The Psychological Effect of Large-Scale Corpus Annotation

Anonymous ACL submission

Abstract

Large-scale corpus annotation could have a psychological burden on annotators, and the repeated task of coding text could influence annotators' own attitudes and preferences. Existing work is qualitative, correlational, or inconclusive. Here, we use a quasi-experimental design to investigate the causal relationship between prolonged engagement in data labeling and psychological outcomes. We examined how 30 trained annotators who labeled a large dataset were affected across two important psychological dimensions: mental health and Post Traumatic Stress Disorder (PTSD) symptoms. We found no evidence that annotations led to declines in mental health or PTSD symptoms when compared with a non-annotating control group. Bayesian analyses provided additional support for the null effect. Our findings provide recommendations for future annotation efforts, demonstrating that integrating annotator-centric design principles can safeguard against psychological harm in annotation projects.

1 Introduction

Annotated datasets are the backbone of supervised machine learning in natural language processing (NLP). Human annotators invest time and effort to label raw text with theoretically defined tags, establishing the "ground truth" for NLP tasks (Kennedy et al., 2022; Geva et al., 2019; Hovy and Lavid, 2010). Specifically, manually annotated corpora are used to train models for applications such as coding linguistic features, sentiment analysis, machine translation, or identifying hate speech on social media platforms (Ljubešić et al., 2023; Pak et al., 2010). Some annotation efforts, often those that require straightforward tasks (e.g., coding word similarity), depend on online crowdsourced workers who receive minimal training or rely on their background knowledge and intuitions to tag text data (Snow et al., 2008). In some cases, however, the process of annotation is complex and subjective (e.g., coding for moral values), making the annotation experience more cognitively and emotionally demanding (Davani et al., 2024; Hoover et al., 2020).

The impact of the annotation process, particularly when it contains emotionally charged or hateful content, on annotators has slowly emerged as a growing focus in the NLP community. Previous research has primarily relied on qualitative interviews and systematic literature reviews to raise awareness about the negative effects of such exposure on annotators (e.g., emotional fatigue, desensitization, and psychological distress, especially in tasks like hate speech detection;(AlEmadi and Zaghouani, 2024; Steiger et al., 2021)). These burdens become pronounced when annotating large-scale psychological corpora, as these tasks demand high levels of cognitive and emotional engagement.

Unlike objective linguistic annotation, coding of text for *psychological* processes often lacks a single "correct" answer. Thus, annotators need to be extensively trained to learn a theoretically-informed typology or an annotation guide to tag text data for its psychological content (Davani et al., 2023). Despite these challenges, such corpora are crucial for diverse applications, including emotion recognition (Poria et al., 2019), moral discourse (Preniqi et al., 2024), social bias inference (Sap et al., 2020), and building generative language models capable of representing cultural values (Kasirzadeh and Gabriel, 2023). Additionally, psychologically annotated corpora may serve as benchmarks for evaluating the biases of generative language models (Rathje et al., 2024; Kristensen-McLachlan et al., 2025; Abdurahman et al., 2024). However, as large-scale psychological annotation grows more in demand (Wolfe, 2000; Beck et al., 2024), the mental health of annotators emerges as a critical, yet often overlooked, aspect of the data annotation pipeline.

Indeed, there is little experimental research on how the annotation process itself affects annotators'

mental well-being or other psychological attributes (Prabhakaran et al., 2021). To bridge this gap, we investigate the psychological impact of annotations by having Research Assistants (RAs) label the recently compiled Culture and Moral Expressions in Language (CAMEL) Corpus (Anonymous Authors, 2025), a collection of ~51,000 texts, including a wide range of data sources from movie reviews to offensive language datasets (see Table 1 in Appendix). Based on prior work, we focused on two domains: (1) mental health and (2) Post Traumatic Stress Disorder (PTSD) symptoms. A host of other psychological outcomes are also reported in the Appendix. The annotation process was converted into a quasi-experimental design (see Figure 1), with a control group, wherein trained annotators completed self-report measures before and after annotating the corpus. A demographically matched control group also completed the identical measures twice, but did not perform any annotations. In summary, this paper turns the spotlight from the *annotation* to the *annotator*, quantifying the often-overlooked psychological impact of annotations.

2 Background

Mental health is characterized by the capacity of individuals to adapt flexibly to stress and adversity, manage emotions effectively, and function well within social contexts (Galderisi et al., 2015). Annotation might impact annotators' mental well-being by contributing to cognitive fatigue, stress from exposure to sensitive content, and potential emotional strain from prolonged engagement with subjective or ethically complex language (AlEmadi and Zaghouani, 2024; Stoev et al., 2023).

The present experimental set-up is grounded in Self-Determination Theory (STD; (Ryan and Deci, 2017)), a pluralistic theory of motivation and well-being suggesting that individuals have three fundamental psychological needs: autonomy, competence, and relatedness. Satisfying these needs is crucial for enhancing well-being. However, the highly controlled, deskilled, and isolating nature of annotation tasks can hinder the fulfillment of these needs, potentially resulting in a negative impact on well-being. In our study, we designed the annotation process to ensure annotators' needs were met. Autonomy, the perceived control of one's actions, could be unmet for annotators due to strict guidelines, such as fixed time limits for each task, rigid hourly quotas, or lack of flexible scheduling (Sansone and Thoman, 2005). Competence, reflecting an individual's need to feel efficient (Ryan and Deci, 2017; Kaveti and Akbar, 2020), is undermined by the overly repetitive nature of annotation tasks, offering no performance feedback for skill development. Relatedness, the connection individuals experience when they feel accepted by others (Walton and Cohen, 2007), is impacted as annotators often work in isolation with no community support, making their "invisible labor" concealed behind algorithms (Steiger et al., 2021). The fulfillment of these needs is a robust predictor of mental health (Ng et al., 2012; Slemp et al., 2024).

Annotation practices that make annotators feel undervalued and replaceable may undermine these basic psychological needs and negatively impact mental health. Additionally, annotation work carries the risk of vicarious traumatization. Vicarious Trauma Theory (McCann and Pearlman, 1990) suggests that constant exposure to others' traumatic experiences can lead to symptoms similar to those of PTSD (known as "the cost of caring" or "compassion fatigue") (Figley, 2002). Annotators who engage with graphic or hateful material may experience PTSD-like symptoms due to a lack of adequate psychological safeguards. Qualitative interviews (Spence et al., 2023) revealed that moderators frequently encounter trauma-related symptoms, such as intrusive thoughts and sleep disturbances. Crowd annotators typically lack training and receive minimal institutional support, making them particularly vulnerable to vicarious trauma.

