の分配に関する「である」論は，実社会で実践可能な分配（「できる」論）に重要な示唆を与える（参考：Greene, 2013）．本稿では，このような行動科学的な「である」論の視点から希少資源の分配に対する人々の判断の予測を行うと同時にその規定因を分析する．特に，感染症予防ワクチンの分配においてどの職業を優先するかについての人々の判断に関して，既存の分析方法の追加的検討を行う．

ワクチンの分配に関する人々の判断を説明するにあたっては，分配原理に関する規範的議論と経験的知見の双方を参考にすることができる．例えば，社会全体における感染症の被害を最小限にすることを考え，医療関係者へのワクチン分配を優先する人がいたとする．これは規範的正義論におけるある種の功利主義（Harsanyi, 1975）に基づく判断であると考えられる．功利主義は規範的正義論の代表的分配原理であるだけでなく，経験的にも，人々が分配される資源の総量の最大化を志向するという意味での功利主義的な好みを示すことが行動実験で観察されている（Charness &

Rabin, 2002; Engelmann & Strobel, 2004; Fisman et al., 2015）．それゆえ，人々の分配判断を功利主義的観点から説明しようとすることは規範理論的にも経験的にも妥当であると思われる．一方で，医療関係者が感染症に罹患した人と接触する機会が多く，高い感染リスクに晒されるという理由で，医療関係者へのワクチン分配を優先する判断は，社会全体の厚生を考えているというよりも，医療関係者にとっての必要性に根ざしており，功利主義とは別の原理で説明することがより適切と考えられる．

このように，看護師や医師などの医療を職業とする人々がワクチン分配を優先されるべきとする判断は，功利主義原理や他の過去に提案されている正義原理から説明できるかもしれない．しかしながら，他の職業の人々に対するワクチン分配がこれらの原理から十分に説明可能とは限らない．例えば，屋内よりも屋外に滞在する時間が長い職業に関しては，密な空間での滞在時間が短いという観点からワクチン分配の優先度が低く判断されるかもしれない．このような分配判断は，仮に功利主義原理の他に，格差原理（Rawls, 1971）や平衡原理（Adams, 1965）などの他の確立した分配原理を考慮したとしても，説明することが困難である．従来の分配原理のみでは，職業のような日常的概念が持つ多様な側面を評価して人々の分配判断を説明することは困難である．

確かに，職業に関しては分配的正義の観点以外からも様々な観点からの評価あるいは定量化が行われている．例えば，職業威信スコアは職業の社会的な地位を定量化する手法である．あるいは，厚生労働省が管理する「職業情報提供サイト（日本版 O-NET）」（https://shigoto.mhlw.go.jp）では，社会的地位，自律性，専門性，社会貢献度をはじめとする，職業に対する様々な印象や意味づけの定量化済みデータが数百種類の職業に関して提供されている．しかしながら，このような定量化の方法では，人間が明示的に概念化することができない側面の定量化が困難である．これらの質問項目は，職業に関する豊かな知識表現や情報を人間が低次元に圧縮して作成

した項目であるため，本来職業が含意する豊かな意味を定量化できていない可能性がある．

本稿では，このような課題を補う手法として，実社会に存在する大規模テクストデータを活用し，人々が職業に持っている印象を汲み取るアプローチに注目する．研究者が職業のある側面について人々の印象を計測するための質問項目を作成する先述のアプローチでは，特定側面について計測するために質問項目が作成される．一方で，実社会（主にインターネット上）の大規模テクストデータは，職業の印象について定量化するために集められたデータではないものの，そのような大規模データを職業の印象の定量化に用いて，分配判断の説明に利用することは可能である．本稿ではこのように必ずしも本稿の目的とは関係のない文脈で計測された大規模データを活用することで，職業に対して付与される豊かな意味づけや印象を定量化し，人々のワクチン分配判断を説明することを試みる（このように研究目的とは異なる文脈で集められたデータを用いるアプローチはレディメイドと呼ばれることがある．参考：Salganik, 2017）．

テクストデータによる意味の定量化の理論的背景には，テクストに出現する語の意味はその語が他のどのような語を伴ってテクスト上で出現するかで明らかになるとする，分布仮説（distributional hypothesis）が存在する（Firth, 1957; Harris, 1954）．例えば，「犬」と「猫」は同様の文脈でテクスト中に出現しやすいためにそれらに同様の数値表現を割り当て，一方で「犬」と「経済」は同様の文脈で相対的にはより出現しにくいため，それらに異なる表現を割り当てることで，各単語の意味的関係性を表現することができる．なお，同様の数値表現を持つ 2 つの単語が文章中で交換可能であることは意味しない．例えば，「戦争」と「平和」という単語は同様の文脈でテクスト中に出現しやすいため，類似した数値表現が割り当てられる可能性があるが，これらを文章中で交換すると文意が変わることが多いだろう．

