Whose Values? Measuring the (Subjective) Expression of Basic Human Values in Social Media Posts

Ziv Epstein[1], Farnaz Jahanbakhsh[2], Tiziano Piccardi [3], Isabel Gallegos [4], Dora Zhao [4], Johan Ugander [5], Michael Bernstein [4]

[1]MIT,
[2]University of Michigan,
[3]John Hopkins University,
[4]Stanford University,
[5]Yale University
zive@mit.edu, farnaz@umich.edu, piccardi@jhu.edu, iogalle@stanford.edu, dorothyz@stanford.edu.
johan.ugander@yale.edu, msb@cs.stanford.edu

Abstract

The value alignment of sociotechnical systems has become a central debate, but progress depends on how human values are perceived in the content these systems surface and how such perceptions can be measured at scale. Social media platforms are a prominent class of sociotechnical systems where algorithmic curation shapes exposure to value-laden content at scale. Large-language models offer new opportunities for measuring expressions of human values (e.g., humility or equality) in social media data, but value expressions can be subjective: different people will annotate the same post with different values. In this paper, we draw on the Schwartz value system as a broadly encompassing and theoretically grounded set of basic human values, and introduce a framework to *personalize* the measurement of expressions of Schwartz values in social media posts at scale. We collect 32,370 ground truth value expression annotations from N=1,079 people on 5,211 social media posts representative of real users' feeds. Due to the subjectivity of the task, we observe low levels of inter-rater agreement between people, and low agreement between human raters and LLM-based methods. In response, we construct a personalization architecture for classifying value expressions by learning from a small number of highly informative calibration annotations per user. In evaluation, we find that modeling these differences successfully yields value expression predictions that people agree with more than they agree with other people. These results contribute new methods and understanding for the measurement of human values in social media data.

Code and data —
https://github.com/zivepstein/icwsm2026-values

Introduction

Social media provides a rich and complex substrate to understand patterns of human life, behavior, and discourse (Lazer et al. 2009). The rise of large language models (LLMs) offers new opportunities for measuring a wide variety of constructs in social media data (Ziems et al. 2024; Jia et al. 2024). One construct of interest for social media, growing out of the focus on value alignment of sociotechnical systems (Russell, Dewey, and Tegmark 2015), is *human values*: the core beliefs that serve as guiding principles and in turn shape decisions, behavior, and perception (Kolluri et al. 2026; Bernstein et al. 2023; Stray et al. 2024). Values expressed on social media, therefore, offer a reflection of discourse and sentiment, that in turn affect behavior. The possibility of measuring human values in social media posts opens new opportunities to build a better understanding of how discourse operates online, and to design the next generation of social media algorithms and platforms. Past work has shown how values and related moral priorities explain patterns of belief and behavior in domains as varied as vaccine hesitancy (Weinzierl and Harabagiu 2022), pro-environmental behavior (Schultz and Zelezny 1998), social norms (Forbes et al. 2020), prosociality (Heilman and Kusev 2020), human development (Inglehart 2020), consumer behavior (Krystallis, Vassallo, and Chryssohoidis 2012; Stathopoulou and Balabanis 2019), and news story framing (Mokhberian et al. 2020). Beyond understanding, the capability to measure values could also enable algorithmic objectives for feed-based ranking that support both individual and societal goals. In particular, there is growing interest in aligning ranking algorithms with societal values (Bail 2022; Bernstein et al. 2023; Stray et al. 2024; Kolluri et al. 2026), focusing on values such as democratic process (Jia et al. 2024; Piccardi et al. 2024b), well-being (Stray 2020), and downranking low-quality information (Epstein, Pennycook, and Rand 2020).

Before these opportunities can be realized, there remain both conceptual and practical challenges in operationalizing and measuring value expressions in social media. By *value expressions*, we mean the extent to which a social media post expresses attributes associated with a given value. Identifying value expression can be straightforward in cases such as the post "Not all disabilities are visible!" (Figure 1), which many perceive as expressing the value of equality but not tradition.

Figure 1: An anonymized example social media post for value annotation. All annotators in our dataset agree that this post strongly expresses Universal Concern, and they all agree that the post does not express Tradition—but, annotators disagree substantially on whether the post expresses Humility. In this paper, we model these interpersonal differences for more accurate value classification.

However, not in all cases are value expressions as straightforward to classify on social media. The core issue is that value expression is a fundamentally *subjective* construct, leading to different labels from different people (Van Der Meer et al. 2023; Basile et al. 2021; Orlikowski et al. 2023). For example, while the Figure 1 post that "Not every disability is visible!" clearly is an expression of equality and clearly is not an expression of tradition, is it an expression of the value of humility? For some, the post strongly expresses a reminder to be humble about the conclusions we draw about other people's (lack of) visible disabilities. For others, their perception of that value expression is much weaker. Both interpretations are valid and shaped by an individual's background and how they both conceptualize these values and understand the social media post itself. As a result, we argue that there exists conceptual and methodological ambiguity about how value expression is operationalized. While existing work (e.g., Qiu et al. (2022); Van Der Meer et al. (2023); Borenstein et al. (2024)) has developed classifiers for detecting basic human values, these models have presumed there exists in all cases one unambiguous correct answer to be learned. But due to the possibility of divergent interpretations of value expressions, a traditional content-level ground truth set to train and evaluate models may obscure disagreements across individuals.

In this paper, we therefore argue for a paradigm shift toward value labeling architectures that directly acknowledge the subjectivity of these constructs. In particular, we first identify disagreements in perceptions of values in social media posts, and then develop personalized methods to identify an individual's perception of a value under a perspectivist framework. To do this, we collect a large annotated dataset where 1,079 people redundantly annotate values in 5,211 posts using a well-studied and theoretically grounded value system drawn from cross-cultural psychology — the Schwartz value system (Schwartz and Cieciuch 2022; Schwartz 1992). We evaluate to what extent large language models (LLMs) can identify the expression of values in these posts, comparing off-the-shelf LLM models with fine-tuned and personalized models, and both to individuals and groups of annotators.

On this task, we find that not only are there low levels of agreement between people, but also that an off-the-shelf LLM exhibits ever lower levels of agreement with people than people's levels of agreement with each other. To make sense of these findings, we draw on the theory of *perspectivism* — that different people will have different subjective experiences of the same situation (Soden, Toombs, and Thomas 2024; Berger and Luckmann 2016) — to suggest that people's divergent perceptions of both the posts and values are driving the low levels of human-human agreement we observe, as well as the lack of performance of the off-the-shelf (base) GPT model.

With this idea in mind, we explicitly model interpersonal differences by combining two components: (1) fine-tuning a large language model to perform value inference, and (2) incorporating recommendation-system–inspired techniques that use a small, user-specific "cold-start" dataset to adapt predictions to individual users.

We find that our personalized approach results in a performant model that individuals agree with more than the rate at which they agree with other people (66% relative improvement in spearman rank correlation (ρ)), or with the consensus vote (28% relative improvement in ρ).

We first contribute a demonstration that perceptions of value expression exhibit high levels of disagreement, which underscores the perils of relying on majority aggregated labels in this task. We then correspondingly contribute a theoretically-grounded personalization framework for calibrating predictions from a fine-tuned large language model to an individual based on their own responses. This results in predictions that align more closely with that person's judgments than either other humans' labels or few-shot–prompted LLM outputs.

Related Work

Schwartz's Theory of Basic Human Values

To operationalize values, we employ Schwartz's theory of Basic Human Values (Schwartz et al. 2012), a theory from Cultural Psychology for mapping individual values. Schwartz defined (basic) values as "trans-situational goals, varying in importance, that serve as guiding principles in the life of a person or group" and sought to identify a set of values that could be recognized in all societies. The original theory (Schwartz 1992) identified 10 basic values from a 56-item questionnaire administered to 9140 students and teachers across 20 countries. The refined theory identified 19 values that exist on a circular continuum with 57-item

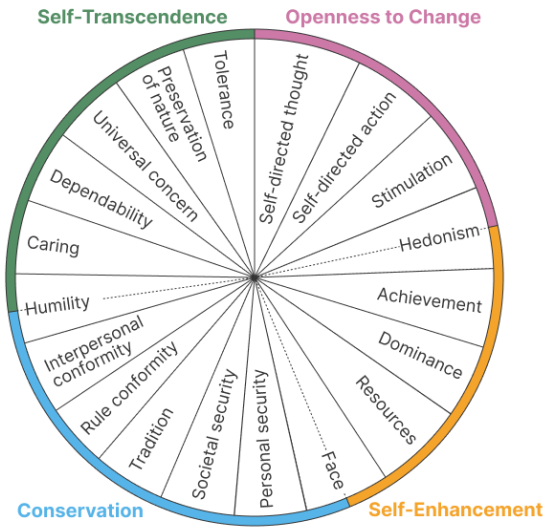

Figure 2: The refined Schwartz value system. The system is hierarchically nested into a highest level: Outcomes for Self (right) vs Outcomes for Others (left), a high level: Self-Transcendence (top left, green), Conservation (bottom left, blue), Self-Enhancement (bottom right, orange) and Openness to Change (top right, red), and the 19 low-level values. Adapted from (Schwartz et al. 2012)

.

questionnaire administered to 6,059 participants across 10 countries (see Figure 2). A core assumption of Schwartz's theory of Basic Human Values is that: "The circular arrangement of values represents a continuum of related motivations, like the circular continuum of colors, rather than a set of discrete motivations" (Davidov, Schmidt, and Schwartz 2008). This circumplex structure is also reflected in the hierarchical branching structure of the values, which starts at the highest level (Outcomes for Self vs Others), then branches further to four high-level values (Openness to Change, Self-Enhancement, Self-Transcendence, and Conservation), and then 19 low-level values. Values close to each other on the wheel are theorized to have similar and compatible underlying motivations, while values on opposite sides are theorized to be in tension. While there are reasons to critique the Schwartz Value System (see e.g. (Chakroff 2015)), the theory ultimately offers conceptual precision, depth, and breadth, as well as offering more theoretical plausibility, analytical utility, and empirical grounding than political values (Goren 2020).

Detecting and Using Schwartz Values

Qiu et al. (2022) create a dataset of human attitudes on over 20,000 text scenarios called VALUENET and a transformer-based regression model for dialogue tasks. Kiesel et al. (2022) create a dataset of the values behind arguments from four geographical cultures. Ponizovskiy et al. (2020) developed a dictionary of words for Schwartz values based on a corpus of Facebook updates, blog posts, essays, and book chapters. Boyd et al. (2015) use free-response survey methods to compare individual values reconstructed from text with traditional survey instruments, showing advancements in understanding in both the structure and content of values Van Der Meer et al. (2023) use data from Qiu et al. (2022) and Kiesel et al. (2022) to train a model for predicting value expression in Reddit posts, which they apply to a dataset of 11.4M comments from 19K users to construct value profiles of individual users. They find that dissimilarity of users' values is associated with the levels of disagreement between those users. Borenstein et al. (2024) train value extraction models to classify both the presence and polarity of Schwartz values in reddit posts. They identify patterns like a strong negative stance towards conformity in subreddits such as r/Vegan and r/AbolishTheMonarchy an overall correlation between tradition values and the conservativeness of U.S. states. Shahid, Zhang, and Vashistha (2025) use LLMs to re-write constructive comments on homophobic and Islamophobic threads, then uses human annotations to evaluate the values represented in these comments. They find that the comments rewritten by LLMs exhibited decreased Conservative values with increased prosocial values such as Benevolence and Universalism.