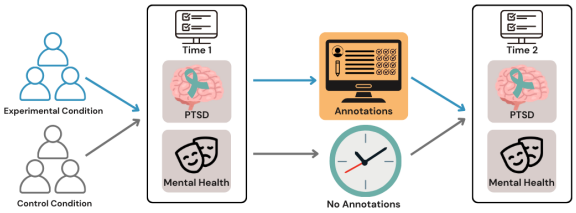

Figure 1: Schematic design of the study

3 Experimental Setup

3.1 Design

The study employed a quasi-experimental pretest-posttest design with a control group. Individuals in the experimental group annotated a random subset of the CAMEL corpus (see Table 1) over 4 months, whereas the control group did not take part in annotations. A demographically matched control group on gender and race strengthens causal inference by

controlling for potential differences in how mental health is affected across these groups (Wallace et al., 2016; Salk et al., 2017). The pretest-posttest design further allows us to measure change over time. Participants in our experimental group were slightly older, but this difference was minimal and primarily driven by a small subset of older RAs (Appendix C). Both groups were blinded to the study's aim and were informed that the study examines their general beliefs about the world. All assessments (Appendix A) were conducted twice: (1) at the beginning of the project, before the annotation process began ($Time_1$), and (2) upon completion of the project ($Time_2$). Annotators were RAs in our institution and received research or course credits. Our Institutional Review Board approved the experiment (IRB #5291; Appendix D).

3.2 Training of Research Assistants

To prepare RAs for the annotations, we incorporated principles from STD into our training process. We developed a comprehensive annotation manual that included detailed explanations for each label. To enhance competence, the training included thorough manual reviews and small-group discussions, each with five annotators, all supervised by experienced graduate students, to clarify any ambiguities. This structured approach promoted a shared understanding of the guidelines. Following the training, RAs consistently completed between 800 and 1,000 annotations per week. To enhance relatedness, weekly meetings were held with supervisors to address emerging challenges. To support autonomy, we emphasized the importance of independent judgment. Annotators were not permitted to confer on labeling any specific text (Appendix E).

3.3 Participants

Two cohorts, Cohort 1 (February – May 2024) and Cohort 2 (August – December 2024), were included to ensure that observed effects were not artifacts of a particular time in the academic year. We observed attrition rates of ~10% in both groups at $Time_2$ assessments, likely due to students' limited availability to retake the post-test psychological measures. In the aggregate sample, the experimental group consisted of 30 annotators (M_{age} = 20.7, SD_{age} = 1.7, *Male* = 6, *Female* = 22, *Nonbinary* = 2; *White* = 14). Annotators met weekly with their supervisor to address any discomfort they may have experienced during the annotation work and had access to a communication channel outside of meetings to reach out if they felt distressed. The control group consisted of 35 undergraduate students (M_{age} = 19.9), SD_{age} = 1.4, *Male* = 10, *Female* = 23, *Nonbinary* = 2; *White* = 16). Data from Cohorts 1 and 2 were combined to increase statistical power.

3.4 Measures & Procedure

Mental health and PTSD symptoms were assessed using psychometrically validated scales (Lukat et al., 2016; Weathers, 2013). Participants in an experimental lab setting (see Figure 1), and completed a 30-minute survey with these scales along with additional measures reported in Appendix.

3.5 Statistical Analyses

We conducted a linear mixed model (group as a between-subject factor and time as a within-subject factor) to examine the psychological impact of annotating a large-scale corpus. Equation 1 demonstrates the effect of annotations on outcome variables (mental health or PTSD symptoms) for annotator *i* at time *t* where u_i represents the annotator-specific random intercept (accounting for individual variation) and ϵ_{it} represents residual. Our key parameter of interest, β_3, estimates the difference in pre–post change in the outcome between the experimental and control groups. It tells us how much the experimental group's outcome changed over time compared with the control group's change.

$$\text{Outcome}_{it} = \beta_0 + \beta_1 \text{Group}_i + \beta_2 \text{Time}_t + \beta_3(\text{Group}_i \times \text{Time}_t) + u_i + \epsilon_{it} \quad (1)$$

We also conducted a Bayesian hierarchical multilevel model using the $\mathcal{N}(0, 10^2)$ distribution as our priors on fixed effects (Bürkner, 2017; Browne and Draper, 2006). This Bayesian approach allows for estimations of the full posterior distributions of the fixed effects and annotator–level random intercepts, derivation of 95% Credible Intervals (CrI) for each parameter estimate, and comparison of alternative model specifications (e.g., with and without the Time×Group interaction) using Bayes factors. Notably, Bayes Factors (BFs) compare the likelihood of the data under the null vs. the alternative, allowing us to quantify evidence for both null and alternative hypotheses, not just a binary "reject" vs. "do not reject" decision (Wagenmakers et al., 2018).

4 Results

Linear mixed-effects models revealed no significant treatment effect for mental health, (β_3 =

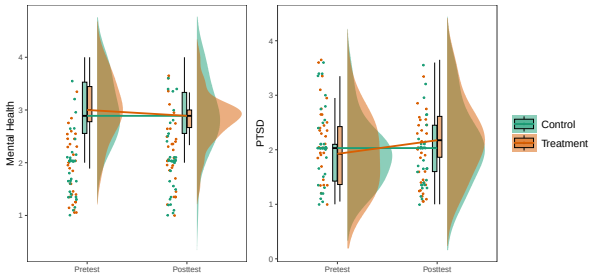

Figure 2: Effect of Group on Changes in Mental Health and PTSD Over Time.

0.11, $SE = 0.09$, $p = 0.223$), or for PTSD, ($\beta_3 = -0.12$, $SE = 0.16$, $p = .455$) (see Figure 2 & Table 2). Bayesian model posterior estimates for the Time×Group interaction suggested a small positive change in mental health for the treatment group ($\beta_3 = 0.11$, 95%CrI [–0.06,0.29]). Bridge-sampling yielded "very strong evidence" for the null hypothesis (BF_{01} = 64.0) (Wagenmakers et al., 2018), indicating that treatment and control changed similarly over time. Moreover, posterior estimates for the time×group interaction depicted a small positive change in PTSD for the treatment group ($\beta_3 = -0.14$, 95%CrI [–0.45,0.20]). Bridge-sampling retained "very strong evidence" for the null hypothesis (BF_{01} = 45.9) (Wagenmakers et al., 2018), highlighting that treatment and control groups changed similarly (see Table 3). Also, no reliable change in any of the additional psychological outcomes among annotators was seen, as shown in Tables 4 - 7.