テクストにおける単語の出現パターンに応じて意味を数値化する手法には認知科学や心理学をはじめとして様々な方法が存在する（Günther et al., 2019; Lake & Murphy, 2021）ことに加え，近年は単語にとどまらない文章の定量化技術（Devlin et al., 2019; Reimers & Gurevych, 2019）も発展しているが，本稿では機械学習分野で提案され既に古典的手法となった word2vec を単語の意味を表現するベクトルを得る手法として取り上げる（Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013）．Google Code Archive で公開されている word2vec リソース（https://code.google.com/archive/p/word2vec/）は，Google ニュースデータの一部を用いて訓練されている．同リソースから 300 万の語とフレーズについてそれぞれ 300 次元ベクトルを得ることができる．このように，テクスト上で元々は離散的な存在であった単語を連続的な数値（ベクトル）で表現する手法を単語埋め込み（word embeddings）と呼ぶ．300 次元は人間が直感的に理解しやすい 1 次元，2 次元，3 次元と比較すると高次元であるが，数多くの単語のそれぞれをたった 300 の数値で表現しているという点で，単語埋め込みは各単語の低次元表現ともいえる．図 1a は単語 “dog” の意味を表現するベクトルの一部である．なお，本稿の以下では単語埋め込みと単語ベクトルを交換可能な用語として用いる．

“dog” と同様の文脈でテクストに出現する語は，同様のベクトル表現となる．Google の word2vec リソースに含まれる単語のうち，“dog” のベクトル表現とコサイン類似度が高い上位 10 単語は “dogs” (0.87)，“puppy” (0.81)，“pit_bull” (0.78)，“pooch” (0.76)，“cat” (0.76)，“golden_retriever” (0.75)，“German_shepherd” (0.75)，“Rottweiler” (0.74)，“beagle” (0.74)，“pup” (0.74) である（“dog” 自体は上位 10 単語から除外した）．かっこ内の数値は意味空間上の単語の類似度の指標として用いられるコサイン類似度である（Jones & Mewhort, 2007; Landauer & Dumais, 1997）．word2vec は日本語などの英語以外の言語にも適用可能である．図 1b は日本語で訓練された word2vec モ

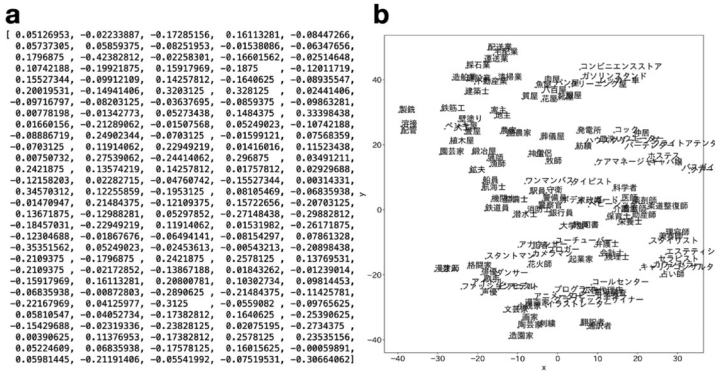


図 1　**a.** Google の word2vec から取得した “dog” の単語ベクトル（単語埋め込み表現）．300 次元のうち 1 次元目から 150 次元目までを表示している．この単語ベクトルを生成した word2vec リソースは Apache 2.0 ライセンスで配布されている．**b.** 日本語 word2vec リソースの一つである chiVe から取得した 132 種類の職業の単語ベクトルの可視化．300 次元のベクトルで表現されている各単語を tSNE（Maaten & Hinton, 2008) で 2 次元に圧縮して可視化した．同様の単語ベクトル表現となっている単語は近くにプロットされている．

デル（真鍋ら，2019）から先行研究（Ueshima & Takikawa, 2022）でワクチン分配の判断対象として使用された 132 種の職業のベクトルを取得し次元縮約を行い可視化している．

Mikolov らの提案した word2vec における単語ベクトルの学習には Continuous bag-of-words（CBOW）または skip-gram と呼ばれるモデルが用いられる．CBOW は，文章が「単語 1 →単語 2 →単語 3 →単語 4 →単語 5」と構成されている場合に，単語 1，単語 2，単語 4，単語 5 の内容から，単語 3 を予測できるようにニューラルネットワークを訓練する．その際，単語 1 が単語 2 より先に出現している時系列情報は考慮しない（bag-of-words）．なお，考慮する前後の単語数はパラメータとして調節可能であるので，単語 1 と単語 5 は含めず，単語 2 と単語 4 だけから単語 3 を予測するモデルも構築可能ではある．逆に，より多くの周辺語から単

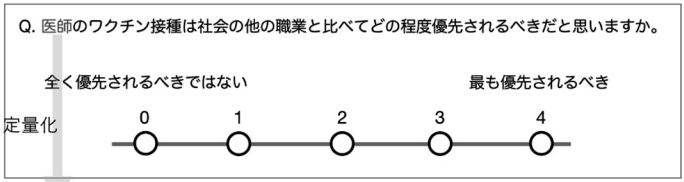
語 3 を予測するモデルを構築することもできる．一方の skip-gram は，単語 3 の内容から，単語 1，単語 2，単語 4，単語 5 の内容を予測できるようにニューラルネットワークを訓練する．CBOW や skip-gram を用いて，大量のテクストデータにおいて単語の予測を行えるようにニューラルネットワークを訓練することで，結果として各単語の意味に対応する数値ベクトル表現（図 1a）が得られる．CBOW あるいは skip-gram で文章に出現する各単語の意味表現が獲得される原理については，（Levy & Goldberg, 2014）などで検討されている．