This body of work uniformly makes the assumption that there exists an unambiguous "ground truth" value label, neglecting the inherent subjectivity of the task. In the next subsection, we discuss work that explores an alternative, perspectivist stance.

Perspectivism

The work highlighted above on automatically detecting Schwartz values offers an exciting lens on the values represented in text. However, in every case, these models assume a fixed meaning for both the content and the values, and a corresponding unambiguous "objective" value label in each case. This is in line with the longstanding paradigm in data labeling (Fleisig et al. 2024) which seeks to collect "ground truth" labels (Nowak and Rüger 2010; Snow et al. 2008) by aggregating labels from annotators. An alternative, the *perspectivist* paradigm (Basile et al. 2021; Fleisig et al. 2024; Plank 2022; McDonald, Schoenebeck, and Forte 2019; Frenda et al. 2025) argues that variation in annotator labels is not purely a source of error to be expunged, but rather is intrinsically meaningful. In this vein, Haghighi et al. (2025) argues that the ontological dimensions underlying generative systems are under-recognized, and makes the case for perspectives that are plural (instead of universal), grounded (instead of abstract), lively (instead of fixed) and enacted (instead of diluted).

We argue that our task of identifying the expression of basic human values in social media posts fits this framework well, because the labels have no singularly correct "ground truth" label that everyone agrees with (they are plural) and are personally calibrated to each individual's behavior (they are grounded).

Personalization

An alternative but complementary lens is personalization, which seeks to build models responsive to heterogeneous

preferences. In the context of recommender systems, classic approaches involve matrix factorization (Koren, Bell, and Volinsky 2009) and collaborative filtering (Resnick et al. 1994). For social media algorithms, Seth and Zhang (2008) use a Bayesian user-model to learn personal recommender systems from social network data. Lerman (2006) uses social filtering of activities as a way to personalize recommendations. See Eg, Tønnesen, and Tennfjord (2023) for a review on social media personalization.

In the context of aligning LLMs, past work has explored aligning models to users' opinions (Hwang, Majumder, and Tandon 2023; Do et al. 2025; Suh et al. 2025) and behaviors (Shaikh et al. 2025a). This is often task specific, with advances in domains as varied as computer usage (Shaikh et al. 2025b), creative writing (Chung et al. 2025) and user interface generation (Wu et al. 2025). In addition, recent work has explored personalized reward models (Chen et al. 2024; Jang et al. 2023; Li et al. 2024; Rame et al. 2023). This is achieved by conditioning of text-based user profiles (Zhang 2024; Ryan et al. 2025), using learned auxiliary predictors to steer towards group preferences (Zhao, Dang, and Grover 2023), and latent variable approaches (Poddar et al. 2024). In this work, we adopt a cold start approach (e.g. identifying and using a small yet compact set of past behaviors) because of its flexibility and interpretability.

Methods

In this section, we first report our methods for curating a dataset of social media posts. Next, we discuss our methods for annotating this dataset for values. Finally, we provide details our approach to modeling and evaluating human and AI-based value expressions.

Curating a Dataset of Social Media Posts

Recruitment We recruited a sample of social media users (N=281) quota-matched to be a representative sample of the US population on age, ethnicity, gender and partisanship using Prolific in May 2024. Participants were all from the United States and were required to be 18 years old or older. We were unable to meet the quota for 17/109 participants in the age range 55-100 so we upsample posts from participants in this age range by 18.4% when constructing our silverset of social media posts. Within this sample, the median age was 43 (min=19, max=80), 58.36% White/Caucasian, 51.2% female and 30.6% Democrat, 28.8% Republican and 40.6% Independent. Participants were paid $\sim$ $\$15.42$ an hour and took approximately 7 minutes to complete the study.

Feed Collection After they consented to the research, participants were asked to install a Chrome Extension that downloaded posts from their X (Twitter) feeds, both their algorithmically curated For You page (FYP) and their in-network Following feeds simultaneously (Piccardi et al. 2024a). We collect and annotate posts from both the Following feed and FYP to account for any systematic differences that may exist between them, and to pave the way for work that uses both to measure algorithmic amplification (see Discussion).

When a browser opens the X page, tweets are fed to the browser client in "batches": a single API request returns a set of tweets at a time. The Chrome Extension aims to download 16 batches of tweets: 7 from the FYP, and 9 from the Following page. Each batch for FYP consists of approximately 30 tweets, whereas the Following feed returns a variable number of tweets up to 100.

From the 281 participants, we collected 265,442 posts, with an average of 207 posts from the FYP and 776 from the Following page per user. Among these 265,442 posts, there are 151,740 posts with unique, non-empty text in English (using Google's Chromium language detection tool `cld2`). We also screen out private posts (n=6,025; 3.97%), deleted posts (n=2,004; 1.32%) and/or reply tweets that are replying to posts not displayed in the user's feed (n=7,581; 4.99%) to focus only on the 142,652 public posts we can render with proper context.

Subsampling feed posts to create an annotation dataset We next prepare these posts for human annotation, since some posts are either not safe for work (NSFW) to show to annotators or incomprehensible to most annotators. We filter the posts for comprehensibility (grounded in the constructs of readability, coherence, spam behavior, and context required for understanding) and NSFW (to comply with our institution's IRB for data annotation) using an automated GPT-based pipeline, (see Table A5 for full text and the Appendix for an evaluation of this model) with `GPT-4o`, temperature=1.0 and seed=0. We filter out any posts that score less than 3 on a comprehensibility scale, resulting in a dataset of 89,383 (62.6%) posts. We then filter out 1,937 posts with NSFW content, resulting in a final dataset of 87,446 posts.

The occurrence rate of different values in social media posts varies, as do the numbers of posts per person across both Following and FYP feeds. Therefore, uniform sampling procedure may yield a sample that underrepresents rarer values and overrepresents certain users for annotation. To construct a final set of posts for human annotation, we stratify on user and post source (FYP vs following) as well as on a preliminary LLM-based screening measure of value expression, designed to err on the side of inclusion and capture as many plausibly value-expressing posts as possible.

In particular, we perform a preliminary rough classification of these posts for all 19 of the Basic Human Values (Schwartz et al. 2012) on a 7-point Likert scale [0-6] using another automated GPT-based pipeline with the prompt that rates all 19 values jointly each on a 0-6 scale based on a short description, with the prompt defined in Table A6 with `GPT-4o`, temperature=1.0 and seed=0 (see Appendix for an evaluation of this method). This allows us to upsample posts that may reflect rarer values. For this upsampling, we say a given post reflects a given value if the GPT annotation pipeline scored it a 4 or above. We then sample a set of 5,227 posts for human annotation by stratifying across users, post source, and value expression. In particular, for each user and each value, we sample a tweet from that user's feed that reflects that value. We sample FYP and Following posts with equal probability, but for a given user, if only one type of

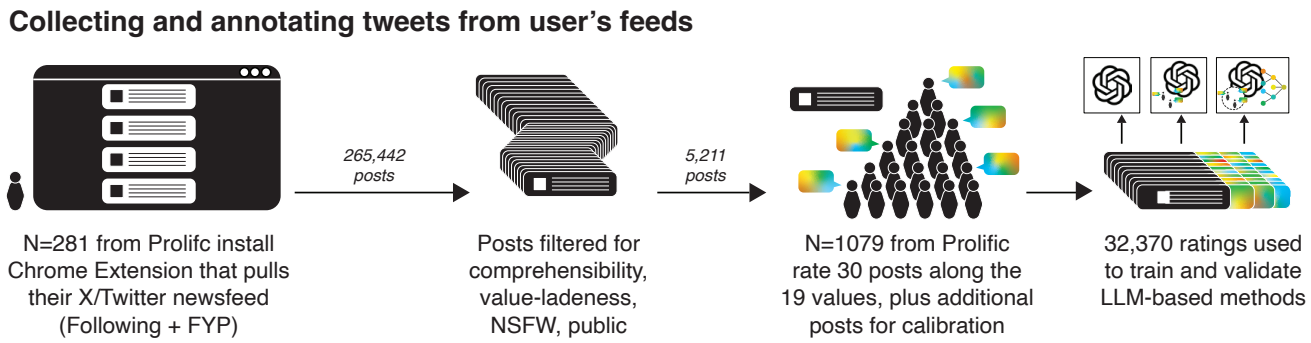

Figure 3: Method for collecting and annotating tweets from user's feeds

post exists, we return that tweet. Because we did not meet our quota for participants over the age of 55 (by 17 participants), we randomly sample (without replacement) 17 of the 76 participants over the age of 55 who completed the survey and repeat this process for each of them to ensure representation of posts from users over the age of 55. Of the 5,227, 3,182 came from users' Following feeds (60.8%) and 2,045 (39.1%) came from users' FYP. This process yielded a dataset of 5,227 posts that we can use to study human value annotation and to predict the values expressed in those posts.

Annotating Social Media Posts for Schwartz Values

In this section, we describe our ground truth data annotation method.

Participants. In total, 1,276 participants began the survey. We filtered out N=153 participants who failed two simple attention checks (see Appendix) or who did not finish the study. N=1,123 completed the study. Our final dataset included ratings from N=1,079 participants who have a full set of ratings for all 30 posts assigned to them and within this sample, the median age was 45 (min=18, max=85), 61.2% White/Caucasian, 50.8% female and 29.6% of participants lean Democrat vs 28.63% lean Republican. Participants were paid $15.00 an hour and took approximately 60 minutes to complete the study.

Procedure Each participant provided value expressions ratings for 30 posts. The posts for each participant were randomly sampled from the dataset of 5,227 posts sourced as described above (each post was rated by a median of 6 people). Given this sampling procedure, we retained ratings of 5,211 posts (since posts are sampled with replacement for each participant).

We used a recursive labeling where raters could traverse the value tree as defined in Figure A3 (top) devised to elicit value ratings for all 19 values in an efficient manner. The branching method leverages the tree structure of the Schwartz Theory of Basic Values, starting with the high-level values, and traversing down only those branches whose parent values they had marked as expressed in the post.

Following this primary task, participants completed a custom developed Value Calibration Questionnaire (VCQ) of 25 additional ratings on a shared set of questions (see Table A3 and the Personalization framework in the Methods sections for a description of how these questions were sourced). Finally, participants answered a number of demographic and political questions.

To familiarize them with the task and to determine eligibility for proceeding to the main labeling task, participants first completed a short training for four synthetic posts followed by a gating task. The training introduced annotators to the meaning of the Schwartz values, while the gating task assessed their understanding. These training posts were created to unambiguously express particular values, and participants could not proceed until they checked the right answer: if they answered incorrectly, a message appeared providing the correct answer and asking them to try again.