5 Discussion

Do human annotators pay a psychological price for NLP progress? Overall, our experimental findings offer reassuring evidence that large-scale corpus annotation does not inherently pose a mental health toll on annotators, at least as long as safeguards are embedded in the workflow to meet annotators' psychological needs. Over two sets of four-month periods of intensive data annotation, we observed no meaningful declines in annotators' self-reported mental health and PTSD symptoms. Although our focus was on mental health and PTSD symptoms, we also found null effects for other psychological variables, including social biases, political attitudes, cognitive flexibility, and moral values.

Notably, null effects merit cautious interpretation as two features of our protocol likely mitigated potential stressors. First, the corpus was intentionally balanced: emotionally neutral movie-review snippets were interleaved with aversive content (e.g., hate-speech excerpts). Randomizing text order prevented prolonged exposure to any single content type, plausibly buffering annotators against cumulative negative emotional reactions. Second, our training and annotator-centric procedures were grounded in theories of motivation and well-being, embedding safeguards that proactively address the psychological needs of annotators. Our protocol thus offers practical recommendations demonstrating how mental health costs of large-scale annotation can be avoided by theory-driven safeguards. By embedding these psychological principles into the annotation pipeline, researchers can not only produce high-quality labeled data, but also uphold their ethical responsibility to the annotators.

This approach improves the transparency of datasets and also allows for further analysis of individual biases and mental health indicators (Prabhakaran et al., 2021). Moreover, we recommend that researchers incorporate steps (e.g., an iterative feedback loop in which annotators can report the psychological impact of their work) to minimize negative effects (Kennedy et al., 2022). Large Language Models (LLMs) are another method of remediation, increasingly taking on annotator roles alongside annotators (Holter and Ell, 2023) or independently (Rathje et al., 2024; Tan et al., 2024). However, researchers should ensure the validation of annotations performed by LLMs, as they may not capture the subjectiveness of certain psychological constructs (Abdurahman et al., 2024). As the field continues to build classifiers and fine-tune existing models using annotated data, it is essential to keep in mind that the data (Atari et al., 2023b) and how annotators interact with it (Davani et al., 2023) shape the quality of machine learning models.

6 Limitations

Our study focused on the psychological impact of large-scale annotations on annotators; however, we did not assess the extent to which annotators engaged with the annotation tasks or how accurate their annotations were. Although we implemented rigorous protocols, we did not directly measure the level of attention given to the annotation process. Future research could incorporate additional measures (e.g., sustained attention checks, task persistence, and response time analysis) (Langner and Eickhoff, 2013) to capture the extent to which annotators efficiently annotated the corpora.

Our annotation corpus also spanned a broad spectrum of texts ranging from often-neutral movie reviews to highly offensive hate-speech excerpts, which contrasts with the decontextualized, token-based annotations that many NLP projects employ (e.g., part-of-speech tagging, syntactic parsing). Importantly, we observed no reliable decline in annotators' mental health when processing such emotionally taxing and cognitively demanding content. Given these results, it is plausible that annotating less emotionally taxing, more linguistically focused corpora would similarly have a minimal psychological impact. Future research should directly compare these different annotation tasks to confirm that lighter workflows remain manageable using similar protocols. Moreover, unlike corpora focused solely on offensive-language detection (Albanyan et al., 2023; Albanyan and Blanco, 2022; Toraman et al., 2022), our design combined neutral texts with offensive and hateful language excerpts; while offensive-language corpora might elicit greater distress, our findings suggest that buffering emotionally heavy content with neutral snippets could potentially mitigate the psychological toll.

Additionally, our study employed a quasi-experimental design with a total sample of 65 participants. While this relatively small sample size limits statistical power and generalizability, we were still able to detect small to medium effect sizes, underscoring the relevance of our findings. Additionally, our sample consisted exclusively of English-speaking undergraduate students from a large public university in the U.S., which may not fully represent the more diverse populations typically engaged in online annotation platforms (Henrich et al., 2010; Park et al., 2025; Nguyen et al., 2020; Lee et al., 2024; Hovy and Prabhumoye, 2021). Moreover, there is a potential selection bias in the experimental group due to non-random assignment (Joyce et al., 2017). For instance, individuals who volunteered as research assistants may have been more hard-working, ambitious, or interested in culture and computational social science, compared with the control group, which we recruited from the subject pool in our department. Future research should address these limitations by implementing true randomization to enhance the external validity of these findings.

We also did not have sufficient statistical power to examine treatment effects within specific subgroups (e.g., we only had two nonbinary individuals in our annotators). For instance, it is possible that individuals from minority demographic or cultural backgrounds may be especially vulnerable to the mental health toll of large-scale annotation tasks. Future studies with larger annotator samples would allow for more robust subgroup analyses and a better understanding of differential effects across populations.

Lastly, our study focused on a short-term evaluation of the psychological effects of the annotation task, without assessing long-term outcomes. This limitation prevents us from determining whether these (null) effects persist beyond task completion. Future research should expand on these findings to investigate the durability of these psychological outcomes over time to better understand their lasting impact.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

CJ Adams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge. `https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`. Kaggle.

Senem Aktas. 2023. Nostalgic sentiment analysis of youtube comments dataset.

Abdullah Albanyan and Eduardo Blanco. 2022. Pinpointing fine-grained relationships between hateful tweets and replies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10418–10426.

Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023. Not all counterhate tweets elicit the same replies: A fine-grained analysis. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 71–88, Toronto, Canada. Association for Computational Linguistics.

Maryam M. AlEmadi and Wajdi Zaghouani. 2024. Emotional toll and coping strategies: Navigating the effects of annotating hate speech data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 66–72, Torino, Italia. ELRA and ICCL.

Bob Altemeyer. 1996. *The Authoritarian Specter*. Harvard University Press.

Anonymous Authors. 2025. The cultural and moral expressions in language (camel) corpus. Manuscript under review.

Mohammad Atari, Jesse Graham, and Morteza Dehghani. 2020. Foundations of morality in iran. *Evolution and Human Behavior*, 41(5):367–384.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023a. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023b. Which humans?

Mohammad Atari and Aliah Zewail. 2025. Morality across time and space. *European Review of Social Psychology*, pages 1–36.

