

Shaped by Values: Language Through the Lens of Culture and Morality

Neele Falk¹, David Jurgens², and Gabriella Lapesa³

¹University of Stuttgart ²University of Michigan

³GESIS - Leibniz Institute for the Social Sciences & Heinrich-Heine University Düsseldorf

¹neele.falk@ims.uni-stuttgart.de ²jurgens@umich.edu

¹gabriella.lapesa@gesis.org

Abstract

Morals and culture are a crucial driver of our behavior and social perception: unsurprisingly, social psychology provides established tools to investigate them in individuals. NLP, on the other hand, has focussed on the interpretation side, with annotations (and models) that, at best, reconstruct the inference an annotator (the reader, in this paper) made of someone else's (the writer) text, knowing little (or nothing) about the reader, nothing about the writer, and focussing on WEIRD countries. This is a dire resource gap, that we fill with the MCWR dataset, covering Mexican Spanish, Chinese, and Arabic, alongside English (US). An absolute novelty with respect to previous work, we adopt a two-step collection process, covering both generation and annotation, and collect value profiles for writers and readers. This enables us to investigate the impact of value mismatches in the annotation output: we find that readers' perceptions of writers' values are strongly shaped by their own values, causing misalignments which in turn raise concerns about generalizations made by models based on reader/writer agnostic data.¹

1 Introduction

An individual's morals and values are fundamental to how they navigate through choices in everyday life (Sun and Goodwin, 2020). They form an important part of individual and social identity (Copp, 2001) and are widely shared across societies, even though their prioritization varies between individuals and cultures (Graham et al., 2013; Haidt, 2013; Hofstede, 2001). In cultural psychology, morals and values² are defined as

¹Anonymized data available on request via <https://github.com/neelefalk/shaped-by-values>.

²For simplicity, we adopt the term "value" to denote an overarching category that includes constructs such as values, morals, and norms, and refer to the core dimensions of the two theoretical frameworks used in this paper.

relatively stable dispositions that guide attitudes, judgments, and behavior across contexts.

However, in practice, values are rarely expressed explicitly in everyday talk (Atari et al., 2023b). This paucity creates an inference gap: while a person may not say much about their values, other people readily draw inferences about the beliefs, identities, and dispositions of others based on language (Brambilla et al., 2012; Goodwin et al., 2014). For example, if a person reports "going to a protest", one person may infer that they do not respect legitimate authority, while another may infer that they highly favor equality. Such inferences are implicit and intuitive, shaped as much by the participants' own value systems as by the linguistic signal itself. Understanding how value systems influence language perception is thus essential for studying the relationship between language and identity in social context.

In this work, we consider two well-established frameworks from cultural psychology: Moral Foundations Theory (MFT; Haidt and Joseph, 2004; Haidt and Graham, 2007; Graham et al., 2013; Haidt, 2013) and Hofstede's Cultural Dimensions (HCD; Hofstede, 1980, 2001). Both frameworks quantify an individual's morals/values along a fixed set of axes like "proportionality" (MFT), i.e., whether or not individuals should be rewarded proportionally to their work.

Prior work in NLP has begun to examine moral and value-related content in text (Atari and Dehghani, 2022), typically focussing on detecting explicit moral signals or extracting value-related categories from language using predefined lexicons or supervised annotation (Reinig et al., 2024; Kiesel et al., 2022). However, these approaches model values as properties of the text itself and often rely on majority-vote annotations, implicitly assuming that moral interpretation is largely objective. As a result, **the subjective nature of value-based interpretation—and the extent**

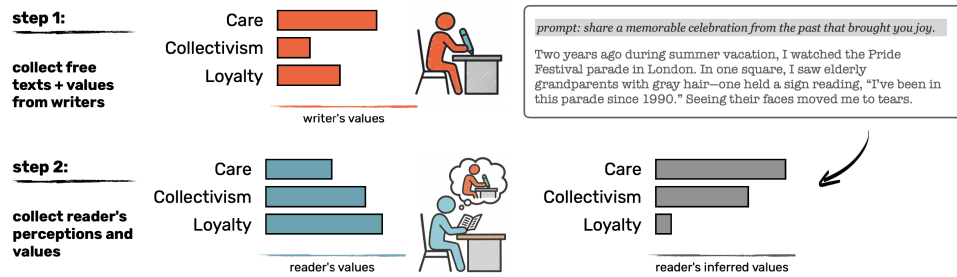


Figure 1: Overview: Collection of the MCWR Dataset

to which annotators' own value systems shape their judgments—remains underexplored.

A second limitation of existing resources concerns the lack of contextual information about the participants involved in the communicative process. In most existing annotation setups, annotators infer properties of the text or its author without any information about who produced the text. Consequently, datasets rarely capture the relationship between three key elements: the inherent values of the person *who* produced the texts (in this paper, the *writer*), the inherent values of *who* is interpreting it (the *reader*, understood here as annotator) and the values inferred by readers. **If we do not know who a writer and a reader are (what are their backgrounds, and own values), all we are left with is the output of the interpretation process – for such a subjective domain, this is extremely limiting** and it raises concerns about what NLP tools trained on limited annotations are *actually* measuring (or aligning models to).