After training, they proceeded to the gating phase, where they rated four posts that, based on a pre-study, either unambiguously contained or did not contain the values of *Self-directed action* or *Face*. Participants who got one or fewer of the four correct were not allowed to continue. We took care in selecting the questions for both pre-tasks and in setting the eligibility threshold for the gating task to ensure a shared baseline understanding. See Figure AA3 Bottom for screenshots for the annotation of one of the gating posts.

After completing basic attention screeners and training, participants who successfully passed the gating phase proceeded to rate 30 posts drawn at random from the larger set of 5,227 posts for value expression using the branching method discussed above. When presenting a tweet to our annotators for value annotation, to convey context, the text of a given tweet (tweet_1) made in reply to or quoting another tweet (tweet_2) was represented as "tweet_2 REPLY TO: tweet_1" and "tweet_2 QUOTED: tweet_1" respectively. The annotation dataset we create is available for researchers on a case by case basis to preserve the privacy of the participants. It is stored securely in a de-identified manner.

Modeling Basic Human Values

In all instances, we compare the predicted value ratings from a range of algorithmic models to ratings from individuals or groups of annotators on a fixed holdout test set of 1496 (28.2%) of posts, which allows us to evaluate the finetuned and personalized models for unseen posts. The remaining 3715 posts are used for training models. We note that we

hold out at the *post*-level, letting train on a subset of the ratings for a given participant and test on their ratings from a different subset of posts.

To improve performance beyond zero-shot GPT-4o, we perform finetuning on high-consensus value labels. To produce this training set, we randomly select 1000 posts for which we have more than 6 ratings per post from our dataset of 5,211 posts annotated by 1,079 participants across the 19 values. Then, for each post, we compute the Spearman rank correlation in value ratings for each pair of raters and average these correlation scores into an aggregate consensus score for each post. We then use the 600 highest consensus annotations of these posts (e.g., average value score rounded to the nearest integer, 11.5% of posts) as labels to fine-tune GPT-4o via the OpenAI API, producing a model that estimates the consensus label of the annotators. The fine-tuned model reached a final error of 0.0722 after 600 steps.

Personalization framework

To capture the subjective nature of value expression annotation, we combine the consensus fine-tuned model with a personalized calibration model to predict the values that a given *individual* would perceive as expressed in a given post. We do so by fitting a series of models (one for each value) trained on the ratings of 3000 posts (with 1496 heldout for evaluation) stratified by number of ratings to predict a given individual's value annotation of a given post using (a) the consensus predictions from our fine-tuned model as well as (b) that individual's responses to the calibration questions designed to explain variation in rater disagreement (see Figure 4). First, we describe how the calibration questions were developed, then we discuss the modeling procedure.

To identify features for modeling that are highly predictive of interpersonal differences, we identify post-value pairs that explain maximal variation in value expression annotation disagreements in a pre-study. In this pre-study, N=51 participants rated the same set of 30 posts on all 19 values using the same procedure as the primary annotation study. But unlike that process, which randomly sampled posts for annotation by each participant, here every participant annotated all 30 posts, resulting in a dense tensor (participant$\times$value$\times$post). We then transform the data into columns for each participant (N=51) and rows for each post/value pair ($19 \times 30 = 570$).

We de-mean this matrix row-wise and conduct principal component analysis (PCA) to compute bases of maximal variation ("eigenraters"). The first 25 eigenraters explain 86% of variation in disagreements across raters. For each of these eigenraters, we find the extremal post/value pair with the highest absolute value and add it to our set of questions, which we dub the Value Calibration Questionnaire (VCQ) to use for the main annotations study to solve the cold start problem. Each of the 25 questions presents a post and asked the participants to rate the extent to which a specific value was reflected in the post, phrased as "To what extent does this post reflect {value}?" The full set is shown in Table A3.

Given this data-driven approach to decomposing variance in value disagreements, post/value pairs are selected for inclusion because they are extremal within an eigenrater, not because they exhibit coverage of the items or values. We therefore note the appearance of both duplicate posts that sparked disagreement across different values (e.g., indices 3/16/22, 5/17, 7/18, 10/11, 2/25, and 13/23 in Table A3) and corresponding duplicate values that sparked disagreement across different posts (*Caring*, *Tradition*, *Equality*, *Achievement*, and *Independent thinking*). This also means certain values do not appear in the VCQ such as *Power*, *Societal security*, *Respect*, *Humility*, *Responsibility*, and *Connection to nature*, which is one limitation of this approach (see Discussion).

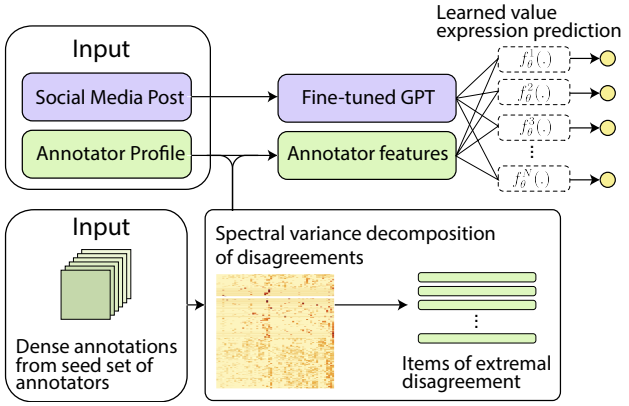

Figure 4: Our personalization framework uses a bespoke set of annotator features to calibrate predictions from the fine-tuned LLM.

Our personalized framework does not directly model individuals—it instead uses the 25 calibration questions as individual-level features to solve the cold start problem (see Figure 4). For each post–user pair, these models predict that user's 0–6 rating for a given value V using a random forest (e.g., one random forest model for each of the 19 values). Each model was trained in R using the `randomForest` package and default hyperparameters (500 trees). Figure A5 shows the total decrease in node impurities (variable importance) of each of the 19 random forest models, revealing how for a given value v the random forest predicting v primarily uses post-level the fine-tuned GPT prediction for value v (shown on the top diagonal).

Evaluation Approach

In settings where there is high levels of subjective variation exists in the labels, a traditional content-level ground truth set is not an appropriate measure and instead evaluations should rely on measure agreement with disaggregated individual labels (Gordon et al. 2022, 2021). For our evaluation metric, we chose Spearman rank correlation for several reasons. For one, a correlation metric allows us to compare the *relative* expression across all 19 low-level Schwartz values, which follows the relational nature of the value system itself. Further, a correlation computation across values is inherently scale-invariant, allowing us to account for individual level differences in how the scale is used. We also chose

Spearman rank correlation because the rank ordering captures the ordinal prioritization and relational nature of values consistent with the circumplex theory of the Schwartz theory of Basic Human Values. However, future work might explore extensions that explicitly use the circular geometry of the circumplex directly (see Discussion). We note that our results are qualitatively similar when using Pearson correlation instead of Spearman. In addition, a correlation metric handles the ordinal nature of the value annotations, which exist on a 7-point Likert scale.

We note that this choice of Spearman rank correlation is an extremely *conservative* measure, providing a rigorous lower bound on the actual levels of agreement. This is because the structure of the rank correlation, computed between two 19-dimensional values taking on values [0...6] with much density on 0, is quite sensitive to small adjustments in values, as incrementing and decrementing one value can cause it to skip up or down many other values in the rank order. Furthermore, in cases of posts with very little value expression (e.g. a vector of 18 zeros and 1 one), calculating Spearman rank within another very similar vector (e.g. a vector of 18 zeros and 1 one in another position) will result in a rank correlation of -0.05. And in the limit, calculating Spearman rank between two vectors of all zeros results in NA (because there is no variance) and is excluded from analysis, even though these cases represent perfect agreement. These reasons indicate that this extremely *conservative* measure, should not be considered in absolute numeric terms of model quality, but rather relative terms to establish that people agree with our LLM-based value annotation model more than they agree with a random other person, or consensus of groups of increasing size.

For a more interpretable evaluation metric, we also report results on MAE over values. In particular, we compute the consensus label of a holdout set of 1,496 posts by taking the mean rating rounded to the nearest integer for each value. Then for each post, we report the mean absolute error (MAE) between the LLM-generated label and the consensus label as LLM-Consensus MAE for each value. We then compare this error to the mean average distance that an individual human annotator has with the overall consensus label (Human-Consensus MAE). To be more precise, for each of the k annotators who labeled a value $\mathcal{V}$ on a given post, we compute the MAE between the label they assigned to the value and the mean labels of the remaining $k-1$ annotators as follows: $\frac{1}{k}\sum_{i=0}^{k}|\mathcal{V}_i - \frac{1}{k-1}\sum_{j=0,j\neq i}^{n}\mathcal{V}_j|$.

Results

To what extent do people agree?

We observe low levels of agreement in value annotations among raters, with the average Spearman rank correlation across posts of 0.201 (Figure 5, white bar). When comparing individual annotations to the consensus annotation for that post (i.e., the average of all other participants' ratings, rounded to the nearest integer), we observe an overall average Spearman rank correlation of 0.260 and an overall positive relationship with the number of raters (p=0.037), indicating a 29% relative improvement in the correlation in

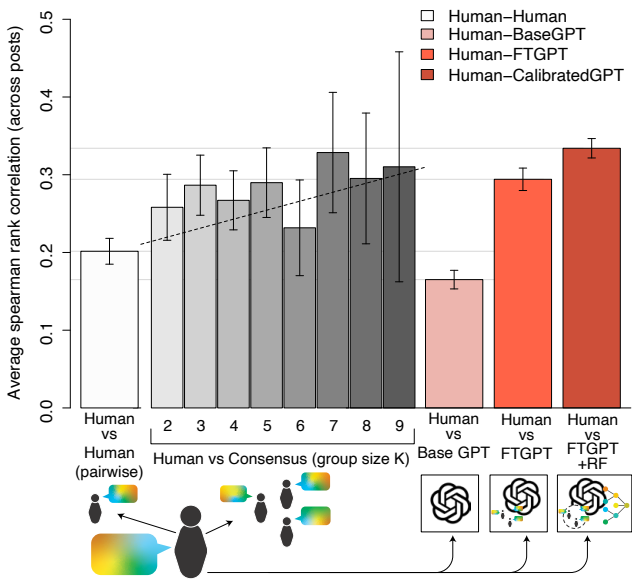

Figure 5: Average Spearman rank correlation (across posts) comparing individual value annotations to 1) other individuals (gray, far left), 2) groups of increasing group size (light gray to dark gray, middle), 3) zero-shot GPT-4o (light red, right), 4) fine-tuned GPT-4o (red, right), and 5) personalized random forest (dark red, right).

human-to-consensus labels compared to human-to-human labels. This suggests that despite the low levels of agreement, some levels of consensus are achieved when aggregating across individuals, particularly more than 3 or 4 people.

We also find subjectivity in value ratings, which helps explain the sources of disagreements we identify above. To test this claim we predict individual's value annotations across the 19 values by regressing, for each value, that particular value against each of the individual values that that individual holds, as well as demographics such as age and partisanship. The resulting regression coefficients are shown in the bottom of Figure A4. We find that for 8/19 of the values, the extent to which an individual holds that value is predictive of the likelihood that they would annotate a given post with that value (*Dominance*, *Resources*, *Achievement*, *Self-directed thoughts*, *Self-directed actions*, *Stimulation*, *Personal security*, and *Societal security*), which suggests people may project their own values into the perception of those values onto novel situations. We observe some partisan differences, where Republicans are more likely to rate a given post as reflecting the values of *Hedonism*, *Self-directed thoughts*, *Self-directed actions*, *Stimulation*, *Humility*, and *Nature*, which may map onto cultural-political divides. Within our sample, we do not detect age effects, except for the value of *Interpersonal conformity*.