Jordan R. Axt. 2018. The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*, 9(8):896–906.

Mahzarin R. Banaji and Anthony G. Greenwald. 2013. *Blindspot: Hidden Biases of Good People*. Delacorte Press, New York, NY.

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024. Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86, St Julians, Malta. Association for Computational Linguistics.

Matt Boraske. 2022. Aita subreddit dataset.

William J. Browne and David Draper. 2006. A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Cristian Candia, Mohammad Atari, Nour Kteily, and Brian Uzzi. 2022. Overuse of moral language dampens content engagement on social media. *Under Review*.

S. Ceylan Batur and N. Sakallı. 2019. Honor scale: A study for reliability and validity.

Nilanjana Dasgupta. 2009. Mechanisms underlying the malleability of implicit prejudice and stereotypes. *Handbook of prejudice, stereotyping, and discrimination*, pages 267–284.

Nilanjana Dasgupta, Debbie E McGhee, Anthony G Greenwald, and Mahzarin R Banaji. 2000. Automatic preference for white americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, 36(3):316–328.

A. M. Davani, M. Atari, B. Kennedy, and M. Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. *arXiv preprint arXiv:2404.10857*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *Preprint*, arXiv:2005.00547.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charles R Figley. 2002. Compassion fatigue: Psychotherapists' chronic lack of self care. *Journal of clinical psychology*, 58(11):1433–1441.

Silvana Galderisi, Andreas Heinz, Marianne Kastrup, Julian Beezhold, and Norman Sartorius. 2015. Toward a new definition of mental health. *World Psychiatry*, 14(2):231.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *Preprint*, arXiv:1908.07898.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464.

Social Grep. 2022. The 2022 trucker strike dataset.

James J Gross and Oliver P John. 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348.

James J. Gross and Oliver P. John. 2012. Emotion regulation questionnaire. *Journal of Personality and Social Psychology*.

Christian Haerpfer, Ronald Inglehart, Alberto Moreno, Christian Welzel, Katerina Kizilova, Juan Diez-Medrano, and Birgitta Puranen. 2020. World values survey wave 7 (2017-2020) cross-national data-set. Retrieved from `https://www.worldvaluessurvey.org/`.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Ole Magnus Holter and Basil Ell. 2023. Human-machine collaborative annotation: A case study with GPT-3. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 193–206, Vienna, Austria. NOVA CLUNL, Portugal.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

Ted Joyce, Dahlia K. Remler, David A. Jaeger, Onur Altindag, Stephen D. O'Connell, and Sean Crockett. 2017. On measuring and reducing selection bias with a quasi-doubly randomized preference trial. *Journal of Policy Analysis and Management*, 36(2):438–459.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27.

Pushyami Kaveti and Md Navid Akbar. 2020. Role of intrinsic motivation in user interface design to enhance worker performance in amazon mturk. In *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 1–7.

B. Kennedy, M. Atari, A. M. Davani, L. Yeh, A. Omrani, Y. Kim, and M. Dehghani. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

M. J. Kirton. 1981. A reanalysis of two scales of tolerance of ambiguity. *Journal of Personality Assessment*, 45(4):407–414.

Ross Deans Kristensen-McLachlan, Miceal Canavan, Marton Kárdos, Mia Jacobsen, and Lene Aarøe. 2025. Are chatbots reliable text annotators? sometimes. *PNAS nexus*, 4(4):pgaf069.

Elizabeth J. Krumrei-Mancuso and Stephen V. Rouse. 2016. The development and validation of the comprehensive intellectual humility scale. *Journal of Personality Assessment*, 98:209–221.

Nour Kteily, Arnold K Ho, and Jim Sidanius. 2012. Hierarchy in the mind: The predictive power of social dominance orientation across social contexts and domains. *Journal of Experimental Social Psychology*, 48(2):543–549.

Robert Langner and Simon B Eickhoff. 2013. Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. *Psychological bulletin*, 139(4):870.

Mark R Leary, Kate J Diebels, Erin K Davisson, Katrina P Jongman-Sereno, Jennifer C Isherwood, Kaitlin T Raimi, Samantha A Deffler, and Rick H Hoyle. 2017. Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, 43(6):793–813.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. *Preprint*, arXiv:2308.16705.

N. Ljubešić, I. Mozetič, and P. K. Novak. 2023. Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, 29(6):1481–1494.

Judith Lukat, Jürgen Margraf, Reinhard Lutz, Willem M. van der Veld, and Eni S. Becker. 2016. Psychometric properties of the positive mental health scale (pmh-scale). *BMC Psychology*, 4:1–14.

Hazel Rose Markus and Shinobu Kitayama. 2014. Culture and the self: Implications for cognition, emotion, and motivation. In *College student development and academic life*, pages 264–293. Routledge.

I Lisa McCann and Laurie Anne Pearlman. 1990. Vicarious traumatization: A framework for understanding the psychological effects of working with victims. *Journal of traumatic stress*, 3:131–149.

Johan YY Ng, Nikos Ntoumanis, Cecilie Thøgersen-Ntoumani, Edward L Deci, Richard M Ryan, Joan L Duda, and Geoffrey C Williams. 2012. Self-determination theory applied to health contexts: A meta-analysis. *Perspectives on psychological science*, 7(4):325–340.

Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, 3.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Math= male, me= female, therefore math≠ me. *Journal of personality and social psychology*, 83(1):44.

Bunmi O. Olatunji, Nathan L. Williams, David F. Tolin, Jonathan S. Abramowitz, Christopher N. Sawchuk, Jeffrey M. Lohr, and Laura S. Elwood. 2007. The disgust scale: Item analysis, factor structure, and suggestions for refinement. *Psychological Assessment*, 19(3):281.

Danny Osborne, Thomas H Costello, John Duckitt, and Chris G Sibley. 2023. The psychological causes and societal consequences of authoritarianism. *Nature reviews psychology*, 2(4):220–232.

Daphna Oyserman. 2011. Culture as situated cognition: Cultural mindsets, cultural fluency, and meaning making. *European review of social psychology*, 22(1):164–214.

Daphna Oyserman. 2017. Culture three ways: Culture and subcultures within countries. *Annual review of psychology*, 68(1):435–463.

Alexander Pak, Patrick Paroubek, et al. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, 2010, pages 1320–1326.

Alessandro Palla. 2024. Chatbot instruction prompts dataset.