大規模テクストから定量化された単語の意味は，人々の単語（概念）に対する意味づけを反映することが報告されている．例えば，単語ベクトルは人々が持つジェンダーバイアスを反映するとされる（Caliskan et al., 2017; Garg et al., 2018）．Caliskan ら (2017) は，男性に多い名前のベクトル表現と比べて，女性に多い名前のベクトル表現は，仕事に関する単語のベクトル表現よりも家庭に関する単語のベクトル表現とより高い類似度を示すことを報告している．同様に，テクストから得られた単語ベクトルは合接の誤謬（conjunction fallacy: Tversky & Kahneman, 1983）などの人間の持つ判断バイアスを反映する傾向が示されている．合接の誤謬を計測する課題の一つでは，実験参加者がリンダという人物についての記述を読んだうえで，リンダが「銀行員である」場合と「銀行員でかつフェミニスト運動に積極的である」場合のどちらの可能性がより高いかを回答する．リンダについての記述には，リンダが差別や社会正義の問題を気にかけている人物であるという内容が含まれている．論理的には「銀行員でかつフェミニスト運動に積極的である」確率は「銀行員である」確率を超えないはずであるが，多くの実験参加者がリンダが「銀行員でかつフェミニスト運動に積極的である」可能性をより高く見積ったという誤謬が報告されている．この合接の誤謬が生じる背景には，ある問題を解決する際に，その問題と最も強く連合する選択肢が人々に回答されやすいというメカニズムが以前

から指摘されてきた．しかしながら，このような仮説を定量的に示す方法がなかった．そこで Bhatia (2017) はテクストデータから得られる単語ベクトルを用いてこの説を示すことを試みた．同論文は，人々が「リンダに関する記述」と「銀行員でかつフェミニスト運動に積極的である」ことの類似度を高く判定することでこの誤謬が生じるという仮説を，「リンダに関する記述」のベクトル表現と「銀行員でかつフェミニスト運動に積極的である」ことのベクトル表現の類似度が，「リンダに関する記述」のベクトル表現と「銀行員である」ことのベクトル表現よりも高いという事実から支持した．単語埋め込みは単語間の意味的関係をテクスト中の単語の出現パターンという統計的規則性に基づいて定量化している．一方で，人々が日常的に行う判断も統計的規則性に基づいている．それゆえに，人々の判断と単語ベクトルの性質には共通点が存在し，人々の判断を意味に基づいてモデル化するにあたって単語ベクトルを用いることは有効なアプローチとなる．

単語の意味的類似度をもとに言語的・意味的課題における人々のふるまいを説明する上記のアプローチの成功を受け，近年の研究はより多様な領域での人々の判断の予測や説明に単語ベクトルの利用を拡張している（Grand et al., 2022; Richie et al., 2019）．Richie ら (2019) は，単語ベクトルを回帰分析の説明変数として利用することで，ジェンダー，職業，リスク，食糧，性格など様々な領域における人々の判断を予測する手法を提案している（マッピングアプローチ）．図 2 はマッピングアプローチを様々な職業がどの程度感染症予防ワクチンを優先的に分配されるべきかを判断する課題に適用する模式図を示している．

図 2 の例では説明変数は職業，従属変数はワクチン分配優先度（：全く優先されるべきではない，4：最も優先されるべきである）である．説明変数である各職業は，それぞれ word2vec から取得された単語ベクトルで表現される．単語ベクトルが 300 次元であれば，それに対応する 30 個の



[0.181, -0.376, 0.158, -0.138, … , -0.299]

「医師」の単語ベクトル（300次元）

$$\hat{y} = w_0 + w_1\underset{0.181}{x_1} + w_2\underset{-0.376}{x_2} + w_3\underset{0.158}{x_3} + w_4\underset{-0.138}{x_4} + \cdots + w_{300}\underset{-0.299}{x_{300}}$$

図 2　ワクチンの分配判断に対してマッピングアプローチを用いた例．回帰分析によって「医師」の単語ベクトルが判断（全く優先されるべきではない最も優先されるべき）にマッピングされる．回帰分析に投入される単語ベクトル（説明変数）は職業によって変化する．例えば，職業が「医師」ではなく「看護師」の場合は，「看護師」の単語ベクトルが判断に変換される．

回帰係数（切片を含めると 301 個）が人々の判断を予測できるように各意味次元への重みづけ（係数の大きさ）を決定する．このようにマッピングアプローチは単語ベクトルを説明変数とした回帰分析であり，300 次元の意味空間に存在している個々の説明変数（この例では職業単語）をワクチン分配優先度（スカラー値）にマッピングする．このアプローチは説明変数が単語ベクトルで表現できさえすれば様々な判断領域に用いることができる．

マッピングアプローチは，大規模なテクストデータから得られた様々な日常的概念の意味を活用できるため人々の判断を予測するうえで有用である．このアプローチのさらなる利点として，人々の判断が判断対象（例えば職業）に対するどのような意味づけをもとに行われているかの解釈を試みることが可能な点が挙げられる（Richie et al., 2019; Zou & Bhatia, 2021）．マッピングアプローチを用いる際に単語ベクトルが 300 次元であ