We address these limitations by introducing MCWR (Morals and Cultural values from Writers and Readers). Figure 1 illustrates our **two-step data collection procedure**: we ask a first set of participants (the writers) to generate texts in scenarios that we expect to trigger value-laden content, i.e., arguments on controversial issues and personal experiences. Next, we ask a second set of participants (the readers) to infer the values of a text's writer without being given any information about the writer. In both steps, we enrich the dataset with contextual information that is typically absent from existing resources: we collect the inherent value profiles of the writers and readers. This allows us to **analyse how readers project their own value systems onto textual interpretation, to examine when readers correctly infer writers' values and when interpretations diverge**. This is an absolute nov-

elty of MCWR. In contrast to existing resources, MCWR is also culturally diverse: beyond the predominantly English and Western data (Vida et al., 2023). In addition to English data, we include Spanish from Mexico, Arabic from multiple countries, and Chinese from multiple regions, allowing us to investigate language-specific and cross-cultural patterns.

The unique setup of MCWR allows us address two research questions. **RQ1:** Which annotator characteristics systematically influence the interpretation of values and culture in text? **RQ2:** In which cases do readers form impressions of the writer that are inaccurate / misaligned?

The contributions of this work are at multiple levels. First, MCWR closes an important resource gap by addressing the Western-centric focus of prior work, largely centered on (U.S.) English. MCWR contains **over 8000 annotations from 200 writers and nearly 500 readers across four languages**: English and three non-WEIRD languages—Arabic, Chinese, and Mexican Spanish— typologically and culturally diverse and among the most widely spoken languages worldwide.³ We combine two complementary value frameworks (MFT and HCD) and collect value profiles for both authors and readers, contextual information that is missing from existing datasets. We contribute to perspectivism (Cabrita et al., 2023) by establishing subjective value interpretation as a factor in annotator variation.

Second, at the level of substantive research on moral and cultural values, thanks to MCWR's novel annotation we uncover a strong egocentric bias in value perception: readers' own values substantially shape how they infer the values of writers (RQ1), and this bias is a major source of misalignment (RQ2). Our multilingual design further allows us to distinguish culture-specific patterns

³ethnologue.com/insights/ethnologue200

from those that generalize across languages.

2 Related Work

Quantifying Morality and Culture. A person’s morals, values, and culture all influence their decisions (Sun and Goodwin, 2020). Among the many proposed frameworks, for morals, Moral Foundation Theory (MFT; Haidt and Joseph, 2004; Haidt and Graham, 2007; Graham et al., 2013; Haidt, 2013) has emerged as a dominant framework for explaining a person’s moral intuitions and judgments. The latest MFT framework (2.0) posits six fundamental axes, which we describe in detail in §3. An individual’s relative prioritization of these foundations unconsciously shape their decisions and emotional reactions, including how they communicate. Prior work has also validated the applicability of MFT across cultures (Atari et al., 2023a), providing translations for the Moral Foundations Questionnaire (MFQ2) survey instrument in multiple languages for measuring an individual’s relative priorities among the foundations. The other leading framework, Schwartz’ Value Theory (SVT; Schwartz, 1992; Schwartz et al., 2012) posits 10 dimensions that describe a person’s general values that motivate them—what they value (Schwartz, 2012). Here, we focus on MFT, which is more widely used than SVT in cross-group settings (Graham et al., 2011), and select the cross-culture adapted MFQ2 questionnaire.

In parallel, complementary work has aimed to quantify culture to describe differences between groups or countries. Early work was introduced by Geert Hofstede (1980; 2001) who identified five cultural dimensions which were refined through large-scale surveys to measure cultural values (Hofstede and Minkov, 2013). More recent work has added additional nuance or specific focus for defining cultural factors, such as the European Values Survey (EVS) or World Values Survey (WVS) (Inglehart et al., 1998; Haerpfer et al., 2024). As with the MFT, the definitions of the cultural dimensions have continued to be refined and expanded to be more applicable to a wider cultural context (Dorfman and Howell, 1988; Furrer et al., 2000; Vitell et al., 2003; Yoo et al., 2011; Taras et al., 2013). While most work on cultural values has restricted itself comparing mean cultural values scores between nations, recent work has adapted these instruments to measure an indi-

vidual’s cultural values; a notable example is the CVSCALE (Yoo et al., 2011). It was adapted from Hofstede’s original tool and has shown good reliability and validity in a meta-study (Taras et al., 2023) compared with simply reusing Hofstede’s instrument (Gerlach and Eriksson, 2021). We thus adopt CVSCALE as a questionnaire, and list the specific dimensions in §3

Egocentric Anchoring and Perspectivism Measuring an individual’s morals or culture is typically done through carefully constructed survey instruments. However, people readily make these inferences about others based on their statements and actions (Winter and Uleman, 1984; Carlson et al., 2022). Moreover, these value judgments are influenced in part by the perceiver’s own morals and culture (e.g., Na and Kitayama, 2011; Napier and Luguri, 2013). Thus, the perception of others’ morals and culture also has egocentric anchoring on the perceiver’s own (Ross et al., 1977; Krueger and Clement, 1994; Epley et al., 2004). Our work here extends this line of research by comparing how annotators’ own values inform their judgment of the values of a text’s author.