Junyeong Park, Seogyeong Jeong, Seyoung Song, Yohan Lee, and Alice Oh. 2025. Llm-c3mod: A human-llm collaborative system for cross-cultural hate speech moderation. *Preprint*, arXiv:2503.07237.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Felicia Pratto, Jim Sidanius, Lonnie M. Stallworth, and Bertram F. Malle. 1994. Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, 67:741–763.

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. 2024. Moralbert: A fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, page 433–442. ACM.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. 2024. Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Gil Rosenthal. 2023. English historical quotes dataset.

Aymeric Roucher. 2023. English historical quotes dataset.

Richard M Ryan and Edward L Deci. 2017. *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford publications.

Rachel H Salk, Janet S Hyde, and Lyn Y Abramson. 2017. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological bulletin*, 143(8):783.

Carol Sansone and Dustin B Thoman. 2005. Interest as the missing motivator in self-regulation. *European Psychologist*, 10(3):175–186.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Ferdinand Schlatt. 2023. Trump tweets dataset.

Simone Schnall, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan. 2008. Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34(8):1096–1109. PMID: 18505801.

Theodore M Singelis, Harry C Triandis, Dharm PS Bhawuk, and Michele J Gelfand. 1995. Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-cultural research*, 29(3):240–275.

Gavin R Slemp, James G Field, Richard M Ryan, Vivien W Forner, Anja Van den Broeck, and Kelsey J Lewis. 2024. Interpersonal supports for basic psychological needs and their relations with motivation, well-being, and performance: A meta-analysis. *Journal of Personality and Social Psychology*.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 17(4).

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

Teodor Stoev, Kristina Yordanova, and Emma L. Tonkin. 2023. Experiencing annotation: Emotion, motivation and bias in annotation tasks. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 534–539.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *Preprint*, arXiv:2402.13446.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The moral foundations reddit corpus. *Preprint*, arXiv:2208.05545.

Harry C. Triandis and Michele J. Gelfand. 1998. Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology*, 74(1):118–128.

Harry C Triandis and Michele J Gelfand. 2012. A theory of individualism and collectivism. *Handbook of theories of social psychology*, 2.

Ayse K Uskul, Susan E Cross, and Ceren Günsoy. 2023. The role of honour in interpersonal, intrapersonal and intergroup processes. *Social and Personality Psychology Compass*, 17(1):e12719.

Eric-Jan Wagenmakers, Jonathon Love, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F Gronau, Damian Dropmann, Bruno Boutin, et al. 2018. Bayesian inference for psychology. part ii: Example applications with jasp. *Psychonomic bulletin & review*, 25:58–76.

Stephanie Wallace, James Nazroo, and Laia Bécares. 2016. Cumulative effect of racial discrimination on the mental health of ethnic minorities in the united kingdom. *American journal of public health*, 106(7):1294–1300.

Gregory M Walton and Geoffrey L Cohen. 2007. A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology*, 92(1):82.

Frank W. Weathers. 2013. The ptsd checklist for dsm-5 (pcl-5).

Joanna L. Wolfe. 2000. Effects of annotations on student readers and writers. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 19–26.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification. *Preprint*, arXiv:1509.01626.

Appendix

A Primary Measures

A.1 Mental Health

Mental Health. Mental health is characterized by individuals' general emotional, psychological, and social well-being (Lukat et al., 2016). We measured mental health using nine items on a 5-point scale anchored by "Not True" (1) to "True" (5). An example item is, "I feel that I am actually well equipped to deal with life and its difficulties."

Responses to these items were averaged to create a composite score (Cronbach's $\alpha = 0.90$).

Post Traumatic Stress Disorder. We used the 20-item PTSD Checklist for DSM-5 (PCL-5) (Weathers, 2013) to assess symptoms of anxiety resulting from stressful or distressing events, commonly associated with PTSD. An example of the item is "Having strong negative feelings such as fear, horror, anger, guilt, or shame." All items were rated on a 5-point scale ranging from "Not at All" (1) to "A great deal" (5). Responses to these items were averaged to create a composite score (Cronbach's $\alpha = 0.94$).

B Additional Measures

In addition to our primary focus on mental health and PTSD-related symptoms, we included a broader set of psychological measures to explore whether large-scale annotation tasks might subtly influence other aspects of cognition, identity, or social perception. We also measured these variables because we intended to collect a variety of individual-difference measures from annotators, inspired by prior recommendations (Prabhakaran et al., 2021). These measures were organized into four superordinate domains, each targeting a theoretically meaningful dimension of psychological functioning.

First, we assessed psychological flexibility, a collection of constructs with significant importance in clinical psychology, reflecting an individual's ability to adapt to difficult thoughts and feelings while maintaining goal-directed behavior. Second, we included a battery of measures tapping into social biases. These included both explicit self-report items and implicit tests designed to capture preferences toward particular racial and gender groups. Third, to explore how participants relate to cultural norms and self-construal, we measured cultural orientation. Finally, we assessed moral values. These values can shape how annotators interpret morally charged language, and may themselves be influenced by prolonged exposure to polarized or norm-violating content. Together, these four domains allowed us to probe whether the annotation experience has downstream psychological effects beyond immediate well-being and PTSD symptoms.

B.1 Psychological Flexibility

Tolerance of Ambiguity. This construct refers to the degree to which individuals prefer ambiguity over familiarity (Kirton, 1981). We measured tolerance of ambiguity using 13 items on a 5-point scale, ranging from "Strongly Disagree" (1) to "Strongly Agree" (5). An example item is, "I generally prefer novelty over familiarity." Responses were averaged to create a composite score (Cronbach's $\alpha = 0.77$).

Emotion Regulation. This construct refers to the ability of individuals to express and modulate both and negative emotions (Gross and John, 2003). We used a 10-item scale to measure emotion regulation (Gross and John, 2012). An example of the item is "I control my emotions by changing the way I think about the situation I'm in." All items were rated on a 7-point scale anchored by "Strongly Disagree" (1) to "Strongly Agree" (7). Responses to these items were averaged to create a total score (Cronbach's $\alpha = 0.79$).

Perspective Taking. This construct refers to the tendency of individuals to understand the thoughts, feelings, and motivations of others in interpersonal situations (Davis, 1983). We used a 14-item scale to measure perspective taking. An example of the item is "Before criticizing somebody, I try to imagine how I would feel if I were in their place." All items were rated on a 5-point scale anchored by "Does not describe me well" (1) to "Describes me well" (5). Responses to these items were averaged to create a composite score (Cronbach's $\alpha = 0.85$).