れば，人々の判断を予測できるように 300 の回帰係数が推定される．その際，300 次元の係数自体も意味空間上のベクトルとして解釈することができる．ここで，300 次元の係数ベクトルを $\boldsymbol{w}$，判断対象（上記の例では職業）i の単語ベクトルを $\boldsymbol{x_i}$ としたとき，回帰分析が出力する判断の予測値は $\boldsymbol{w} \cdot \boldsymbol{x_i}$ と比例する．$\boldsymbol{x_i}$ を大きさ 1 に正規化している場合，$\boldsymbol{w} \cdot \boldsymbol{x_i}$ の大きさは，$\boldsymbol{w}$ と $\boldsymbol{x_i}$ の意味空間上のコサイン類似度である $\frac{\boldsymbol{w} \cdot \boldsymbol{x_i}}{\|\boldsymbol{w}\|\|\boldsymbol{x_i}\|}$ と線形の関係にある（$\|\boldsymbol{x_i}\|$ は常に 1）．それゆえ，マッピングアプローチでは意味空間上で $\boldsymbol{w}$ と類似度が高い判断対象 $\boldsymbol{x_i}$ に対する出力ほど高い値（例えば，高いワクチン優先度）を取る．職業へのワクチン分配の例で，医療関係者へのワクチン分配優先度の予測を高く出力する回帰モデルでは，推定された $\boldsymbol{w}$ と「医師」や「看護師」などの医療関係の職業単語のベクトルの意味空間上での類似度が高くなる．重要なことに，職業以外の一般の単語のベクトルと $\boldsymbol{w}$ の類似度を計算することも可能である．そのようにすると，「医師」や「看護師」と意味空間上で類似度が高い「病院」や「福祉」，「介護」などの単語も，係数 $\boldsymbol{w}$ との類似度が高くなると考えられる．このように，説明変数として用いていない単語（この例の場合は非職業単語）を探索することで，人々の判断が判断対象（例えば職業）に対するどのような意味づけをもとに行われているかの解釈を試みることができる．

このように，先行研究ではデータから学習された回帰係数と意味空間上で類似度の高い単語を探索することで，判断の予測のみならず判断の心理次元の解釈が行われている．ただし，判断次元の解釈に際しては，マッピングアプローチが数百次元の単語ベクトルを説明変数としており，回帰モデルがデータに過剰適合することを防ぐために正則化を行う必要がある点に注意する必要がある．正則化の方法によって 300 次元の係数が異なる値を取るため，それに伴って係数と意味空間上で類似度が高い単語も多少は変化する可能性がある．本研究では，回帰分析を用いたマッピングアプローチにおける判断次元の解釈が正則化パラメータを変化させた際にどの

程度安定して行えるのかを検討する.

方法

参加者

様々な職業へのワクチン分配をマッピングアプローチで分析した Ueshima & Takikawa（2022）の公開データの再分析を行う．同論文は 2021 年 5 月に 243 名の参加者をクラウドソーシングサイト上でリクルートしたオンライン実験であり（性 136 名，男性 105 名，回答なし 2 名」コ平均年齢：40.88 歳），東北大学文学研究科調査・実験倫理委員会による承認を受けて実施された研究に基づいている.

従属変数

各参加者は 132 種類の職業からランダムに選ばれた 5 つの職業について，「もし特定の職業から優先して新型コロナウイルス感染症のワクチンが接種される場合，各職業は社会の他の様々な職業と比べてどの程度接種を優先されるべきだと思いますか.」という質問に回答した．各職業に対する優先度評定は，「全く優先されるべきではないと思う」（）から「最も優先されるべきだと思う」（4）の 5 段階で行われた.

説明変数

132 種類の職業は，総務省の日本標準職業分類を参考に選択された．各職業の単語ベクトル表現（300 次元）は，事前訓練済みの日本語 word2vec リソース（真鍋ら，2019）から取得された.

回帰分析

マッピングアプローチを用いた先行研究（Richie et al., 2019）に基づき，リッジ回帰（L2 正則化）により参加者の 5 段階の評定を予測した．モデ

ルは以下になる．

$$\hat{y_i} = w_0 + \boldsymbol{w}\boldsymbol{x_i}$$

$\hat{y_i}$ は職業 i に対する人々のワクチン分配判断（：全く優先されるべきではないと思う〜4：最も優先されるべきだと思う）の予測平均値を示す．$\boldsymbol{x_i}$ は職業 i の 300 次元のベクトルを表す（$\|\boldsymbol{x_i}\| = 1$）．x_{ij} はスカラー値で職業 i の単語ベクトルの j 次元目の値を示す．$\boldsymbol{w}$ は回帰係数であり 300 次元ベクトルを表す．$\boldsymbol{w}\boldsymbol{x_i}$ はこれら二つのドット積として計算され，スカラー値となる．w_0 は切片を表す．リッジ回帰における損失は以下のように定義される．

$$\textit{平均二乗誤差} + \lambda \sum_j w_j^2$$

平均二乗誤差に加えて，回帰係数 $\boldsymbol{w}$ の二乗和に λ（正則化項）を掛けた値が損失に含まれる．λ は以上の値を取る．λ がより大きい場合に，損失を抑えるために回帰係数が小さな値となるように推定されるため，正則化項がない場合に比べると訓練データへの過剰適合が抑制される．なお，平均二乗誤差は以下のように計算される．

$$\sum_i (y_i - w_0 - \sum_j x_{ij} w_j)^2$$

本研究では λ の大きさを変えることが，判断次元の解釈にどの程度の影響を与えるのかを探索する．

ソフトウェア

回帰分析には python パッケージの scikit-learn v.1.3.0（Pedregosa et al., 2011）を用いた．作図には Python パッケージの matplotlib v.3.7.2（Hunter, 2007）を用いた．