Predicting Basic Human Values with LLMs

Next, we compare rates of agreement among people with those between people and LLM-based methods. First, we find that people agree with the zero-shot GPT-4o with an av-

erage Spearman rank correlation of 0.165 (light red in Figure 5), which is a 18% relative decrease in the correlation with which they agree with the annotations from a random other person ($\bar{\rho} = 0.201$, white bar in Figure 5, $p < 0.001$).

By fine-tuning GPT-4o on the consensus annotations from the highest consensus posts, we construct an AI model that generates value annotations with which people agree more than they agree with a random other person ($p < 0.001$), with an average Spearman rank correlation of 0.294. This improvement reflects a 46% relative improvement in correlation compared to the human-to-human correlation. Comparing the fine-tuned GPT-4o to these wisdom of the crowds estimates, we see a rate of agreement similar to taking the average (consensus) rating of 8 or more other people (medium red in Figure 5).

When aggregating across values, we find the average mean absolute error (MAE) between the fine-tuned GPT-4o and human consensus (crowd) is 0.763 ± 0.037 compared to an average MAE between human consensus and individual raters of 1.113 ± 0.0743 (see Table 1). This result suggests that the LLM is predicting the consensus better, since its distance from the consensus is less than from a random individual annotator.

Value	Human-Crowd MAE	LLM-Crowd MAE
Humility	1.620 ± 0.035	0.971 ± 0.024
Caring	1.622 ± 0.039	1.027 ± 0.025
Dependability	1.512 ± 0.037	0.917 ± 0.025
Universal concern	1.389 ± 0.039	0.865 ± 0.025
Hedonism	1.325 ± 0.037	0.927 ± 0.026
Face	1.208 ± 0.037	0.918 ± 0.024
Tolerance	1.208 ± 0.036	0.740 ± 0.023
Societal security	1.206 ± 0.036	0.779 ± 0.024
Rule conformity	1.175 ± 0.037	0.737 ± 0.024
Tradition	1.169 ± 0.035	0.777 ± 0.024
Self-dir. actions	1.118 ± 0.038	0.822 ± 0.025
Self-dir. thoughts	1.104 ± 0.038	0.800 ± 0.024
Interp. conformity	1.048 ± 0.035	0.699 ± 0.022
Achievement	1.005 ± 0.037	0.749 ± 0.022
Stimulation	0.959 ± 0.035	0.729 ± 0.024
Preservation of nature	0.738 ± 0.031	0.497 ± 0.022
Resources	0.694 ± 0.030	0.599 ± 0.020
Dominance	0.650 ± 0.029	0.566 ± 0.018
Personal security	0.414 ± 0.026	0.395 ± 0.017
Overall	1.113 ± 0.0743	0.763 ± 0.037

Table 1: Mean Absolute Error (MAE) between human value expression annotations and consensus (crowd) aggregations (center) and MAE between the fine-tuned gpt-4o model and consensus (crowd) aggregations (center) on a holdout set of posts (N=1,496)

Personalization Results

While fine-tuning an LLM on high-consensus labels yields a model that people agree with more than they agree with other people, the inherent subjectivity of the task underscores the point that any single model will not be performant for everyone. Therefore, next we train a *perspectivist* model that attempts to model an individual's subjective perception of value expression. In particular, the personalized calibration model uses a series of random forest models to predict the values that a given individual would see reflected in a given post. Using only an individual's calibration ratings and the target post's aggregate GPT-predicted values as inputs, the random forests achieves an average Spearman rank correlation of 0.334 with human ratings, a rate of agreement surpassing the wisdom of the crowds ($p < 0.001$, dark red in Figure 4). This most performant model reflects a 66% relative improvement in correlation vs. the human-to-human baseline. This model also reflects a 28% relative improvement in correlation vs. the more difficult human-to-consensus condition. We find that for each value, the random forest models typically use the underling GPT consensus prediction of that value as the most important variable (see Figure A5) but also make use of the calibration questions from Table A3. These results (Table 2) support the claim that people agree with the LLM-based method at a level at or above the extent to which they agree with other people, for both other individuals as well as wisdom of the crowds.

Condition	ρ	**%Δ vs. H–H**
Human vs. Base GPT	0.165	-18%
Human vs. Human	0.201	–
Human vs. Consensus	0.260	+29%
Human vs. Fine-tuned GPT	0.294	+46%
Human vs. Personalized GPT	0.334	+66%

Table 2: Correlation results and relative improvement vs. human–human baseline.

What Drives Personalization Improvements?

Adjusting differentially subjective values In Table 1, we see that the most overall subjective values in our sample were humility, caring, and dependability, while some of the most agreed upon values were resources, dominance, and personal security. To understand which values are being adjusted and how that effects personalization improvements, for a given value v, we compute the average absolute difference (Δ) between the predictions of the fine-tuned consensus and the personalized models for value v. We find that the top adjusting values are humility ($\Delta = 0.771$), caring ($\Delta = 0.655$) and dependability ($\Delta = 0.655$), and the least adjusting values are personal security ($\delta = 0.344$) and preservation of nature ($\Delta = 0.429$) with an overall value-wise spearman rank correlation of R=0.833. This indicates that personalization improves agreement by differentially adjusting subjective values where perceptions differ.

Personalized to whom? For each participant, we compute the average spearman rank correlation between that participant's ratings of a given post and the model predictions and then compute the difference in average spearman rank correlation between the consensus model and the personalized. In other words, we compute a participant-level measure of

the delta between the red bar in Figure 5 and the dark red bar in Figure 5. We then use a linear regression to compute the association between this delta (e.g. performance gain due to personalization) with individual differences. We find that personalization gains are relatively higher for men ($p = 0.0102$), Republicans ($p = 0.0297$) and no effect across ages ($p = 0.824$).

When comparing posts sourced from the For You page versus the Following page, we see no differences in either personalization gains ($p = 0.709$) or individuals average level of disagreement with the consensus annotation ($p = 0.8$).

Discussion

In this paper, we introduce and validate a framework for measuring the expression of basic human values on social media paradigmatic of a shift towards perspectivism. From a methodological perspective, we note that the large variation in value perceptions across individuals suggests that while indeed there is a degree of shared understanding, there is no singular ground truth in every case: the low levels of agreement in value annotations we observe appear even after training and gating to encourage shared understanding.

To acknowledge the inherent subjectivity of the task, we introduce a specialized framework for using LLMs for annotating subjective social science constructs. With this framework, we showcase a personalized LLM-based model that people agree with more than the base model, other people, or even consensus annotations. This approach represents a perspectivist approach (Fleisig et al. 2024) to modeling subjective constructs, which we believe is critical in the age of ontologically monolithic LLMs (Haghighi et al. 2025).

We note that in this paper, we were primarily focused on arguing for a paradigm shift towards personalization, and our personalization framework using random forests and a spectral decomposition of variations in disagreement is intended to demonstrate one possible instantiation that yields improvements beyond consensus annotations.

The personalization framework presented here therefore represents one exemplary approach as a proof-of-concept but future work is needed to evaluate this approach more deeply and contrast it to a more comprehensive design space. For example, as discussed above, the method of identifying highly predictive features to solve the cold start problem does not include certain values and thus requires future work to understand how items identified generalize from one sample of raters (and items) to another and develop alternative methods for solving the cold start problem efficiently.

In addition, we note that we only considered perspectivist models for the Schwartz value system. Future work should explore adapting this approach to other value systems that may be well-suited for the particular domain (such as the social media context). Depending on the domain, this might include other existing theories of values such as the World Values Survey or Moral Foundations Theory, or methods for deriving bottom-up theories of values via attentional probes (Klingefjord, Lowe, and Edelman 2024), or inverse reinforcement learning (Oliveira et al. 2023), or combinatorially combining "moral molecules" (Curry et al. 2022). In addition, we note that the metric we used, Spearman rank correlaiton, does not explicitley capture the circualar geometry of the Shwartz value system and future work could explore bespoke measures that capture angular distance more precisely.

Another limitation is that this study recruited Americans to annotate X posts. How this generalizes beyond the US content and to other social media platforms remains important future work. We also note we used two preliminary GPT-based filters (comprehensibility, NSFW) in addition to a GPT-based upsampling by predicted value to identify relevant posts that conformed to our IRB. Future work should investigate how value labeling extends to incomprehensible, value-sparse and NSFW posts.

There are several both epistemic and practical implications of this shift towards personalized values that are important to keep in mind. Epistemically, this underscores the need for further research into the subjectivity of a variety of social constructs (such as "helpfulness" or "toxicity"), particularly when they are measured and deployed in high stakes sociotechnical contexts. These constructs may operate differently for different people, and therefore these systems may inadvertently exhibit bias based on divergent interpretations.

One practical implication of this work is for social media feed design. If a platform can carefully understand how a user perceives values in content, it could tune that user's feed to highlight posts that resonate with their values. This opens the door to new feed algorithms that provide increased user control and alignment. In addition, this shift also provides a more precise way to audit what values are currently being amplified by existing feed ranking algorithms.

Ethical Considerations

Our approach imagines a world in which value expressions can be readily assessed and measured. This may open the door to new forms of persuasion and microtargeting based on value-based rhetoric. This could be mitigated by mandated disclosure for advertisers and opt-out mechanisms for consumers. It could also include labels for value-laden language, to help consumers identify and contextualize value aligned or value misaligned content.

This may also enable new forms of ranking algorithms that rank by values, which represents a departure from the engagement-based ranking paradigm of existing platforms. This possibility raises important questions for who gets to determine this information (platforms? users?) and how are these value profiles stored. We note that our method for personalizing value predictions and the corresponding VCQ do not rely on any privacy-compromising or sensitive individual attributes, and therefore representing a promising and procedurally legitimate approach to store value profiles.

Furthermore, there is the concern that the value-based ranking approach outlined above could induce value-based filter bubbles. This could be mitigated by identifying and design for *bridging values* that are held across partisan divides. However future work is needed to explore how what these

bridging values are and how effective they are in mitigating value-based filter bubbles.

We note that our approach relies on the OpenAI API which, if used in a real social media setting, could cause privacy concerns for users who have not explicitly consented to sharing their feeds' content or values with a third-party service. This could be mitigated by local LLM implementations that do not share data with a third-party, or by sufficiently explicit data-sharing agreements with users.

This work has been approved by the IRB of our institution to ensure adherence to ethical standards. In all the reproductions of social media posts from our dataset in this paper, we have reworded the post, blurred meta-data and replaced meta-data to preserve anonymity while maintaining value saliency.

Acknowledgements

Thanks to Jeanne Tsai and Jennifer Allen for helpful feedback on this paper. This project was supported by the Stanford Institute for Human-Centered Artificial Intelligence and National Science Foundation. Special thanks to the Digital Economy lab at Stanford University and Stochastic Labs.