B.2 Social Biases

Implicit gender and racial biases. Implicit biases are automatic, unconscious associations that influence human attitudes and behavior (Banaji and Greenwald, 2013). We used a Science–Gender Implicit Association Test to assess the automatic association between science (e.g., "hypothesis," "data") versus liberal-arts concepts (e.g., "essay," "grammar") and male ("he," "male," "boy") versus female pronouns ("she," "woman," "girl") (Greenwald et al., 1998). After three practice rounds, participants completed four alternating test blocks of stereotype-congruent and incongruent pairings. The order of these blocks was counterbalanced following the procedure (Nosek et al., 2002). Positive *D*-scores indicated a stronger implicit stereotype, resulting from faster responses when male terms were paired with science and female terms with liberal arts. Conversely, negative scores reflected faster responses for female-science/male-arts pairings. We also administered a Race IAT to evaluate automatic associations between racial categories ("European American" vs. "African American")

and evaluative attributes ("pleasant" vs. "unpleasant") (Dasgupta et al., 2000). Participants categorized faces and words across seven blocks: three for practice and four for testing. Congruent (e.g., European American + pleasant) and incongruent (e.g., African American + pleasant) trial types were counterbalanced across participants. An individual's *D*-score was calculated as the difference in average response latencies between congruent and incongruent blocks, standardized by the pooled variance. Higher positive scores reflected a stronger implicit preference for European American over African American stimuli (Dasgupta, 2009).

Explicit Racial Bias. We used a 2-item scale (Axt, 2018) to measure individuals' beliefs about Black and White Americans. An example item is: "The How warm or cold do you feel toward Black Americans?" All items were rated on a 10-point scale, ranging from "Extremely Cold" (1) to "Extremely Warm" (5). Responses were averaged to create a composite score (Cronbach's $\alpha = 0.84$).

Right-Wing Authoritarianism. This construct refers to an ideological orientation that prefers stricter authority, traditional norms, and punitive measures (Altemeyer, 1996; Osborne et al., 2023). We used a 15-item scale (Altemeyer, 1996) to measure Right-Wing Authoritarianism. An example item is: "The "old-fashioned ways" and "old-fashioned" values' still show the best way to live." All items were rated on a 5-point scale, ranging from "Strongly Disagree" (1) to "Strongly Agree" (5). Responses were averaged to create a composite score (Cronbach's $\alpha = 0.84$).

Social Dominance Orientation. This construct refers to the extent to which individuals prefer group hierarchies and inequality (Kteily et al., 2012). We used a 17-item scale (Pratto et al., 1994) to measure social dominance orientation. An example item is: "Some groups of people are simply not the equals of others." All items were rated on a 7-point scale, ranging from "Strongly Disagree" (1) to "Strongly Agree" (7). Responses were averaged to create a composite score (Cronbach's $\alpha = 0.86$).

B.3 Cultural Orientations

Individualism & Collectivism. The cultural orientations of individualism and collectivism are cental to understanding how people define themselves, relate to others, and make sense of the social world (Markus and Kitayama, 2014; Oyserman, 2017, 2011; Triandis and Gelfand, 2012). These two constructs can be categorized into four key components: horizontal individualism (the extent to which individuals perceive themselves as independent from their groups; e.g., "I want to do my own thing"; (Cronbach's $\alpha = 0.82$)), vertical individualism (the extent to which individuals seek power and status through competition; e.g., "I want to be the best"; (Cronbach's $\alpha = 0.58$)), horizontal collectivism (the extent to which individuals see themselves as part of a group; e.g., "The well-being of my co-workers is important to me"; (Cronbach's $\alpha = 0.62$)), and vertical collectivism (the extent to which individuals are willing to sacrifice personal goals for the benefit of their group; e.g., "I would sacrifice an enjoyable activity if my family disapproved"; (Cronbach's$\alpha = 0.80$)) (Triandis and Gelfand, 1998; Singelis et al., 1995). We used a 15-item scale (Triandis and Gelfand, 1998) to measure these components. All items were rated on a 10-point scale, ranging from "Never" (1) to "Always" (10). Responses were averaged to create composite scores.

Honor. This construct assess the extent to which individuals experience distress over damaging their own reputation or that of their family, as well as their perceived inability to protect women's honor and maintain moral integrity (Uskul et al., 2023). We used a 30-item scale (Ceylan Batur and Sakallı, 2019) to measure honor. An example of the item is "If you were unable to protect your partner when someone harasses him/her." All items were rated on a 5-point scale anchored by "Not at All bad" (1) to "Extremely bad" (5). Responses to these items were averaged to create a composite measure(Cronbach's$\alpha = 0.83$)).

Moral Values

Moral Foundations. Moral Foundations Theory posits that there are several dimensions to human morality, each of which serves an evolutionary purpose(Haidt and Joseph, 2004). These moral foundations are intuitive, culturally shaped, and deeply rooted in evolved psychological systems that enabled cooperation (Graham et al., 2013). We used the 36-item Moral Foundation Questionnaire-2 (Atari et al., 2023a) to assess moral concerns: care (e.g., "It pains me when I see someone ignoring the needs of another human being."; (Cronbach's $\alpha = 0.92$)), equality (e.g., "I believe that everyone should be given the same quantity of resources in life."; (Cronbach's $\alpha = 0.75$)), loyalty (e.g., "I think children should be taught to be loyal to their country."; (Cronbach's $\alpha = 0.69$)), authority (e.g., "I believe that one of the most important

values to teach children is to have respect for authority."; (Cronbach's $\alpha = 0.79$)), purity (e.g., "I believe chastity is an important virtue."; (Cronbach's $\alpha = 0.59$)), and proportionality (e.g., "It makes me happy when people are recognized for their merits."; (Cronbach's $\alpha = 0.75$)). All items were rated on a 5-point scale, ranging from "Does not describe me at all" (1) to "Describes me extremely well" (5).

Qeirat. This construct is characterized by moral concerns for one's family, especially female kin, extended family members, and by extension, their country. This construct, closely related to but different from honor, can be considered an additional moral foundation (Atari and Zewail, 2025). We used an 8-item scale (Atari et al., 2020) to quantify participants' endorsement of Qeirat values. An example of the item is "A man should be responsive to his family's needs." All items were rated on a 7-point scale anchored by "Strongly Disagree" (1) to "Strongly Agree" (7). Responses to these items were averaged to create a composite score (Cronbach's $\alpha = 0.80$).