結果

以下で示す分析の一部は先行研究（Ueshima & Takikawa, 2022）の再現を含む.

基礎的な結果

参加者の平均的な判断値は 2.03（標準偏差 =1.06）であった（：全く優先されるべきではないと思う〜4：最も優先されるべきだと思う）. 図 3a は 1215 回答（243 人*5 回答）の分布を示している.

最もワクチン分配を優先されるべきであると判断された 1 職業は,「医師」(3.80),「看護師」(3.70),「介護士」(3.67),「消防士」(3.43),「ケアマネージャー」(3.42),「保育士」(3.25),「宅配業」(3.23),「助産師」(3.22),「薬剤師」(3.17),「清掃業」(3.14）であった（数値は各職業に対する平均判断値を示す）. 一方で, 最もワクチン分配を優先されるべきでないと判断された 10 職業は,「画家」(0.67),「地主」(0.75),「イラストレーター」(0.83),「文芸家」(0.86),「アイドル」(1.00),「ユーチューバー」(1.00),「占い師」(1.00),「ブロガー」(1.13),「タイピスト」(1.18),「靴屋」(1.20) であった.

回帰分析（マッピングアプローチ）

次に, 各職業に対する平均的な判断値を回帰分析で予測するために, リッジ回帰において単語ベクトルを説明変数として用いた. 予測は 1 職業抜きの交差検証で行った. これは, 例えば「医師」に対する人々の平均判断値を予測する際には,「医師」以外の 131 職業でモデルを訓練し,「医師」に関する判断値は訓練に利用しないことを意味する.「タイピスト」に対する判断値を予測する際には,「タイピスト」以外の 131 職業でもう一度モデルを訓練し直す. このようにして, 各職業についてその職業に対する判断値のデータは使わずに予測値を出力した. 図 3b は参加者の各職業に

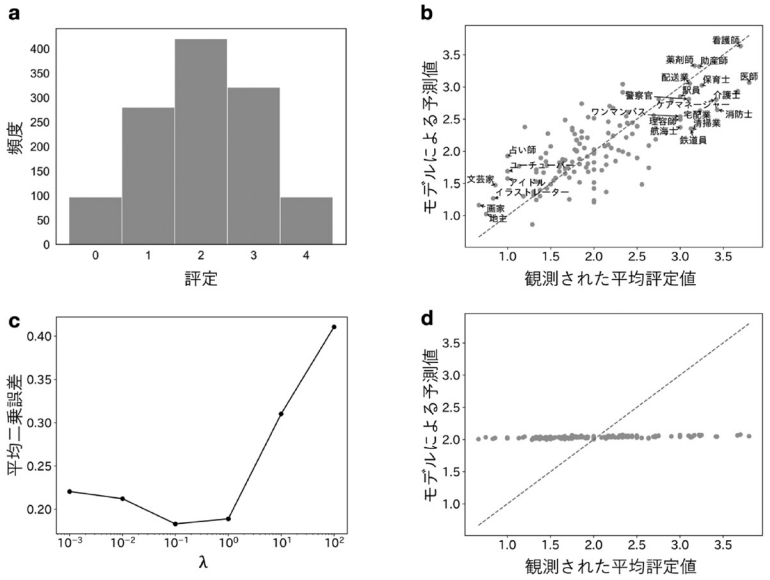


図 3　**a.** 全回答の分布．**b.** 各職業に対する観測された平均評定値（x 軸）と 1 職業抜き交差検証の予測値（y 軸）の散布図．点線は y=x を示す．「看護師」のように，データポイントが点線上に近い職業ほど，観測された値（3.7）に対して予測（3.64）が正確に行えている．「占い師」のように，点線から遠いデータポイントほど，観測値をモデルが正確に予測できていないことを示す．なお，観測された平均評定が 1.0 以下の職業と 3.0 以上の職業について職業名を表示している．また，ここでは正則化パラメータ λ は 0.1 に設定している．**c.** λ が予測精度に与える影響．λ を 0.001,0.01,0.1,1,10,100 に設定した各場合において，観測値と予測値の平均二乗誤差を算出した．図は λ が 0.1（10^{-1}）の場合に誤差は最も低かったことを示している．**d.** λ を 100 に設定した場合の平均評定値と予測値の散布図（職業名をラベリングしていないこと以外の描画方法は図 3b と同じ）．λ が過剰に大きいことで，回帰係数が小さくなる方向に過度な罰則が課されているため，ほぼ切片のみで予測値が決まっていることがわかる．