Analogous research in NLP has also contextualized the role of the annotator’s own identity in informing their judgments (Basile et al., 2021; Prabhakaran et al., 2021; Davani et al., 2022). This perspectivist view argues that there is no single ground truth and that multiple legitimate differences in interpretation which are informed by the annotators culture, background, and values (Plepi et al., 2022; Fleisig et al., 2024; Frenda et al., 2025). Work in this area has shown that such factors can explain the variation in multiple types of tasks, such as rating toxicity, politeness, or moral judgments (e.g., Sap et al., 2022; Gordon et al., 2022; Pei and Jurgens, 2023; Xu et al., 2025). Our work builds on this tradition, showing how individual’s own morals influences their perception of a subjective quality (values of a text’s author), but goes significantly further by quantifying the gap between the author’s actual values.

Research on Morality and Culture in NLP The most-related strand of NLP research focuses on inferring values or moral from text. This task is commonly framed as *reader-focused*: what values can be inferred from text. These computationally-labeled values (cf. Zangari et al. (2025) for an overview of modeling approaches) are used to address research questions from social sciences, i.e., the impact of morally framed arguments on per-

suasion (Kobbe et al., 2020; Alshomary et al., 2022), their relationship to political ideology (Fulgoni et al., 2016; Johnson and Goldwasser, 2018), value conflicts in disagreements in online discussions (van der Meer et al., 2023), relationship between language variation and variation in cultural background (Havaladar et al., 2024).

Datasets with value annotations are mostly in English (Vida et al., 2023), e.g., the Moral Foundations Twitter Corpus (Hoover et al., 2020) or Moral Foundations Reddit Corpus (Trager et al., 2022). Some collections extend exploration to other languages, e.g. the Moral Foundations Weibo Corpus in Chinese (Cao et al., 2024) or cultural contexts, e.g., the original corpus on Human Values (Kiesel et al., 2022) was extended to contain more arguments from India, China or Islamic religious contexts (Mirzakhmedova et al., 2024). Human annotation typically uses extensive guidelines to label texts, but a recurring challenge is that interpretations of values are highly context-dependent and influenced by individual or cultural backgrounds (Alvarez Nogales and Araque, 2024). Detailed information about instances where annotators disagree—and the value systems that shape their judgments—is a critical gap within the field, which our work fills.

Few NLP works have investigated *author-focused* tasks for assessing how values manifest implicitly in language, none have collected the “ground truth” on both sides, writer and reader, and none in the value literature. The closest analogs are Troiano et al. (2022), Wegge et al. (2022), and Kross et al. (2019), who study emotions. Our work fills this key gap for values, connecting readers’ perceptions to the writer’s reality.

3 Data Collection

We build MCWR in two steps to collect: (1) written free texts from writers, along with their self-reported values, and (2) readers’ perceptions about their inferences of the values of those texts’ writers, along with readers’ self-reported values.

Participants We use Prolific to collect the data. For English we sample from the US which offers a large pool of participants, and we further stratify the pool based on political ideology, which is known to be predictive of different values (Hatemi et al., 2019). For Spanish, we restrict participants to Mexico with Spanish as their first and primary language. Collecting data for Chinese and Arabic

is more challenging, since Prolific is not available in many countries. We can only screen by first or primary language, resulting in a smaller and more culturally diverse participant pool, including people living abroad for study or work.⁴

Psychometric instruments For both writers and readers we collect self-reported moral foundations and cultural dimensions, relying on standard psychometric instruments, the MFQ2 from (Atari et al., 2023a) for MFT and the CVSCALE from Yoo et al. (2011) for HCD.

The MFQ measures six foundations: **Care** involves protecting others from harm. **Equality** focuses on equal outcomes and fair treatment. **Proportionality** relates to the belief that rewards or punishments should match a person’s contribution. **Authority** emphasizes respect for tradition and legitimate leadership. **Loyalty** relates to solidarity with one’s group, and **Purity** to maintaining the purity and balance of body and soul.

The CVSCALE captures five cultural dimensions, each dimension represents a continuum between two value orientations: **Individualism vs. Collectivism** contrasts personal achievement with group harmony. **Uncertainty Avoidance** reflects a preference for structure versus ambiguity. **Long- vs. Short-Term Orientation** distinguishes present-focused from future-oriented thinking. **Gender Egalitarianism** concerns the extent of gender-specific roles. **Power Distance** describes whether societies are more hierarchical or egalitarian (see Appendix, Tab. 3 for more detailed definitions).

These questionnaires contain statements such as “I think people should be rewarded in proportion to what they contribute.” (MFT:Proportionality) or “Group welfare is more important than individual rewards.” (CVSCALE:Proportionality) Participants rate each statement on a 1–5 scale, from ‘It does not describe me at all’ to ‘It describes me extremely well’ for MFT and ‘Strongly Disagree’ to ‘Strongly Agree’ for HCD. Each value is measured using several such items, and the score for a dimension is computed by averaging the corresponding item responses. In total, participants complete 50 items once at the end of the study to report their values.