Disgust. Disgust is a moral emotion and has been found to be highly related to moral judgments (Schnall et al., 2008). We used a 13-item scale (Olatunji et al., 2007) to assess the extent to which individuals feel an aversion to various unpleasant or aversive contexts. An example item is: "It would not upset me at all to watch a person with a glass eye take the eye out of the socket." All items were rated on a 5-point scale, ranging from "Strongly Disagree" (1) to "Strongly Agree" (5). Responses were averaged to create a composite score (Cronbach's $\alpha = 0.86$).

Trust. This construct refers to the extent to which individuals expect others to act in an honest and reliable manner. We used a 6-item scale (Haerpfer et al., 2020) to assess the extent to which individuals trust other people. An example item is: "Most people are basically good and kind." All items were rated on a 5-point scale, ranging from "Strongly Disagree" (1) to "Strongly Agree" (5). Responses were averaged to create a composite measure (Cronbach's $\alpha = 0.78$).

Intellectual Humility. Intellectual humility is characterized by the recognition that one's knowledge and beliefs might be limited or incorrect, and being open to revise one's views in light of new evidence (Leary et al., 2017). We used a 22-item scale (Krumrei-Mancuso and Rouse, 2016) to quantify intellectual humility. An example of the item is "When I am really confident in a belief, there is very little chance that belief is wrong." All items were rated on a 5-point scale anchored by "Strongly Disagree" (1) to "Strongly Agree" (5). Responses to these items were averaged to create a composite score (Cronbach's $\alpha = 0.86$).

C Demographic Match Analysis

To assess whether the experimental and control groups were comparable in terms of demographic characteristics, chi-square tests were conducted for gender and race, and an independent samples t-test was performed for age. A chi-square test of independence indicated that gender distribution did not significantly differ between the experimental and control groups, $\chi^2(1, N = 65) = 0.169$, $p = .681$. Similarly, race distribution did not significantly differ between groups, $\chi^2(1, N = 65) = 0.482$, $p = .488$. However, an independent-samples *t*-test revealed a significant difference in age between the experimental group ($M = 20.80$, $SD = 1.73$) and the control group ($M = 19.91$, $SD = 1.36$), $t(63) = 2.31$, $p = .024$, suggesting that participants in the experimental group were slightly older than those in the control group. This difference was primarily driven by a subset of research assistants (10% of the sample) in the experimental group who were between 24 and 26 years old.

D Institutional Review Board (IRB

Our IRB includes the anonymity of the participants' data and ensures minimal risk to them during and after the experiment.

E Data & Code

Data, code and annotation materials are publicly available at `https://anonymous.4open.science/r/Annotations-DCA6`.

Table 1: The CAMEL Corpus contains data extracted from various datasets. Below is a summary of the sources.

Corpus	Description	*N*	Source
Moral Foundations Reddit Corpus	Reddit posts related to moral discourse	8,000	Trager et al. (2022)
AITA Subreddit	Submissions from the "Am I The A-hole¿' subreddit along with the top 10 comments for each submission	2,000	Boraske (2022)
The 2022 Trucker Strike	Reddit posts capturing the discourse of the 2022 trucker strike	1,000	Grep (2022)
Hate Speech and Offensive Language Dataset	Dataset containing X (Twitter) posts considered racist, sexist, homophobic, or generally offensive.	2,000	Davidson et al. (2017)
Jigsaw Toxic Comment Classification Challenge	A subset of Wikipedia comments focusing on negative online behaviors such as toxic comment	2,000	Adams et al. (2017)
AG News Dataset	News articles spanning four topics: "World," "Sports," "Business," "Science and Technology"	4,000	Zhang et al. (2016)
English Historical Quotes Dataset	A collection of random historical quotes	5,000	Roucher (2023)
Stanford Sentiment Treebank	Movie reviews	8,000	Socher et al. (2013)
Chatbot Instruction Prompts Dataset	Prompts and responses for chatbot instructions	2,000	Palla (2024)
Nostalgic Sentiment Analysis of YouTube Comments Dataset	Nostalgic sentiment expressed in Youtube comments	1,000	Aktas (2023)
Moral Stories Dataset	A series of short stories showcasing moral and ethical dilemmas	4,000	Emelin et al. (2021)
Candia et al. Dataset	online posts from X (Twitter), 8Chan, and Reddit	9,000	Candia et al. (2022)
Latin-English Parallel Corpus	English translations of historical Latin texts	1,000	Rosenthal (2023)
Trump Tweets Dataset	A collection of tweets from former President Donald Trump	1,000	Schlatt (2023)
GoEmotions Dataset	Texts spanning 27 emotion categories such as admiration, disappointment, and relief	1,000	Demszky et al. (2020)

Table 2: Effects of Annotation Tasks on Annotator's Psychological Outcomes

	B	SE	p
Mental Health	0.11	0.09	0.223
PTSD	-0.12	0.16	0.455
			$^{*}p<0.05$; $^{**}p<0.01$

Note. Frequentist analysis of the interaction effects of time×group on the listed variables.

Table 3: Bayesian Hierarchical Model Results: Estimates, 95% Credible Intervals, Bayes Factors, and Interpretation

Outcome	Estimate	95% Credible Interval	BF_{10}	BF_{01}	Interpretation
Mental Health	0.11	$[-0.06,\ 0.29]$	0.02	50.00	Very strong evidence for null
PTSD	-0.12	$[-0.45,\ 0.20]$	0.02	50.00	Very strong evidence for null

Note. BF_{10} = Bayes factor favoring the alternative hypothesis; BF_{01} = $1/BF_{10}$ favoring the null hypothesis. 1–3 "anecdotal," 3–10 "moderate," 10–30 "strong," 30–100 "very strong," >100 "extreme."