対する平均評定と，リッジ回帰において λ =0.1 とした場合のモデルの予測する平均評定の散布図を示している．実測値と予測値のピアソン相関係

数は 0.76 であった．図 3c は λ を 0.001,0.01,0.1,1,10,100 に設定した各場合における，観測値と予測値の平均二乗誤差を示す．λ が 0.1 のとき誤差は最も低いが，λ が 1 の場合でも誤差には定性的には大きな変化がみられなかった．ただし，λ を 100 に設定した場合は，予測誤差に大きな変化がみられることを確認した（図 3d）．

判断メカニズムの解釈

次に，人々の判断が判断対象（職業）に対するどのような意味づけをもとに行われているかを解釈するために，推定された回帰係数 $\boldsymbol{w}$（300 次元ベクトル）と，意味空間上で類似度の高い一般的な単語を探索した（Richie et al., 2019）．$\boldsymbol{x_i}$ は全て大きさ 1 に正規化しているため，$\boldsymbol{wx_i}$ のドット積を計算し，その結果が高い単語を探索することで $\boldsymbol{w}$ と類似度の高い単語を探索することができる．なお，この分析では 1 職業抜き交差検証で得られる $\boldsymbol{w}$ ではなく，全データを用いて推定した $\boldsymbol{w}$ を用いた（λ は 0.1 とした）．

132 の職業単語に限らない一般単語を『現代本語書き葉均衡コーパス』（国国語研究所）から抽出した．助詞，助動詞，副詞，接続詞，間投詞，代名詞，連体詞，接頭辞を除いた頻出 5,000 語を選択し，重複する単語を除き，本稿で使用している word2vec リソース（chiVe）で単語ベクトルが提供されている単語に絞ると，4,196 単語が残った．この 4,196 単語のそれぞれと $\boldsymbol{w}$ のドット積を計算することで，意味空間上で $\boldsymbol{w}$ との類似度が高い一般単語を探索した（図 4）．医療関係や行政，交通機関，保育に関わる単語が出現していることから，参加者が特定の職業へのワクチン優先度を高く判断したこととこれらの概念が関連することが示唆される．

λ の影響

上記の判断メカニズムの解釈に対して回帰分析のハイパーパラメータがどの程度影響しうるのかを検討した．上記の分析では λ を．1 に設定して

1病院	11福祉	21出産	31警察	41上司
2職員	12患者	22責任者	32鬱病	42命
3医療機関	13市役所	23事業所	33タクシー	43役人
4医療	14課長	24入院	34幼稚園	44労働時間
5厚生労働省	15認可	25行政	35休日	45交付
6便	16看護婦	26施設	36業務	46緊急
7薬	17乗客	27職場	37公務員	47届け出
8介護	18災害	28定員	38部長	48予防
9保育園	19副作用	29自衛隊	39長	49医療費
10医者	20役所	30処置	40老人	50事務

図 4　意味空間上で $\boldsymbol{w}$ との類似度が高かった上位 50 単語.

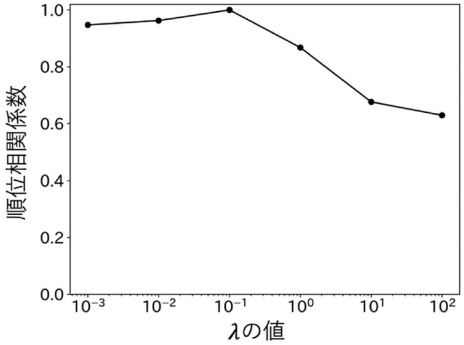


図 5　λ の値が 4,196 単語の順位に与える影響．手順として，まず λ が 10^{-1}（0.1）の場合に得られた $\boldsymbol{w}$ と 4,196 単語の意味空間上での類似度で順位をつけた．図 4 で示した通り，病院は 1 位，職員は 2 位などとなった．次に，λ が他の値の場合に得られた $\boldsymbol{w}$ と 4,196 単語の意味空間上での類似度で順位をつけた．その上で，λ が 0.1 の場合とその他のそれぞれの場合で順位がどの程度変動するかを順位相関で評価した．

回帰係数 $\boldsymbol{w}$ を推定し，$\boldsymbol{w}$ と 4,196 単語の意味空間上の類似度を算出した．ここでは新たに，λ を 0.001,0.01,10,100 に設定した各場合で $\boldsymbol{w}$ を推定し，同じく 4,196 単語と $\boldsymbol{w}$ の類似度を算出した際に，4,196 単語と $\boldsymbol{w}$ の類似度の順位が λ を 0.1 に設定した場合と比較してどの程度変化するかを順位相関係数で検討した（図 5）.