For MFT, higher scores indicate that a participant places greater importance on a given moral

⁴Crowdworkers received a minimum of 13 EUR/hour (plus bonuses), corresponding to the minimum wage in Germany and higher than minimum wages in the other countries.

principle (e.g., a score ≥ 4 on equality suggests that this value is particularly important to the individual). For HCD, scores represent positions along a continuum between two poles. For example, the dimension individualism–collectivism ranges from more collectivist orientations (lower scores) to more individualist orientations (higher scores). In the annotation step, readers are shown the same questions, using a simplified version of the questionnaires (§3.2). We measure questionnaire reliability with Cronbach’s α , as customary in survey research (details in App. §C).

Sociodemographics For both readers and writers, we collect sociodemographic data (age, gender, ethnicity, mother tongue, education, country of residence, occupation). Most participants also report “cultural influence”—all countries they’ve lived in or have close family, years lived in China (Chinese speakers), and Arabic variety/usage context (Arabic speakers). We also collect and normalize political leaning as free text. App. Tables 7 and 8 show the sociodemographic distributions.

3.1 Part I: Writers

Writing prompts were designed to encourage writers to produce more value-laden texts in five categories. *Personal experiences* asks about the writer’s childhood or everyday life, for instance recounting a family celebration or conflict at work. Using *interview questions*, writers are presented with questions commonly found in job interviews, designed to reveal personal values, personality traits, or to encourage more authentic responses (e.g. “What will you never spend money on?”; or “Who is the smartest person you know personally? Why?”). In *arguments* we ask writers to formulate an argument on a controversial claim (e.g. “We have no right to make animals suffer, even though there are benefits for us.”) expressing a stance of their choice. Finally, we use two types of *dialogue prompts*, asking writers to respond to an existing comment as if in a conversation with a friend. Shorter dialogue prompts are drawn from the ValueNet Corpus (Qiu et al., 2022), based on value-related scenarios from SOCIAL-CHEM-101 (Forbes et al., 2020) (e.g., “My mother encouraged me to take a day off from work, and I’m extremely happy about the decision”) Prompts for the *long dialogue* categories were selected from the AITA Subreddit in which people present conflict situations and ask how others would interpret them. Appendix §E reports the full list of writing

prompts, grouped by prompt type.

For the other languages (Spanish, Chinese, Arabic) we automatically translate the prompts and task two native speakers to annotate the prompts for (a) translation quality and (b) plausibility in the corresponding cultural context to filter out low quality prompts. The resulting list of unique prompts ranges between 66 and 86, depending on the language. Each writer is prompted with 10 different writing prompts and tasked to write at least 4 (or more) sentences. The writer was able to skip a prompt if writing about an issue made them feel uncomfortable. The written texts were manually checked by native speakers and a few cases of LLM-generated texts were discarded. All writing prompts are listed in App. E.

3.2 Part II: Readers

The goal of collecting perceptions of the reader was to investigate how recipients intuitively make assumptions about writers based on the content (or style) of text while at the same time potentially being influenced by their own values.

To capture a wider range of perspectives, we collect approximately 5 reader interpretations for each text. Each reader receives 20 texts from different writers to annotate. We encourage readers to rely on their intuition and avoid overthinking. To reduce annotator burden, we use a simplified version of the self-report questionnaire with 1–4 questions per value for each text (full value-items list and more details on scale development, annotator quality checks and validity of the annotator questions can be found in App. §C). Each item is presented as a statement, rated by readers on how much they think the writer agrees or disagrees (5-point Likert scale). As with the original psychometric instrument, the final scores for each dimension are calculated by averaging the relevant item ratings. We implemented the annotation using Potato (Jurgens et al., 2026, see screenshots in App. Figs. 4 and 5). Table 1 and Appendix Table 14 display MCWR examples, one for text type.

3.3 Quantitative properties

Table 2 provides an overview of the quantitative properties of our dataset, while Figure 2 visualizes the distribution of self-reported scores for five different values (see Fig 6, Appendix for the remaining values).

Type	Prompt	Text	Value	W	R	A
arg	Express your opinion about the following sentence: Regulations on free speech protect individuals from harm.	Unfortunately, there are regulations on everything Because of certain individuals and groups who have harmful intentions, it is necessary to some degree It's impossible to make everyone happy all the time It is a double-edged sword.	collectivism	1.00	4.00	4.00
exp	Share about a time you overcame a challenging situation. What was the situation and how did you overcome it? What did you learn from that experience?	This occurred when I struggled with an opiate addiction. I was addicted to heroin for years and struggled to live a normal life Nobody trusted me and I was financially, mentally, and spiritually unstable I lost a lot of friends and normalcy in my life I overcame it by going to rehab and exchanging that bad habit for better ones like my career and fitness	purity	2.17	4.50	4.50