References

Bail, C. 2022. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Basile, V.; Fell, M.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M.; Uma, A.; et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, 15–21. Association for Computational Linguistics.

Berger, P.; and Luckmann, T. 2016. The social construction of reality. In *Social theory re-wired*, 110–122. Routledge.

Bernstein, M.; Christin, A.; Hancock, J.; Hashimoto, T.; Jia, C.; Lam, M.; Meister, N.; Persily, N.; Piccardi, T.; Saveski, M.; et al. 2023. Embedding societal values into social media algorithms. *Journal of Online Trust and Safety*, 2(1).

Borenstein, N.; Arora, A.; Kaffee, L.-A.; and Augenstein, I. 2024. Investigating Human Values in Online Communities. *arXiv preprint arXiv:2402.14177*.

Boyd, R.; Wilson, S.; Pennebaker, J.; Kosinski, M.; Stillwell, D.; and Mihalcea, R. 2015. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 31–40.

Chakroff, A. 2015. *Discovering structure in the moral domain*. Ph.D. thesis.

Chen, R.; Zhang, X.; Luo, M.; Chai, W.; and Liu, Z. 2024. Pad: Personalized alignment of llms at decoding-time. *arXiv preprint arXiv:2410.04070*.

Chung, J. J. Y.; Padmakumar, V.; Roemmele, M.; Wang, Y.; Sun, Y.; Wang, T.; Almeda, S. G.; Halperin, B. A.; Lu, Y.; and Kreminski, M. 2025. LiteraryTaste: A Preference Dataset for Creative Writing Personalization. *arXiv preprint arXiv:2511.09310*.

Curry, O. S.; Alfano, M.; Brandt, M. J.; and Pelican, C. 2022. Moral molecules: Morality as a combinatorial system. *Review of Philosophy and Psychology*, 13(4): 1039–1058.

Davidov, E.; Schmidt, P.; and Schwartz, S. H. 2008. Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public opinion quarterly*, 72(3): 420–445.

Do, X. L.; Kawaguchi, K.; Kan, M.-Y.; and Chen, N. 2025. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2526–2547.

Eg, R.; Tønnesen, Ö. D.; and Tennfjord, M. K. 2023. A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, 9: 100253.

Epstein, Z.; Pennycook, G.; and Rand, D. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–11.

Fleisig, E.; Blodgett, S. L.; Klein, D.; and Talat, Z. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. *arXiv preprint arXiv:2405.05860*.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Frenda, S.; Abercrombie, G.; Basile, V.; Pedrani, A.; Panizzon, R.; Cignarella, A. T.; Marco, C.; and Bernardi, D. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2): 1719–1746.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

Gordon, M. L.; Zhou, K.; Patel, K.; Hashimoto, T.; and Bernstein, M. S. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 chi conference on human factors in computing systems*, 1–14.

Goren, P. 2020. Values and Public opinion. In *Oxford Research Encyclopedia of Politics*.

Haghighi, N.; Yu, S.; Landay, J.; and Rosner, D. 2025. Ontologies in Design: How Imagining a Tree Reveals Possibilites and Assumptions in Large Language Models. *arXiv preprint arXiv:2504.03029*.

Heilman, R. M.; and Kusev, P. 2020. Personal values associated with prosocial decisions. *Behavioral Sciences*, 10(4): 77.

Hwang, E.; Majumder, B. P.; and Tandon, N. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

Inglehart, R. 2020. *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton university press.

Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettlemoyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Jia, C.; Lam, M. S.; Mai, M. C.; Hancock, J. T.; and Bernstein, M. S. 2024. Embedding democratic values into social media AIs via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–36.

Kiesel, J.; Alshomary, M.; Handke, N.; Cai, X.; Wachsmuth, H.; and Stein, B. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4459–4471.

Klingefjord, O.; Lowe, R.; and Edelman, J. 2024. What are human values, and how do we align AI to them? *arXiv preprint arXiv:2404.10636*.

Kolluri, A.; Su, R.; Jahanbakhsh, F.; Zhao, D.; Piccardi, T.; and Bernstein, M. S. 2026. Alexandria: A Library of Pluralistic Values for Realtime Re-Ranking of Social Media Feeds. *ICWSM 2026*.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.

Krystallis, A.; Vassallo, M.; and Chryssohoidis, G. 2012. The usefulness of Schwartz's 'Values Theory'in understanding consumer behaviour towards differentiated products. *Journal of Marketing Management*, 28(11-12): 1438–1463.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Computational social science. *Science*, 323(5915): 721–723.

Lerman, K. 2006. Social networks and social information filtering on digg. *arXiv preprint cs/0612046*.

Li, X.; Zhou, R.; Lipton, Z. C.; and Leqi, L. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–23.

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K. 2020. Moral framing and ideological bias of news. In *International conference on social informatics*, 206–219. Springer.

Nowak, S.; and Rüger, S. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, 557–566.

Oliveira, N.; Li, J.; Khalvati, K.; Barragan, R. C.; Reinecke, K.; Meltzoff, A. N.; and Rao, R. P. 2023. Culturally-Attuned Moral Machines: Implicit Learning of Human Value Systems by AI through Inverse Reinforcement Learning. *arXiv preprint arXiv:2312.17479*.

Orlikowski, M.; Röttger, P.; Cimiano, P.; and Hovy, D. 2023. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*.

Piccardi, T.; Saveski, M.; Jia, C.; Hancock, J.; Tsai, J. L.; and Bernstein, M. S. 2024a. Reranking Social Media Feeds: A Practical Guide for Field Experiments. *arXiv preprint arXiv:2406.19571*.

Piccardi, T.; Saveski, M.; Jia, C.; Hancock, J. T.; Tsai, J. L.; and Bernstein, M. 2024b. Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity. *arXiv preprint arXiv:2411.14652*.

Plank, B. 2022. The'Problem'of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *arXiv preprint arXiv:2211.02570*.

Poddar, S.; Wan, Y.; Ivison, H.; Gupta, A.; and Jaques, N. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37: 52516–52544.

Ponizovskiy, V.; Ardag, M.; Grigoryan, L.; Boyd, R.; Dobewall, H.; and Holtz, P. 2020. Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text. *European Journal of Personality*, 34(5): 885–902.

Qiu, L.; Zhao, Y.; Li, J.; Lu, P.; Peng, B.; Gao, J.; and Zhu, S.-C. 2022. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11183–11191.

Rame, A.; Couairon, G.; Dancette, C.; Gaya, J.-B.; Shukor, M.; Soulier, L.; and Cord, M. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36: 71095–71134.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; and Riedl, J. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 175–186.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *AI magazine*, 36(4): 105–114.

Ryan, M. J.; Shaikh, O.; Bhagirath, A.; Frees, D.; Held, W. B.; and Yang, D. 2025. Synthesizeme! inducing persona-guided prompts for personalized reward models in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8045–8078.

Schultz, P. W.; and Zelezny, L. C. 1998. Values and proenvironmental behavior: A five-country survey. *Journal of cross-cultural psychology*, 29(4): 540–558.

Schwartz, S. H. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, 1–65. Elsevier.

Schwartz, S. H.; and Cieciuch, J. 2022. Measuring the refined theory of individual values in 49 cultural groups: psychometrics of the revised portrait value questionnaire. *Assessment*, 29(5): 1005–1019.

Schwartz, S. H.; Cieciuch, J.; Vecchione, M.; Davidov, E.; Fischer, R.; Beierlein, C.; Ramos, A.; Verkasalo, M.; Lönnqvist, J.-E.; Demirutku, K.; et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4): 663.

Seth, A.; and Zhang, J. 2008. A social network based approach to personalized recommendation of participatory media content. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 2, 109–117.

Shahid, F.; Zhang, S.; and Vashistha, A. 2025. LLMs Homogenize Values in Constructive Arguments on Value-Laden Topics. *arXiv preprint arXiv:2509.10637*.

Shaikh, O.; Lam, M. S.; Hejna, J.; Shao, Y.; Cho, H. J.; Bernstein, M. S.; and Yang, D. 2025a. Aligning Language Models with Demonstrated Feedback. In *The Thirteenth International Conference on Learning Representations*.

Shaikh, O.; Sapkota, S.; Rizvi, S.; Horvitz, E.; Park, J. S.; Yang, D.; and Bernstein, M. S. 2025b. Creating general user models from computer use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, 1–23.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 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*, 254–263.

Soden, R.; Toombs, A.; and Thomas, M. 2024. Evaluating interpretive research in HCI. *Interactions*, 31(1): 38–42.

Stathopoulou, A.; and Balabanis, G. 2019. The effect of cultural value orientation on consumers' perceptions of luxury value and proclivity for luxury consumption. *Journal of Business Research*, 102: 298–312.

Stray, J. 2020. Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, 3(4): 443–463.

Stray, J.; Halevy, A.; Assar, P.; Hadfield-Menell, D.; Boutilier, C.; Ashar, A.; Bakalar, C.; Beattie, L.; Ekstrand, M.; Leibowicz, C.; et al. 2024. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, 2(3): 1–57.

Suh, J.; Jahanparast, E.; Moon, S.; Kang, M.; and Chang, S. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*.

Van Der Meer, M.; Vossen, P.; Jonker, C. M.; and Murukannaiah, P. K. 2023. Do Differences in Values Influence Disagreements in Online Discussions? *arXiv preprint arXiv:2310.15757*.

Weinzierl, M. A.; and Harabagiu, S. M. 2022. From hesitancy framings to vaccine hesitancy profiles: A journey of stance, ontological commitments and moral foundations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1087–1097.

Wu, J.; Swearngin, A.; Vajjala, A. K.; Leung, A.; Nichols, J.; and Barik, T. 2025. Improving User Interface Generation Models from Designer Feedback. *arXiv preprint arXiv:2509.16779*.

Zhang, J. 2024. Guided profile generation improves personalization with llms. *arXiv preprint arXiv:2409.13093*.

Zhao, S.; Dang, J.; and Grover, A. 2023. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes because we do not do these things.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes.
   (b) Have you provided justifications for all theoretical results? Yes.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes.

(e) Did you address potential biases or limitations in your theoretical framework? Yes.

(f) Have you related your theoretical results to the existing literature in social science? Yes.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, prompts and details necessary to reproduce the results are provided in the Appendix. The data and code are provided at the GitHub repository link.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? N/A

(b) Did you mention the license of the assets? N/A

(c) Did you include any new assets in the supplemental material or as a URL? N/A

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes but in order to comply with our IRB and protect the privacy of our participants, the data will be shared via a Data Rights Agreement (DRA) to other researchers on a one by one basis.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, provided in GitHub repository.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, annotation interface and screenshots provided in the Appendix.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes.

(d) Did you discuss how data is stored, shared, and deidentified? Yes.

Appendix

Evaluating LLM's for data processing

NSFW Classifier

To evaluate our classifier for the comparatively straightforward, unambiguous and sparse task of identifying content that is not safe for work (NSFW), we compute agreement on a random sample of 100 positive and 100 negative examples tagged by the classifier. Two authors (A1, A2) annotated the 200 posts with a binary label for NSFW based on the codebook shown in Table A5. A1 and A2 agree with each other with Cohen's $\kappa = 0.65$. A1 agreed with the model with Cohen's $\kappa = 0.70$ corresponding to an accuracy of 84.9%. A2 agreed with the model with Cohen's $\kappa = 0.86$ corresponding to an accuracy of 93.0%.