Table 4: Frequentist Treatment Effects in Additional Psychological Measures

Variable	β_3	SE	p
Psychological Flexibility			
Tolerance of Ambiguity	-0.16	0.12	0.188
Emotion Regulation	0.01	0.22	0.961
Intellectual Humility	0.00	0.10	0.917
Perspective Taking	-0.00	0.10	0.990
Social Biases			
Right-Wing Authoritarianism	-0.19	0.07	0.008^{**}
Social Dominance Orientation	-0.15	0.14	0.375
Explicit Black Preference	0.03	0.36	0.913
Explicit White Preference	-0.14	0.41	0.500
Implicit Gender Bias	0.01	0.10	0.917
Implicit Racial Bias	-0.00	0.10	0.990
Culture Orientations			
Horizontal Individualism	-0.16	0.43	0.765
Vertical Individualism	-0.17	0.38	0.975
Horizontal Collectivism	-0.12	0.27	0.760
Vertical Collectivism	-0.88	0.38	0.038^{*}
Male Honor	0.17	0.11	0.094
Female Honor	0.06	0.10	0.642
Moral Values			
Care	0.03	0.17	0.846
Equality	0.17	0.14	0.303
Loyalty	0.17	0.13	0.305
Proportionality	-0.07	0.16	0.500
Authority	0.04	0.14	0.925
Purity	-0.09	0.12	0.901
Qeirat	-0.09	0.22	0.569

Note. Each entry shows the time×group interaction effect. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 5: Bayesian Treatment Effects in Additional Psychological Measures

Variable	Estimate	95% Credible Interval	BF_{10}	BF_{01}	Interpretation
Psychological Flexibility					
Tolerance of Ambiguity	-0.16	$[-0.39, 0.08]$	0.03	33.33	Very strong evidence for null
Emotion Regulation	0.02	$[-0.53, 0.58]$	0.02	50.00	Very strong evidence for null
Social Biases					
Right-Wing Authoritarianism	-0.20	$[-0.34, -0.06]$	0.25	4.00	Moderate evidence for null
Social Dominance Orientation	-0.14	$[-0.42, 0.15]$	0.02	50.00	Very strong evidence for null
Explicit Black Preference	0.00	$[-0.74, 0.75]$	0.03	33.33	Very strong evidence for null
Explicit White Preference	-0.27	$[-1.09, 0.56]$	0.04	25.00	Strong evidence for null
Implicit Gender Bias	0.01	$[-0.17, 0.19]$	0.02	50.00	Very strong evidence for null
Implicit Racial Bias	0.00	$[-0.18, 0.18]$	0.03	33.33	Very strong evidence for null
Culture Orientations					
Horizontal Individualism	-0.16	$[-0.44, -1.02]$	0.05	20.00	Strong evidence for null
Vertical Individualism	-0.03	$[-0.78, 0.71]$	0.04	25.00	Strong evidence for null
Horizontal Collectivism	-0.10	$[-0.87, 0.67]$	0.04	25.00	Strong evidence for null
Vertical Collectivism	-0.78	$[-1.54, -0.02]$	0.29	3.45	Moderate evidence for null
Male Honor	0.18	$[-0.04, 0.39]$	0.04	25.00	Strong evidence for null
Female Honor	0.04	$[-0.16, 0.25]$	0.01	100.00	Very strong evidence for null
Moral Values					
Care	0.03	$[-0.32, 0.39]$	0.02	50.00	Very strong evidence for null
Equality	0.21	$[-0.09, 0.50]$	0.04	25.00	Strong evidence for null
Loyalty	0.20	$[-0.09, 0.49]$	0.04	25.00	Strong evidence for null
Proportionality	-0.07	$[-0.41, 0.27]$	0.02	50.00	Very strong evidence for null
Authority	0.05	$[-0.26, 0.35]$	0.02	50.00	Very strong evidence for null
Purity	-0.05	$[-0.29, 0.20]$	0.01	100.00	Very strong evidence for null
Qeirat	-0.12	$[-0.57, 0.33]$	0.02	50.00	Very strong evidence for null

Note. BF_{10}= Bayes factor favoring the alternative; BF_{01} = $1/BF_{10}$ favoring the null. 1–3 "anecdotal," 3–10 "moderate," 10–30 "strong," 30–100 "very strong," $>$100 "extreme."

Table 6: Descriptive Statistics for Mental Health, Psychological Flexibility, and Social Biases

Mental Health and Psychological Flexibility					Social Biases				
	M_{Time_1}	SD_{Time_1}	M_{Time_2}	SD_{Time_2}		M_{Time_1}	SD_{Time_1}	M_{Time_2}	SD_{Time_2}
Mental Health	3.04	.56	2.88	.49	Right- Wing Authoritarianism	2.22	.52	2.15	.57
PTSD	1.91	.64	2.17	.71	Social Dominance Orientation	2.11	.73	2.24	.75
Intellectual Humility	3.76	0.41	3.73	.41	Explicit White Preference	7.94	2.18	7.79	1.81
Perspective Taking	3.27	.32	3.19	0.41	Explicit Black Preference	8.56	2.15	8.34	1.72
Emotion Regulation	4.85	0.86	4.98	0.84	Implicit Gender Bias	0.32	0.36	0.25	0.53
Perspective Taking	3.27	1.11	3.19	1.26	Implicit Racial Bias	0.07	0.44	0.06	0.43

Note. Well-being was measured on a 1-4 scale, while PTSD symptoms, Intellectual Humility, Perspective Taking, and Right-Wing Authoritarianism (RWA) were measured on a 1-5 scale. Emotion Reappraisal, Suppression, and Social Dominance Orientation (SDO) were measured on a 1-7 scale. Explicit racial preferences were measured on a 1-10 scale, and implicit biases were measured as association scores ranging from -1 to +1. For all variables, lower values indicate low endorsement and higher values indicate high endorsement.

Table 7: Descriptive Statistics for Moral Values and Cultural Orientations

Moral Values					Cultural Orientation				
	M_{Time_1}	SD_{Time_1}	M_{Time_2}	SD_{Time_2}		M_{Time_1}	SD_{Time_1}	M_{Time_2}	SD_{Time_2}
Care	3.98	.86	3.91	0.72	Horizontal Individualism	8.13	1.55	8.12	1.49
Equality	2.65	.83	2.67	.84	Vertical Individualism	8.56	1.04	8.36	1.10
Proportionality	3.67	.64	3.60	.70	Horizontal Collectivism	7.10	1.79	7.06	1.61
Loyalty	2.83	.64	2.86	.66	Vertical Collectivism	5.98	1.59	5.97	1.28
Authority	2.65	0.67	2.64	.84	Male Honor	4.35	0.27	4.14	.37
Purity	2.29	0.66	2.23	.66	Female Honor	3.41	.53	3.36	.49
Trust	3.36	.67	3.33	.70	Moral Integrity	4.35	.31	4.18	.42
Disgust	3.05	.51	3.10	.52	Qeirat	4.75	.94	4.56	1.02

Note. Moral Values, Trust, Disgust, Honor, and Qeirat were measured on a 1–5 scale, and Individualism and Collectivism were measured on a 1–10 scale. For all variables, lower values indicate low endorsement and higher values indicate high endorsement.