Table 1: Examples from MCWR with an annotation example. Type = stimulus type: arg(ument), int(erview), exp(erience). Prompt = instruction prompt for the writer. Value = moral foundation or cultural dimension selected for the example. W = writer self-reported value. R = reader self-reported value. A = Annotation.

source	#(ann)	#(w)	#(r)	#(t)	arg	int.	exp	dial
US	2097	50	138	493	99	98	148	98
MEX	1990	50	120	495	99	99	149	99
ZH	2094	50	120	495	100	99	147	100
AR	1842	50	120	498	100	100	150	98
Total	8023	200	498	1981	398	396	594	395

Table 2: Overview of MCWR composition, including the number of annotations (#ann), w(riters), r(eaders), and t(exts) for each language / population group (US, MEX (Mexico), ZH (Chinese), AR (Arabic)). The table also details the distribution of text types: arg(uments), int(erview), exp(erience), and dial(ogue).

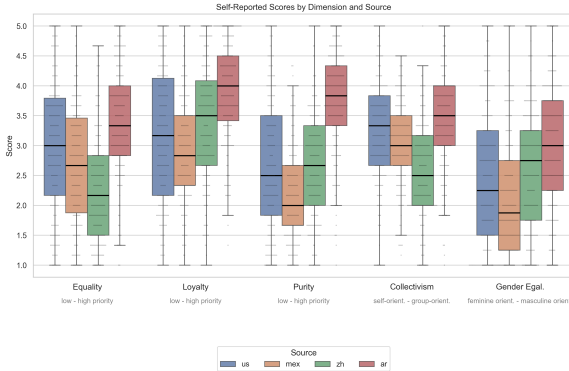


Figure 2: Distribution of self-reported scores for all participants on 5 different values, grouped by source (language). Left: moral foundations (score of 5 = full endorsement). Right: cultural dimensions (scores range from 1 to 5, e.g., 1 = individualistic or self-oriented, 5 = collectivistic or group-oriented).

Figure 2 reveals several distinct differences between the language groups. The US shows a larger variance in values than the other language groups, likely due to political stratification in the sample, leading to a wide mix of both, participants scoring high on binding and individualizing foundations. While they score lowest on power distance,

they are more collectivist than expected. **Mexico** generally shows the lowest mean across most values and is surprisingly low on Purity, with only a few outliers scoring high; they also lean the most toward a feminine orientation with few participants leaning toward traditional roles. The **Chinese** data aligns most with existing work, appearing as the most collectivist group with high scores on binding foundations and proportionality, but lower scores on individualizing and feminine orientations. The **Arabic** group consistently shows a higher mean across all values, standing out most on Purity and Loyalty, though they also score high on individualizing foundations. Interestingly, the Arabic sample spans a wider range of power distance, with some participants being more hierarchical and others more egalitarian. A quantitative analysis of the alignment between writer values and the interpretation of the readers can be found in Appendix H.

4 Influence of Readers' Backgrounds

In what follows, we investigate how annotator identity (their self-reported values and socio-demographic variables) influence their value annotations to tackle **RQ1: Which annotator characteristics systematically influence the interpretation of values and culture in text?**

Experimental Setup To identify which characteristics drive value perception, we fit 44 linear mixed-effects models: one per value per language group. The dependent variable is the annotator's inferred score for that value on a text (e.g. Care); fixed effects are the annotator's eleven self-reported values and their socio-demographics. We include random effects for Annotator ID and Text ID, capturing systematic differences between

readers and between texts. The specification is identical across all 44 models; predictors and outcomes are standardized within language.

We report **marginal** R^2 (R_m^2), the variance explained by the fixed effects, and **conditional** R^2 (R_c^2), the variance explained by fixed and random effects jointly. A large gap between R_m^2 and R_c^2 indicates that much of the explained variance comes from stable differences not captured by the predictors; we decompose this variance using the **intraclass correlations** (ICCs) of the two random effects, which distinguish annotator- from text-driven variance. While R_m^2 summarizes the fixed effects jointly, individual predictors are characterized by their standardized coefficients (direction and magnitude, holding other predictors constant) and **semi-partial** R^2 (sr^2), which captures the drop in R_m^2 when that predictor is removed and thus its unique contribution. Given correlated predictors, a large coefficient with small sr^2 indicates variance shared with other values rather than a predictor-specific effect.

4.1 Results

App. Table 15 reports marginal and conditional R^2 together with the intraclass correlations for Text ID and Annotator ID. **Value inference depends more on the annotator than on the text.** Across all languages and values, the fixed effects explain much less variance than the random effects: R_m^2 ranges from 3% to 23%, while R_c^2 ranges from 32% to 60%. When looking at the random-effect variance separately, the annotator accounts for most of it. Annotator ICCs (mean 28–40% across languages) are always larger than text ICCs (mean 7–14%). Consequently, who reads a text matters more for the inferred values than which text is read, highlighting the subjective nature of the task. **Annotators agree more on some values than on others.** Care has the highest text ICC in every language (13–31%), indicating that ratings are relatively consistent across annotators: different annotators tend to agree on how caring an author appears to be based on the text. In contrast, Long-Term Orientation, Uncertainty Avoidance, and Purity have the lowest text ICCs (3–13%) and the highest annotator ICCs (up to 52%). This suggests that these dimensions are less consistently signaled by the text itself. Instead, ratings vary more systematically across annotators, meaning that the inferred score depends strongly on who is doing the reading.