Comprehensibility Classifier

Next, we evaluate our classifier for the relatively ambiguous, context-based and label balanced task of identifying content that is comprehensible and therefore capable of being judged for values. Comprehensibility is a highly subjective task because annotators are coming from different baselines, where what is comprehensible to them may differ based on their backgrounds, experiences and contexts. To evaluate this model, we also compute agreement on a random sample of 100 positive and 100 negative examples tagged by the classifier. Three authors (A1, A2, A3) annotated the 200 posts with a binary label for comprehensibility based on the codebook shown in Table A4. We then computed an overall majority label from the three rater's responses (121 labeled as comprehensible, 79 labeled as incomprehensible) and downsample the comprehensible posts to balance the labels with respect to the human-rated majority labels.

For this balanced set of annotations, the classifier has a recall of 0.695 and a precision of 0.721 (see Table A1). However many of the misclassifications occur for posts where raters disagreed on if the post was comprehensible: when

considering the subset of ratings where all three raters unanimously agreed on the label (85 labeled as comprehensible, 48 labeled as incomprehensible, similarly downsampling the comprehensible posts to balance the labels with respect to the ground truth), the classifier has a recall of 0.709 and a precision of 0.813.

Metric	Majority	Unanimous Majority
Accuracy	0.703	0.740
κ	0.405	0.479
Recall	0.695	0.709
Precision	0.722	0.813
F1 Score	0.708	0.757
Number of posts	158	96

Table A1: Evaluation metrics for comprehensibility classifier grounded in 1) majority annotations from the three raters, and 2) majority annotations where the raters unanimously agreed.

Preliminary Values Classifier

To evaluate the preliminary values classifier we used to upsample infrequent values into the final annotation dataset, we leverage the fact that our annotated dataset contains examples of positive and negative labels of the preliminary values classifier for each of the 19 values. In particular, for each value v, we compare the mean human rating for posts we sampled as positive examples (i.e. the classifier identified as containing the value $\hat{y}_v \geq 4$) to the other posts sampled for other values (i.e. $\hat{y}_v < 4$), the results of which are shown in Table A2. We see that for 18/19 of the values, the average rating of the posts sampled via the preliminary values classifier is statistically greater than the those that were not.

Value	Average ($\hat{y}_v \geq 4$)	Average ($\hat{y}_v < 4$)	p-val
Face	1.174	0.761	< 0.001
Dominance	0.472	0.371	0.001
Resources	0.844	0.390	< 0.001
Achievement	1.245	0.517	< 0.001
Hedonism	2.457	0.955	< 0.001
Self-Dir. Thoughts	1.423	0.719	< 0.001
Self-Dir.Actions	1.322	0.734	< 0.001
Stimulation	1.238	0.583	< 0.001
Personal Security	0.627	0.208	< 0.001
Societal Security	1.389	0.649	< 0.001
Tradition	1.677	0.683	< 0.001
Rule Conformity	1.989	0.620	< 0.001
Interp. Conformity	0.787	0.631	< 0.001
Humility	1.426	1.345	0.171
Dependability	1.865	1.235	< 0.001
Caring	2.551	1.235	< 0.001
Universal Concern	2.446	0.851	< 0.001
Pres. of Nature	3.076	0.465	< 0.001
Tolerance	1.803	0.754	< 0.001

Table A2: Schwartz Value Ratings (Human) by preliminary values classifier

Case studies of value disagreements in social media posts

What kind of disagreements did we notice in social media posts? While we did not observe any global linguistic or contextual trends associated disagreements, we did observe many disagreements where certain value dimensions of a post where salient to some participants, while others dimensions were salient to others. In addition to annotating the posts for values, our survey also provided participants an optional space to provide their rationale. These rationales help shed light on some disagreements and divergent perspectives within the posts.

Case Study 1:

Figure A1 depicts one post from our dataset, describing a magnetic storm hitting earth. On one hand, One participant read the post as a socially meaningful act—inferring care or universal concern from the decision to share timely information about a potentially impactful event, stating "great post to give others information about world shifting weather events." On the other hand, another participant grounds value inference in the described phenomenon itself, focusing on the natural event's aesthetic and reminder of humans' smallness in the face of nature, mapping it onto humility and appreciation of nature, stating "The northern lights are just beautiful and a good reminder of the beauty of earth."

Case Study 2:

Figure A2 depicts another post from our dataset, describing an NFL playing giving a shout out to their 4th grade teacher. One participant centers the relational and moral dimension of the story, interpreting the public acknowledgment of a teacher as an expression of dependability and caring toward those who supported the athlete's development, stating: "it's important to thank those who inspire us". The other focuses on the athlete's success itself, interpreting the post as a story about achievement and public recognition, stating: "It is a positive story about the guy who was drafted. It is about his success."

Attention checks

Our first attention check included an image of the digits 15 and a textbox asking "Please enter the number you see in the image (use numerical digits)" Our second attention check asked the participant: "Help us keep track of who is paying attention, please select - "Somewhat disagree" in the options below.")

Index	Post Text	Value
1	Pat Cummins decided to take a break from IPL 2023 due to hectic cricket season in International cricket then; - Won the WTC as captain. - Retained Ashes as captain. - Won the WC as captain. Captain, Leader, Legend, Cummins.	Achievement: success according to social standards
2	TOMORROW: I have an hourlong exclusive that we are broadcasting live for @BloombergTV with Citadel founder Ken Griffin, whose hedge fund recently surpassed Bridgewater as the world's most profitable. Send questions!! There's a lot on the agenda...	Achievement: success according to social standards
3	Being pro-palestine doesn't EVER mean that it gives you the right to be antisemitic. Jewish people have been our greatest allies, brothers, sisters since 7th october than anyone else on this fucking earth. if you're being antisemitic, youre not pro-palestine. we dont want you	Caring: devotion to the welfare of ingroup members
4	This morning, I proudly signed the Korean American VALOR Act into law – providing a pathway for thousands of Korean American Vietnam War veterans to access VA health care. Because of this law, they'll have the care and benefits they earned.	Caring: devotion to the welfare of ingroup members
5	Over 2000 years ago a child was born that came to die for all of our sins. He is the Christ, the living God, the Saviour of the world. Love Him with all your heart, because He loves you. Have a blessed Christmas everyone.	Caring: devotion to the welfare of ingroup members
6	R.I.P. My father, Marine Staff Sergeant Randolph Elder; my little brother Dennis Elder, Army Vietnam Vet; my older brother Kirk Randolph Elder, Navy Vietnam Era Vet; my nephew Eric Randolph Elder, Army Congressional Gold Medal: #VeteransDay2023	Caring: devotion to the welfare of ingroup members
7	The choice for GOP primary voters: Do we want "reform" or revolution? Do we aspire to "normalcy" or excellence? Do we want Super PAC puppets or patriots who speak the TRUTH? Clinton, IA .	Lawfulness: compliance with rules, laws, and formal obligations
8	After years, the People's Republic of China and the United States are restarting cooperation on counternarcotics. In particular, we seek to reduce the flow of precursor chemicals and pill presses fueling the fentanyl crisis.	Rule conformity: compliance with rules, laws, and formal obligations
9	A man has been arrested on suspicion of manslaughter following the death of ice hockey player Adam Johnson during a game in October, UK police say.	Personal security: safety in one's immediate environment
10	getting the nice plates out for the holidays	Novelty: excitement, novelty, and change
11	getting the nice plates out for the holidays	Hedonism: pleasure and sensuous gratification
12	Darius Jackson allegedly denies abusing Keke Palmer. A "source" tells TMZ they were arguing over custody and photos in restraining order were him and Palmer wrestling over her phone. Also included is a video of Palmer's mom threatening to put a bullet in his head :... Keke Palmer cradling her baby son Leodis as she steps out for the first time since being granted temporary sole custody and a restraining order against her allegedly 'abusive' ex Darius Jackson:	Reputation: security and power through maintaining one's public image and avoiding
13	Merry Christmas to ALL my followers on I am GUILTY of loving America too much I am GUILTY of being ULTRA MAGA I am GUILTY of loving all of you Am I on the Naughty List ?	Resources: power through control of material and social resources
14	"Every day is a gift." – Art Loveley	Self-directed actions: freedom to determine one's own actions
15	Couple breaks down when their cat returns after going missing	Stimulation: excitement, novelty, and change
16	Being pro-palestine doesn't EVER mean that it gives you the right to be antisemitic. Jewish people have been our greatest allies, brothers, sisters since 7th october than anyone else on this fucking earth. if you're being antisemitic, youre not pro-palestine. we dont want you	Tradition: maintaining and preserving cultural, family, or religious tradition
17	Over 2000 years ago a child was born that came to die for all of our sins. He is the Christ, the living God, the Saviour of the world. Love Him with all your heart, because He loves you. Have a blessed Christmas everyone.	Tradition: maintaining and preserving cultural, family, or religious tradition
18	The choice for GOP primary voters: Do we want "reform" or revolution? Do we aspire to "normalcy" or excellence? Do we want Super PAC puppets or patriots who speak the TRUTH? Clinton, IA .	Tradition: maintaining and preserving cultural, family, or religious tradition
19	UFC Champion Islam Makhachev just shared this pro-Palestine graphic!	Universal concern: commitment to equality, justice, and protection for all people
20	I'm proud to establish a new White House Initiative on Women's Health Research, an effort led by the First Lady and my Gender Policy Council. Together, they'll work to ensure our Administration does everything it can to drive innovation in women's health and close research gaps.	Equality: commitment to equality, justice, and protection for all people
21	Hey, remember when everyone endured like two straight years of trauma and then did nothing to address it and we all collectively and institutionally buried it way down deep under the constant productive pressures of capitalism? Probably fine	Equality: commitment to equality, justice, and protection for all people
22	Being pro-palestine doesn't EVER mean that it gives you the right to be antisemitic. Jewish people have been our greatest allies, brothers, sisters since 7th october than anyone else on this fucking earth. if you're being antisemitic, youre not pro-palestine. we dont want you	Equality: commitment to equality, justice, and protection for all people
23	Merry Christmas to ALL my followers on I am GUILTY of loving America too much I am GUILTY of being ULTRA MAGA I am GUILTY of loving all of you Am I on the Naughty List ?	Independent thinking: freedom to cultivate one's own ideas and abilities
24	It's ironic to me that people who deny Jesus still observe Christmas and Easter by taking those days off. Devout atheists should work right through the holidays if they are serious about their denial.	Independent thinking: freedom to cultivate one's own ideas and abilities
25	TOMORROW: I have an hourlong exclusive that we are broadcasting live for @BloombergTV with Citadel founder Ken Griffin, whose hedge fund recently surpassed Bridgewater as the world's most profitable. Send questions!! There's a lot on the agenda...	Wealth: control of material and social resources