分析の結果，λ が 10^{2}（100）の場合の順位と，λ が 10^{-1}（0.1）の場合の順位の順位相関係数は，0.6 以上の値となることが明らかとなった．λ

が 10^2 の場合に $\boldsymbol{w}$ と最も距離が近い 2 単語は「職員」,「看護婦」,「医療機関」,「介護」,「病院」,「医療」,「患者」,「運転手」,「乗客」,「処置」,「業務」,「職場」,「利用者」,「タクシー」,「医者」,「事業所」,「厚生労働省」,「事務」,「血圧」,「高齢者」であった．これらの単語は図 4 で 1 位から 2 位に示されている単語と異なるものの，医療や交通機関などの同様のトピックを捉えていると解釈できる．これらの結果は，λ が過剰に大きく予測が成立していない場合（図 3d 参照）でも，結果の解釈に与える影響が非常に大きいとは言えない可能性を示唆している．

考察

本稿では，職業を判断対象としたワクチン分配判断について，単語ベクトルを用いた回帰分析でアプローチする方法を取り上げ，単語ベクトルが人々の職業に対する意味づけを定量化する有効な手段である可能性を検討した．特に，先行研究では十分に扱われていなかった，正則化パラメータ（λ）が結果の解釈に与える影響を新たに検討した．分析の結果，正則化パラメータが妥当な値に設定された場合と比較して，正則化パラメータの値が不必要に大きい場合においても，一般単語と $\boldsymbol{w}$ の意味空間上での類似度の順位はある程度保全されることが明らかになった．この結果は，本稿や先行研究で用いられた種類の回帰分析においては，正則化パラメータを変化させたところで，どのような意味次元が判断に関連していたかに関する解釈が大きくは変化しない可能性を示している．

本稿の結果を解釈する際にはいくつか注意すべき点が存在する．1 つめに，本稿では正則化パラメータの変化に伴って一般単語と $\boldsymbol{w}$ の意味空間上での類似度が変化する様子を順位相関係数で定量化しているがこれは唯一の定量化方法ではない．本稿が順位相関係数を用いたのは，判断次元を解釈する際に一般単語と重みパラメータ $\boldsymbol{w}$ との類似度の順位を定性的に確認する先行研究の手法を前提としているためである．順位ではなく重みパ

ラメータ $\boldsymbol{w}$ と各単語の意味空間上での類似度それ自体に関心がある場合には，別の定量化による評価が適切である．2 つめに，リッジ回帰で推定される回帰係数は不偏推定量ではないため，$\boldsymbol{w}$ の意味空間上での位置を正確に推定することはできない．この事実は，本稿の内容にとどまらず，先行研究で提案された解釈手法の妥当性を評価する上で改めて考慮すべきであると考えられる．3 つめに，本稿や先行研究ではあくまで予測モデルとしてリッジ回帰を用いており，人間が実際にどのようなプロセスで判断を下すかをモデル上に反映していない点も考慮する必要がある．本稿で用いたモデルは，300 の意味次元のそれぞれに対する重みづけを決定するが，それぞれの意味次元を人間が行うように重みづけするわけではない．例えば，実際の人間が判断を下す際には特定の意味次元群を最初に考慮する可能性も考えられるが，本稿の回帰モデルはそのような心理プロセスを考慮しない．4 つめに，本稿のモデルは各職業に対して人々が下した判断の平均を予測するモデルであるため，個々人が各職業に対してどのような意味づけを行って判断したかは明らかにしない．実際には個人によって看護師がどのような職業かに関する意味づけは大きく異なる可能性がある．例えば，看護師が知り合いにいるか否かで「看護師」という単語からある人が何を心的に思い浮かべるかは異なるだろう．それゆえ，本稿のアプローチは個人が職業をどのように意味づけるかを十分に明らかにするものではない．最後に 5 つめとして，一般単語と $\boldsymbol{w}$ の意味空間上での類似度を利用して判断の心理次元を解釈する方法を，どのように応用するかを議論しない限り，正則化パラメータの実際的（実用的）な影響は当然ながら明らかではないという限界も挙げられる．

以上の限界の一部は，大規模テクストデータから得られる 300 次元の単語ベクトルを用いる予測モデルではなく，解釈可能な少数の意味次元を個人ごとに測定して実際の人間心理を説明することに重きを置いたモデルを構築すれば，解決できるかもしれない．しかしながら，人々が下す判断を

予測する上では，本稿のように数百次元の単語ベクトルを説明変数に用いることに利点がある．近年の心理学方法論ではこれまでの心理学の研究が心理現象の説明あるいは理解に偏ってきたことの問題点と予測の重要性が指摘されている（Rocca & Yarkoni, 2021）．現象の理解を重視する研究の課題としては，そのような研究は単純で表現力の低いモデルを用いる傾向があるが，人間が直感的に理解できるモデルで近似できるほど心理現象が単純である保証はないことが挙げられる（Yarkoni & Westfall, 2017）（社会学においても社会現象に関して同様の議論がなされている：Watts, 2014）．社会科学一般を対象とした方法論に関する研究においても，社会科学では基本的に線形モデルの使用が前提とされやすいことと，現象の複雑さに対してモデルが単純であることは誤った結論を導く可能性が議論されている（Vowels, 2023）．一方で，人々の現在の行動を理解するだけでなく，予測できることは，予防的な政策介入にもつながりうる点で重要である（例えば Salganik et al., 2020）．これらを踏まえると，直感的な解釈のしやすさよりも予測を重視することはある程度妥当であると考えられる．