Annotator background still shapes value perception. Fixed effects explain 3–23% of the variance. Strongest models are Authority (22.6% in AR), Uncertainty Avoidance (21.9% in AR), and Purity (21.5% in US); the weakest are Care (3.0% in ZH) and Uncertainty Avoidance (3.9% in ZH). Explanatory power is highest in AR, lowest in ZH.

Which values influence annotators’ interpretation? Because each of the eleven values serves both as a predictor and as an outcome, the coefficients form an 11×11 matrix per language. Diagonal cells capture *projection*, the effect of an annotator’s own score on the matching inferred value. Off-diagonal cells capture *cross-value influence*. We correct for multiple comparisons within each language.⁵ Fig. 3 shows the diagonal effects for the values with strongest effects (full plot in App. 9).

Annotators project their own values onto writers. Diagonal effects are positive for most values (Fig. 3), with the exception of Care in AR and Proportionality in MEX. Equality and Collectivism are the most robust: both have narrow confidence intervals and survive correction in three of the four languages. Purity and Power Distance survive in two languages each, Purity in MEX and US and Power Distance in AR and US. Authority survives only in MEX. No value is significant in all four languages. The largest single effect is Purity in US, followed by Uncertainty Avoidance and Power Distance in AR. These are also the predictors with the most unique variance, with sr^2 reaching 7.2% (Uncertainty Avoidance, AR), 5.7% (Power Distance, AR), and 5.5% (Purity, US). Care, Proportionality, and Long-Term Orientation show no diagonal effect in any language.

Effects from cross-value influence are small and not consistent. For example, Uncertainty Avoidance raises perceived Long-Term Orientation in AR ($sr^2 = 3.7\%$), and Collectivism raises perceived Purity in MEX ($sr^2 = 1\%$). Most off-diagonal cells contribute below 1% unique variance each and vary in size and direction across languages. With 120 annotators per language, these estimates are too imprecise to separate language-specific effects from noise.

Socio-demographics. When accounting for annotators’ values, individual socio-demographic

⁵We treat the 11 diagonal terms as confirmatory and apply Holm correction; the 110 off-diagonal terms and demographic terms are exploratory and corrected using Benjamini–Hochberg. Adjusted within each language.

variables do not provide additional explanatory power and are generally non-significant. The exception is AR, where combinations of socio-demographic variables improve model fit for some values (App. Tab. 16). Thus, annotators' values appear to matter more for value perception than the socio-demographic variables typically used to study annotator variation in NLP.

5 What predicts misalignment?

In this section, we model *misalignment*: cases where an annotator attributes a value to a text that differs from the writer's self-report—for example, perceiving a writer as group-oriented (high on Collectivism) when they describe themselves as self-oriented (low on Collectivism). These are not annotation errors, but differences between annotator inference and writer self-report. We ask which annotator characteristics predict such differences, hypothesizing that greater reader–writer dissimilarity in values or socio-demographics leads to greater misalignment, thereby addressing **RQ2: In which cases do readers form impressions of the writer that are misaligned?**

Experimental Setup We perform a mixed-effect logistic regression analysis, similar in design to the previous setup with one model per value. Our dependent variable is a binary variable indicating whether the annotator incorrectly inferred the stance on a certain value. We include a random effect for the writer ID and reader ID. We define fixed effects that capture the relationship between reader and writer. These include the absolute distance on each value, allowing us to identify which values—when similar or different—are associated with greater misalignment. Additionally, we include alignment on socio-demographics as a binary categorical variable (e.g., considering reader/writer aligned if they belong to the same age group, and not aligned otherwise).

5.1 Results

To what extent does annotator-writer distance predict misinterpretation? App. Table 22 displays marginal and conditional R^2 by value and language. Distance on values and socio-demographics accounts for 2–9% of the misinterpretation variance. **Authority is among the best explained values in every language (6–9%):** misinterpretation on it is the one most tied to the characteristics of annotator and writer. Purity is

comparably well explained in US and AR (around 8%). Collectivism is less well explained everywhere (2.0–5.0%), with conditional R^2 above 70% in MEX and ZH, indicating that most variation reflects factors which are not measured.

The variance decomposition shows whether misinterpretation is driven more by the writer or annotator. For Care, Equality, Loyalty, and Collectivism, writer ICC is consistently much larger than annotator ICC (up to 73.7% for Collectivism in MEX), suggesting that misinterpretation depends mainly on whose text is being read. For Uncertainty Avoidance, Long-Term Orientation, and Power Distance, the pattern is reversed (up to 44.8% annotator ICC for Uncertainty Avoidance in US), indicating that misinterpretation depends more on who is reading. As in the first Analysis we find that some values are more strongly tied to the writer here, while other values are more subjective in their interpretation.