Table A3: Value Calibration Questionnaire (VCQ) developed to personalize value predictions

```
Task -- I will provide a Twitter post alongside a codebook that describes in which settings Twitter posts are
comprehensible Tweets, and in which settings Tweets lack understandability. I want you to apply each concept in
the codebook to determine why and to what degree the concept the applies to the post. As you answer, please take
the following steps:
Step 1) For each concept in the codebook, describe whether and to what degree the Tweet illustrates
comprehensible, or uncomprehensible behavior, with the following format: "<CONCEPT 1>": "Why": "<explanation
of how the concept applies to the post>", "Rating": <rating>, ..., "<CONCEPT N>": "Why": "<explanation of how the
concept applies to the post>", "Rating": <rating>. ("Codebook Application")
Use the following scale to assign your rating:
0="the post strongly exhibits uncomprehensible behavior for the given concept"
1="the post somewhat exhibits uncomprehensible behavior for the given concept"
2="the post somewhat exhibits comprehensible behavior for the given concept"
3="the post strongly exhibits comprehensible behavior for the given concept"
Step 2) Summarizing your reasoning in Steps 1 and 2, determine a single rating for whether the post is
comprehensible, or uncomprehensible: "Final Rating": "Why": "<explanation of final rating>", "Rating": <rating>.
("Agreement Rating")
Use the following scale to assign your rating: 0="the post strongly exhibits uncomprehensible behavior" 1="the
post somewhat exhibits uncomprehensible behavior" 2="the post somewhat exhibits comprehensible behavior" 3="the
post strongly exhibits comprehensible behavior"
Codebook --
READABILITY: Readable -- Well-structured, coherent, and easily understandable language that facilitates the
comprehension of the expressed values;Unreadable -- Poorly structured, incoherent, or overly complex language
that hinders understanding of the message
COHERENCE: Coherent -- Logically structured, with a clear progression of ideas that underscore the expressed
meaning; Incoherent -- Disjointed or lacking logical flow, making it difficult to discern any underlying meaning
SPAM BEHAVIOR: Authentic -- Content is original, personal, and specifically targets relevant issues; Spammy --
Repetitive, unsolicited, or irrelevant content or appears automated
CONTEXT REQUIRED FOR UNDERSTANDING: SelfContained -- Minimal to no additional context required, as the post is
self-contained and clear; RequiresContext -- Requires external information or additional context for understanding
due to vagueness or lack of focus
Output Format -- Output your message in JSON, where you provide the following:  "Codebook Application":  "<CONCEPT
1>":  "Why": "<explanation of how the concept applies to the post>", "Rating": <integer from 0-3>, ..., "<CONCEPT
N>":  "Why": "<explanation of how the concept applies to the post>", "Rating": <integer from 0-3> , "Final
Rating":  "Why": "<explanation of final rating>", "Rating": <integer from 0-3>
Social Media Post --
### {post} ###
```

Table A4: Prompt for comprehension check

```
Task -- I will provide a Twitter post alongside a codebook that describes in which settings Twitter posts are
comprehensible Tweets, and in which settings Tweets lack understandability. I want you to apply each concept in
the codebook to determine why and to what degree the concept the applies to the post. As you answer, please take
the following steps:
Step 1) For each concept in the codebook, describe whether and to what degree the Tweet illustrates
comprehensible, or uncomprehensible behavior, with the following format: "<CONCEPT 1>": "Why": "<explanation
of how the concept applies to the post>", "Rating": <rating>, ..., "<CONCEPT N>": "Why": "<explanation of how the
concept applies to the post>", "Rating": <rating>. ("Codebook Application")
Use the following scale to assign your rating:
0="This post does not contain NSFW content at all"
1="This post contains a little bit of NSFW content"
2="This post moderately contains NSFW content"
3="This post strongly contains NSFW content"
Step 2) Summarizing your reasoning in Steps 1 and 2, determine a single rating for whether the post is NSFW:
"Final Rating": "Why": "<explanation of final rating>", "Rating": <rating>. ("Agreement Rating")
Use the following scale to assign your rating:
0="This post does not contain NSFW content at all"
1="This post contains a little bit of NSFW content"
2="This post moderately contains NSFW content"
3="This post strongly contains NSFW content"
Codebook --
SEXUAL: Sexual -- Contains graphic imagery or text that is sexually explicit; NotSexual -- Does not contain
sexually explicit imagery. Safe for the workplace.
VIOLENT: Violent -- Contains graphic imagery or text that describes violence in an explicit or egregieous way;
NonViolnet -- Contains no reference to violence
DEROGATORY: Derogatory -- Contains derogatory imagery or text that could make someone uncomfortable; NonDerogatory
-- Does not contain any derogatory imagery or text
Output Format -- Output your message in JSON, where you provide the following:
"Codebook Application":  "<CONCEPT 1>":  "Why": "<explanation of how the concept applies to the post>", "Rating":
<integer from 0-3>, ..., "<CONCEPT N>":  "Why": "<explanation of how the concept applies to the post>", "Rating":
<integer from 0-3> , "Final Rating":  "Why": "<explanation of final rating>", "Rating": <integer from 0-3>
Social Media Post --
### {post} ###
```

Table A5: Prompt for NSFW check

```
'Consider the following set of concepts, listed as !<CONCEPT>! : !<DEFINITION>!
- FACE_SCHWARTZ: Security and power through maintaining one's public image and avoiding humiliation; not wanting
to be shamed by anyone, protecting one's public image, wanting to always be treated with respect and dignity by
people
- DOMINANCE_SCHWARTZ: Power through exercising control over people; wanting people to do what one says, wanting to
be the most influential person in any group, wanting to be the one who tells others what to do
- RESOURCES_SCHWARTZ: Power through control of material and social resources; wanting to have the feeling of power
that money can bring, wanting to be wealthy, pursuit of high status and power
- ACHIEVEMENT_SCHWARTZ: Success according to social standards; being ambitious, wanting to be successful, wanting
people to admire oneś achievements
- HEDONISM_SCHWARTZ: Pleasure and sensuous gratification; having a good time, enjoying lifeś pleasures, taking
advantage of every opportunity to have fun
- SELF_DIRECTED_THOUGHTS_SCHWARTZ: Freedom to cultivate one's own ideas and abilities; being creative, forming
one's own opinions and having original ideas, learning things for oneself and improving one's abilities
- SELF_DIRECTED_ACTIONS_SCHWARTZ: Freedom to determine one's own actions; making oneś own decisions about one's
life, doing everything independently, freedom to choose what one does
- STIMULATION_SCHWARTZ: Excitement, novelty, and change; looking for different kinds of things to do, excitement
in life, wanting to have all sorts of new experiences
- PERSONAL_SECURITY_SCHWARTZ: Safety in one's immediate environment; avoiding anything that might endanger oneś
safety, personal security, wanting to live in secure surroundings
- SOCIETAL_SECURITY_SCHWARTZ: Safety and stability in the wider society; wanting oneś country to protect itself
against all threats, wanting the state to be strong so it can defend its citizens, wanting to have order and
stability in society
- TRADITION_SCHWARTZ: Maintaining and preserving cultural, family, or religious traditions; maintaining
traditional values's or beliefs, following one's family customs or the customs of a religion, valuing the
traditional practices of one's culture
- RULES_CONFORMITY_SCHWARTZ: Compliance with rules, laws, and formal obligations; wanting to do what people in
authority say, following rules even when no one is watching, obeying all the laws
- INTERPERSONAL_CONFORMITY_SCHWARTZ: Avoidance of upsetting or harming other people; avoiding upsetting other
people, not being annoying to anyone, trying to be tactful and avoid irritating people
- HUMILITY_SCHWARTZ: Recognizing one's insignificance in the larger scheme of things; not drawing attention to
oneself, being humble, being satisfied with what one has and not to ask for more
- DEPENDABILITY_SCHWARTZ: Being a reliable and trustworthy member of the ingroup; being loyal to those who are
close, going out of one's way to be a dependable and trustworthy friend, wanting those one spends time with to be
able to rely on one completely
- CARING_SCHWARTZ: Devotion to the welfare of ingroup members; helping people who are dear to oneself, caring for
the well-being of people one is close to, trying to always be responsive to the needs of family and friends
- UNIVERSAL_CONCERN_SCHWARTZ: Commitment to equality, justice, and protection for all people; protecting society's
weak and vulnerable members, wanting every person in the world to have equal opportunities in life, wanting
everyone to be treated justly even people one doesn't know
- PRESERVATION_OF_NATURE_SCHWARTZ: Preservation of the natural environment; believing that one should care for
nature, working against threats to the world of nature, protecting the natural environment from destruction or
pollution
- TOLERANCE_SCHWARTZ: Acceptance and understanding of those who are different from oneself; promoting harmony and
peace among diverse groups, listening to people who are different from oneself, wanting to understand people even
when disagreeing with them
For each concept, rate whether the message reflects the concept on a Likert scale with options [0, 1, 2, 3, 4, 5,
6], where: 0 = This post does not reflect this concept at all 1 = This post reflects this concept a little bit 6 =
This post reflects this concept strongly
A concept can be reflected if the message either supports or opposes the concept.
Output: One JSON dictionary of the following format. Rating is a dictionary of key value pairs, with each key
being a concept and each value your rating for that concept:
"Rating": \'Concept\': !<RATING>!
```

Table A6: Prompt for value expression

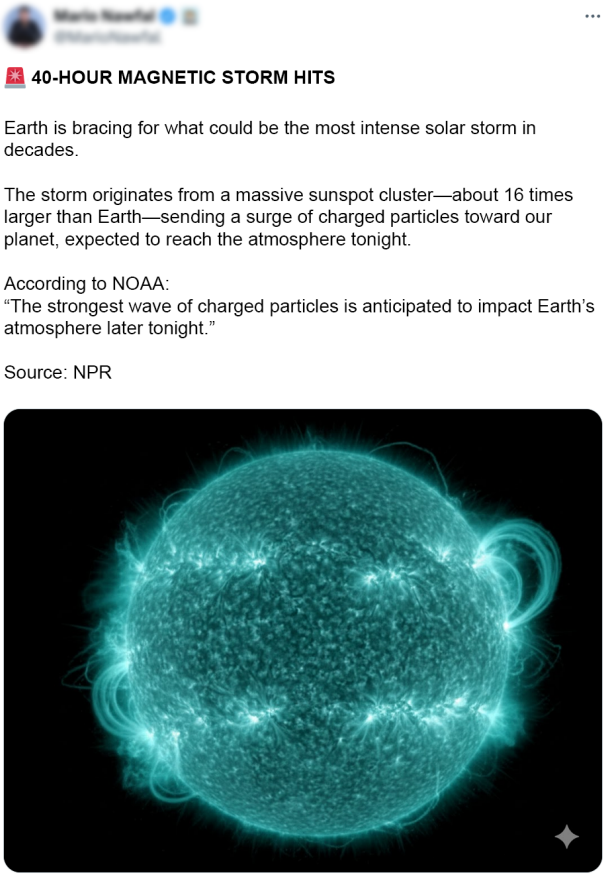

Figure A1: Anonymized example post from our dataset.