幸い，予測を重視することで理解を重視した心理学的知見を必ずしも諦める必要はない．予測重視のアプローチと理解重視のアプローチの対照的な点（Breiman, 2001）への注目に加えて，近年は予測と理解の相補性（Hofman et al., 2021）も議論されている（共通する問題点も議論されている：Hullman et al., 2022）．実際の経験的研究においても，例えば（本来は心理学的な解釈が困難である）ニューラルネットワークを用いて予測力を重視しつつ，心理学・認知科学的理解も得ようとする研究がなされている（Agrawal et al., 2020; Miller et al., 2023; Peterson et al., 2021）．これらの研究は，人々の行動の予測モデルの構築を通じて理解に迫ることが可能であり，かつ有益である可能性も示している．例えば本稿では予測モデルを用いつつ，一般単語と $\boldsymbol{w}$ の意味空間上での類似度の分析から判断の心理的次元を解釈したが，このような解釈をもとにこれまで議論されていなかった

分配に関する価値規準を理解することで予測と理解の両立が可能かもしれない．

結論

本稿では職業という日常的概念を対象にした分配判断の予測を単語埋め込みを利用して行った．それと同時に，判断に伴う意味次元の解釈を行い，正則化パラメータの影響を探索した．様々な限界があるものの，本稿では予測を重視した心理学研究の実践の一例を示した（もしこれを心理学研究と呼ぶなら）．本稿では単語の定量化に word2vec という古典的技術を用いたが，自然言語処理は近年急速な進展を遂げている．それに応じて，心理学において予測を重視するアプローチは今後さらに発展すると考えられる．心理学は再現可能性危機（Open Science Collaboration, 2015）を経験し，その後も一般化可能性危機（Yarkoni, 2022）を含む様々な方法論的変革（平石 & 中村，2022）の下にある．近年の大規模言語モデル（Brown et al., 2020）の発展で生まれつつある科学研究の新しい潮流は，予測を重視するアプローチに限らない更なる機会を心理学研究にもたらすと考えられる．

利益相反事項

本稿に関して著者が報告すべき利益相反はない．

オープンマテリアル

本稿が分析に用いたコードは次のリンクにて共有されている：https://osf.io/fzgd7/?view_only=5068110698af47be9fa74989d97a8397

本稿が分析に用いたデータは先行研究によって次のリンクで共有されている：https://osf.io/mgpws/

補足

要旨の校正に OpenAI の GPT-4 を用いた.

謝辞

本研究は特別研究員奨励費の助成を受けている（21J00403）.

引 用 文 献

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 2, pp. 267–299). Academic Press.

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, *117*(16), 8825–8835.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3) 199–231

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language models are few-shot learners* arXiv. https://doi.org/10.48550/arXiv.2005.14165

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). BERT: Pre-training of deep bidirectional transformers for language understanding [Conference session]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, United States.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857–869.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Studies in Linguistic Analysis. Oxford: Philological Society. Reprinted in F. R. Palmer (ed.), Selected Papers of J. R. Firth 1952–1959, London: Longman (1968), 1–32

Fisman, R., Jakiela, P., Kariv, S., & Markovits, D. (2015). The distributional preferences of an elite. *Science*, *349*(6254), aab0096.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Atlantic Books

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2–3), 146–162.

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review*, *69*(2), 594–606.

平石界，& 中村大輝．(2022)．心理学における再現性危機の 10 年．科学哲学，*54*(2), 27–50.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). *The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning* arXiv. https://doi.org/10.48550/arXiv.2203.06498

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature*, *41*(4), 1188–1239.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*, *130*(2), 401–431.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Levy, O., & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, &

K. Weinberger (Eds.), Advances in Neural Information Processing Systems (NIPS) 2014 (Vol. 27, pp. 2177–2185).

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

真鍋陽俊，岡照晃，海川祥毅，髙岡一馬，内田佳孝，& 浅原正幸．(2019). 複数粒度の分割結果に基づく日本語単語分散表現．言語処理学会第 25 回年次大会（NLP2019），NLP2019-P8-5.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* arXiv. https://doi.org/10.48550/arXiv.1310.4546

Miller, K. J., Eckstein, M., Botvinick, M. M., & Kurth-Nelson, Z. (2023). *Cognitive model discovery via disentangled RNNs*. bioRxiv. https://doi.org/10.1101/2023.06.23.546250

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-Networks* arXiv. https://doi.org/10.48550/arXiv.1908.10084

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1), 50.

Rocca, R., & Yarkoni, T. (2021). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211026864

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., ... McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Ueshima, A., & Takikawa, H. (2022). Analyzing vaccination priority judgments for 132 occupations using word vector models. *IEEE/WIC/ACM International Conference on Web Intelligence*, 76–82.

Ueshima, A., & Takikawa, H. (2022). Analyzing vaccination priority judgments for 132 occupations using word vector models. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (pp. 76–82). Association for Computing Machinery.

Vowels, M. J. (2023). Misspecification and unreliable interpretations in psychology

and social science. *Psychological Methods*, *28*(3), 507–526.

Watts, D. J. (2014). Common sense and sociological explanations. *American Journal of Sociology*, *120*(2), 313–351.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Zou, W., & Bhatia, S. (2021). Judgment errors in naturalistic numerical estimation. *Cognition*, *211*, 104647.