How does annotator-writer distance influence concrete values? Fig. 3 shows the diagonal effects, so the effect of distance on a value on misinterpretation of that same value. The full matrices and the semi-partial R^2 tables are in App. Fig. 10 and Tables 24–27.

Distance on a value predicts misinterpretation of that same value. The greater the annotator–writer distance on a value, the more likely the annotator is to misread it. This holds across all four languages for Equality and Authority, and in three of four for Purity, Collectivism, and Gender Egalitarianism. Uncertainty Avoidance shows no effect, while Care and Proportionality show only weak effects. Distance on the matching value accounts for much of the explanatory power: in several cases, it explains more than half of the variance explained by all eleven values combined ($sr^2 = 5.4\%$ of 8.7% for Purity in US; 4.7% of 6.5% for Power Distance in US). Compared with Analysis 1, these effects are larger and cleaner: value distance is a stronger predictor of misinterpretation than annotators' own values are of their perceptions.

Cross-value influence is weak. Again, cross-value effects are weaker and not significant after correction. Language-specific effects are for example a negative relationship between Care and Purity in AR ($sr^2 = 2\%$) or Equality and Long-Term ($sr^2 = 2\%$).

Socio-demographic alignment matters jointly. No individual socio-demographic contrast

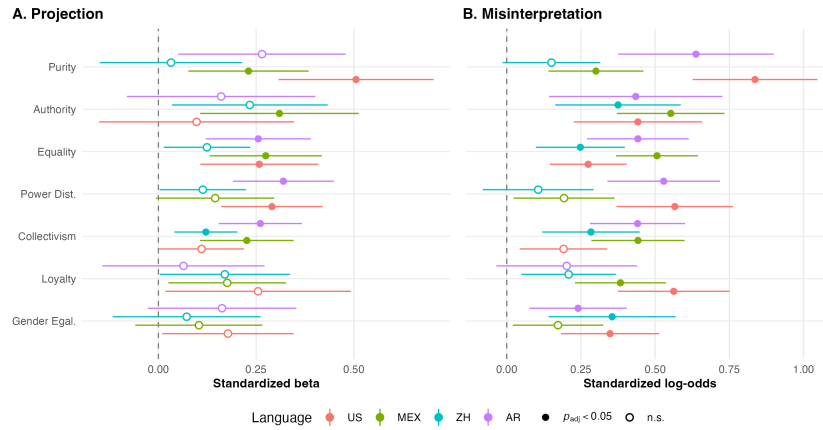


Figure 3: Own-value (diagonal) effects by language. **A:** annotator value predicting the same perceived writer value (standardized beta). **B:** annotator–writer misalignment predicting conflict on the same value (standardized log-odds). Filled points are significant after Holm correction; bars are 95% CIs. Axis units differ between panels.

survives correction in any language, but the demographic block significantly improves model fit, suggesting that background differences matter in combination. For instance, age alignment in ZH is associated with lower misinterpretation on ten of eleven values, while in AR, alignment in the context of Arabic use is associated with higher misinterpretation (App. Fig. 11).

Morals and cultural dimensions behave alike. Across both analyses, moral foundations and cultural dimensions show the same pattern: Power Distance and Gender Egalitarianism behave like Authority and Purity, while Care and Proportionality show no effect in either analysis. The key distinction is therefore not between moral and cultural values, but between values that texts signal clearly and those they do not. Values with high text ICCs (e.g., Care) show little projection or misinterpretation related to annotator characteristics, whereas values with low text ICCs (e.g., Purity and Power Distance) are more strongly shaped by annotators’ own values.

6 Conclusion

NLP models for sociodemographic and psychosocial properties assume that the content and style an author’s text is influenced by these properties. However, research in this space has rarely captured an author’s actual properties and instead used third parties to infer them. Here, we introduce a new dataset, MCWR, for studying the linguistic expression and perception of morals and culture, which enables analyzing how closely readers’ interpretations align with writers’ under-

lying values. We find significant egocentric bias in reader’s perceptions where readers project their own values and attitudes onto others. This projection leads to misinterpretations when the assumed value does not align with the other person’s actual value. Further, projection is value- and context-sensitive: Where a text signals a value clearly annotators converge and their own values matter little, but where the signal is weak annotators fill the gap by projecting their own values onto the text.

These findings raise concerns for NLP research on morals and values, where annotator differences are often treated as noise to be reduced. Our results show which values annotators are most likely to project onto texts, suggesting that guidelines alone are unlikely to eliminate this variation. Writer-agnostic datasets may therefore be misaligned, limiting models’ ability to make aligned moral judgments (Talat et al., 2022). Instead, the rich annotator profiles we collect can inform multi-annotator models through value-based groupings that go beyond socio-demographics.

Our findings and MCWR also support perspectivist approaches to NLP by enabling alignment evaluation through both constrained (close-ended questionnaires) and unconstrained (open-ended texts) measures of DSM (Röttger et al., 2024). As the field moves toward open-ended evaluation (Ma et al., 2026), datasets like MCWR, which measure values both directly and through their latent expression in open-ended text, will become increasingly important.