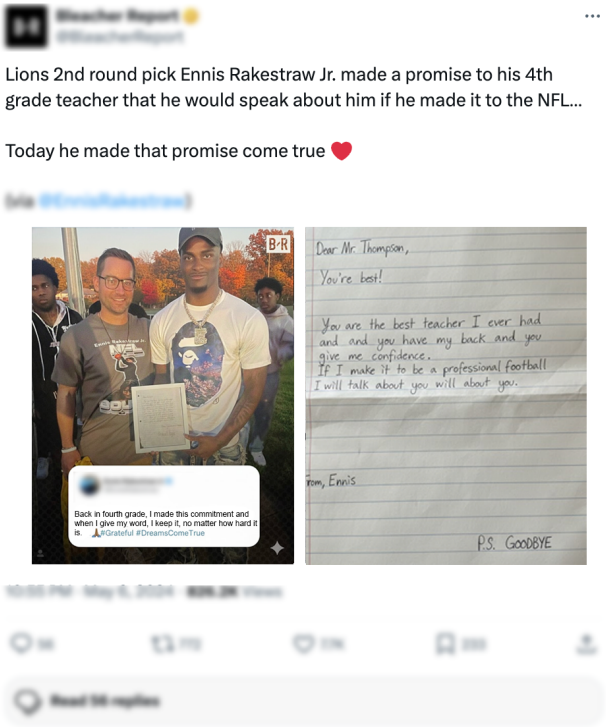

Figure A2: Anonymized example post from our dataset.

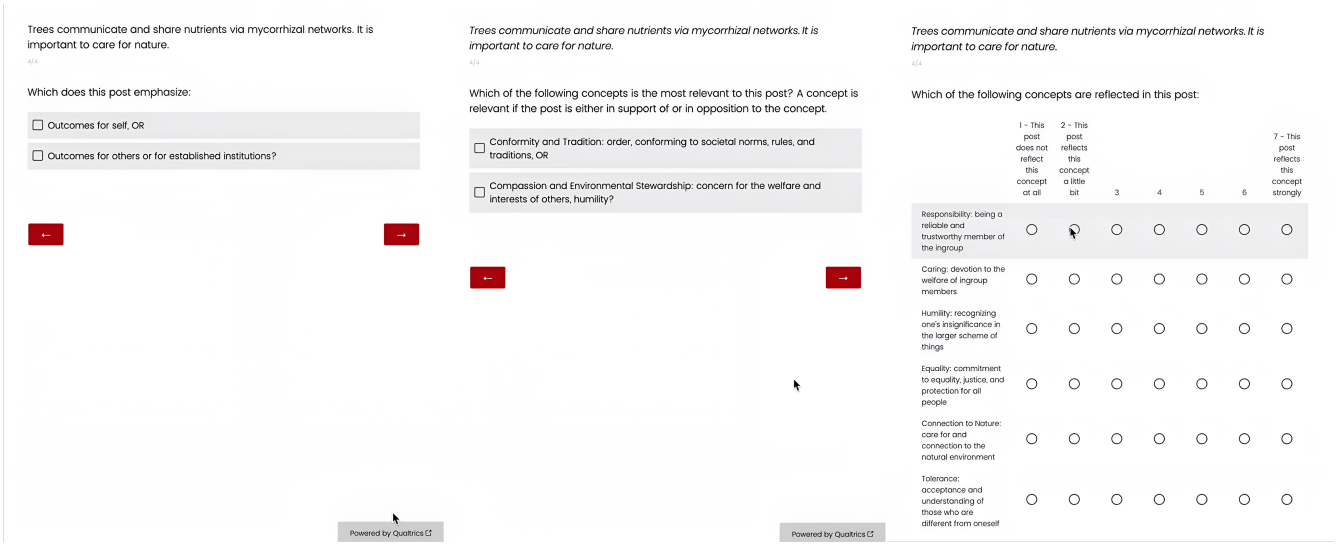

Figure A3: Top: Recursive tree structure for value annotation. Red indicates modifications we took to make the categories more interpretable for raters and salient to the social media context. Bottom: Example of recursive labeling scheme for gating post

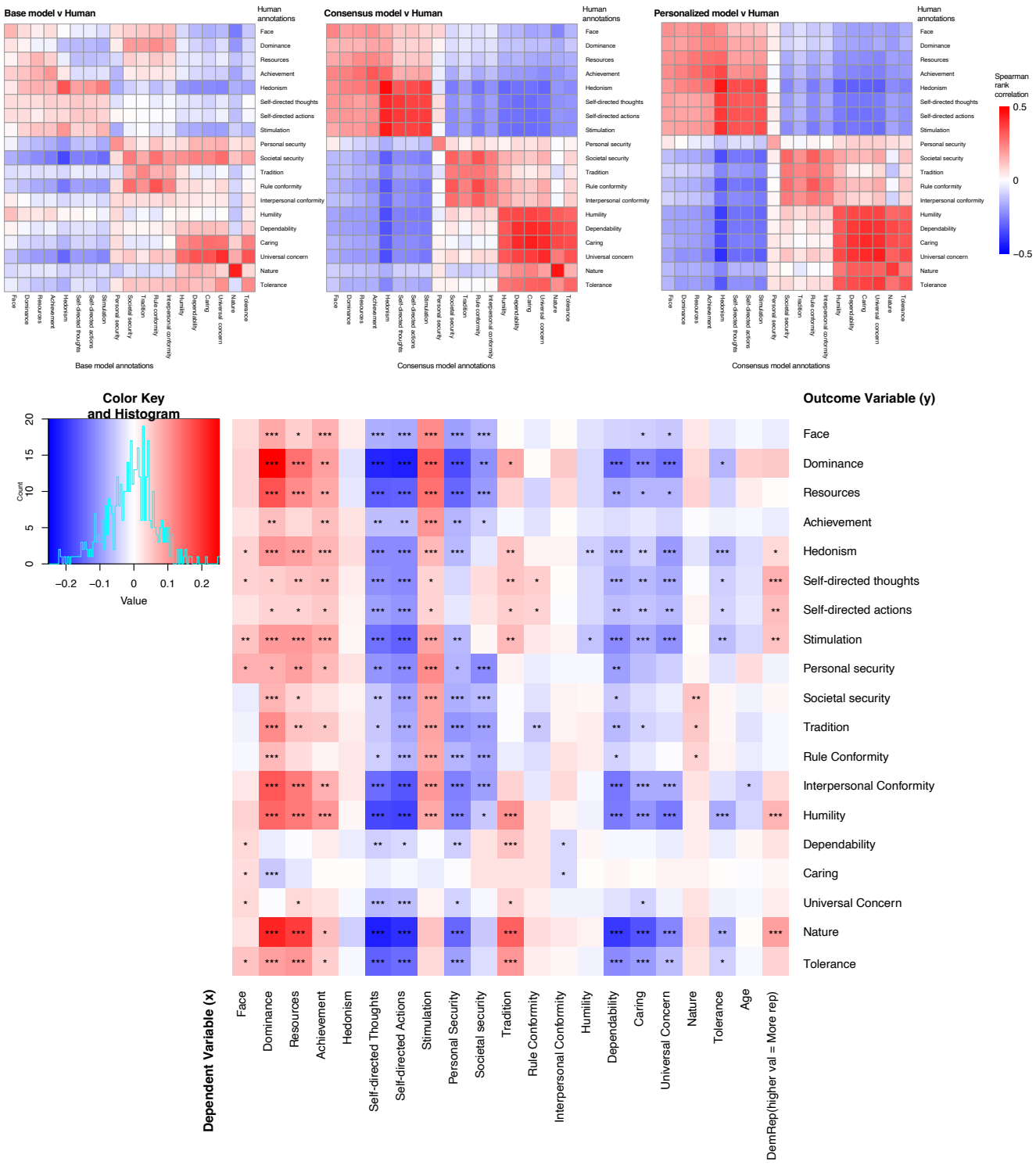

Figure A4: Top: Value-wise spearman rank correlation between raters and base model (left), consensus model (center) and personalized model (right). Bottom: Regression coefficients in predicting value annotations (y) from a given value that that individual holds (x), as well as demographics such as age and partisanship.

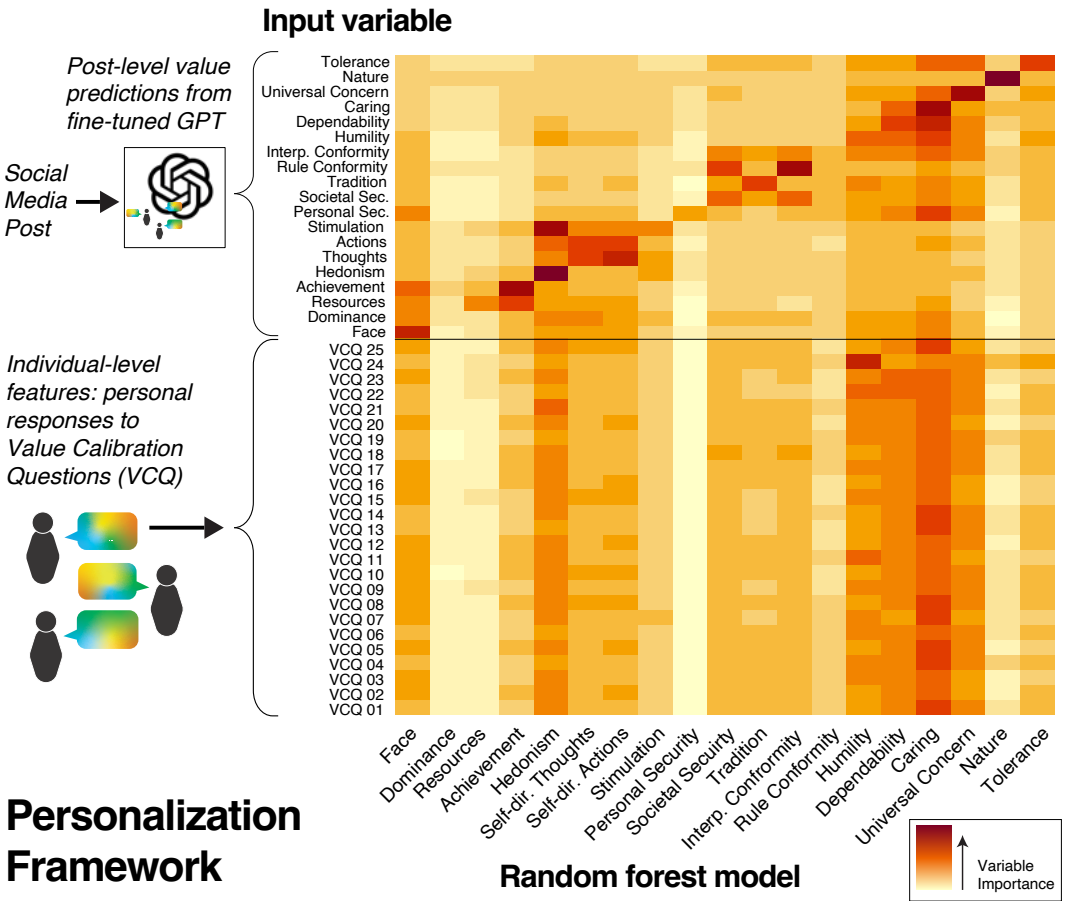

Figure A5: Heatmap of variable importance for the 19 random forest models (measured by total decrease in node impurities).