7 Limitations

The first limitation of our work is that, even with our best efforts of recruiting participants who are speakers of Chinese and Arabic, their status is different from that of speakers of English and Mexican Spanish as an immigration background needs to be taken into account. This is a limitation of Prolific, and we tried to compensate it with additional questions in the questionnaire.

Another factor we addressed with native speakers checks, but that we want by no means oversimplify, is the contrast between standard Arabic and the different varieties of Arabic spoken by our participants. If crowdsourcing wants to go beyond WEIRD, this is of extreme relevance. Another limitation we faced, in particular with Arabic, is the smaller number of annotators registered on the platform, which makes sampling annotators – and potentially, replication studies – more challenging.

Another limitation is the fact that, for the annotation to be feasible in a crowdsourcing scenario, we had to reduce the number of item questions in our instruments. While this limits comparison to "standard" use of these instruments, the careful procedure we employed for subsampling item questions and evaluating reliability is one of our contributions to the community – as other researchers can use the two instruments to conduct their own crowdsourcing studies.

While we measure and briefly discuss agreement among readers in the annotation scenario, we have decided, for this paper, to focus on the (mis)alignment between self-reported writer's values and interpreted values in the annotation scenario. This is due to the fact that the phenomenon we are interested in is "why a human can misunderstand another human". The question of "why two humans disagree in their (mis)understanding of another human" is a natural next step.

Despite the large scale of our collection compared to existing work, the generalization power of our sample is still limited. Our analyses show that egocentric bias is a robust effect, but cross-value effects almost never reach significance after correction; with 120 annotators per language, we cannot separate language-specific effects from noise. For that, we would need considerably more data. Two further constraints concern the samples and the instruments. The ZH and AR samples are sub-optimal in composition, which limits how far our findings extend to those populations.

The psychometric instruments also do not behave equally well across languages: some values show ceiling effects, as with Long-Term Orientation in MEX, and several correlate strongly enough that their effects cannot be separated, as with Authority, Loyalty, and Purity in US and AR. The way forward is to collect more annotations, while keeping in mind that the process we model is inherently noisy. A further annotation setup could be to show several texts from the same writer, to test whether misalignment stems from the limited information available to support the interpretation.

8 Ethical Reflection

Even though the data collection includes data from non-WEIRD contexts (Spanish in Mexico, Arabic, Chinese), the writers and the theoretical frameworks are from WEIRD contexts, and it is often found that the instruments yield lower validity outside WEIRD. The population of participants recruited via Prolific is also biased towards WEIRD (Henrich et al., 2010). As stated by Goodwin et al. (2020), to make progress in reducing power asymmetries and epistemic injustice, studies, theoretical frameworks and data collections need to be expanded to more regions of the world and to languages that are underinvestigated. Collecting more data from non-WEIRD contexts is crucial given the significant imbalance and overrepresentation of dominant groups in LLMs—a finding increasingly documented in recent research. The **intended use** of this data is to investigate the relationship between an individual's values and the way they express themselves or perceive texts written by others. Models trained on such data must be carefully evaluated and should not be used for profiling—that is, inferring individuals' moral beliefs or norms from their language without their knowledge. A more appropriate application could be to reveal potential misunderstandings and improve transparency in communication.

An important component of our contribution is the breadth of MCWR in terms of number of involved annotators and the depth of the questions we asked them, but we are aware that running such an expensive study (with adequate participant remuneration) is a privilege. We strongly support perspectivism, but we also believe that running expensive studies cannot become an "entry level" requirement – this would be yet another unethical economic bottleneck, after compute costs.

Acknowledgments

We acknowledge funding by the Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB, which funded part of the data collection costs, N.Falk and her research visit at the University of Michigan where this project kicked off. This work was also supported in part by the National Science Foundation under Grant No. IIS-2143529. We thank Milad Alshomary, Mohammad Sharabati, Hongyu Chen and Yarik Menchaca Resendiz for their support in the native speaker checks of the collection.

References

- Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. [The moral debater: A study on the computational generation of morally framed arguments](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.
- Anny D. Alvarez Nogales and Oscar Araque. 2024. [Moral disagreement over serious matters: Discovering the knowledge hidden in the perspectives](#). In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 67–77, Torino, Italia. ELRA and ICCL.
- Mohammad Atari and Morteza Dehghani. 2022. Language analysis in moral psychology. *The atlas of language analysis in psychology*, pages 207–228.
- Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023a. [Morality beyond the weird: How the nomological network of morality varies across cultures](#). *Journal of Personality and Social Psychology*, 125(5):1157–1188.
- Mohammad Atari, Matthias R Mehl, Jesse Graham, John M Doris, Norbert Schwarz, Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Elaine Gonzalez, Nikki Jafarzadeh, Alyzeh Hussain, Arineh Mirinjian, Annabelle Madden, Rhea Bhatia, Alexander Burch, Allison Harlan, David A. Sbarra, Charles L. Raison, Suzanne A. Moseley, and 2 others. 2023b. The paucity of morality in everyday talk. *Scientific Reports*, 13(1):5967.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21.
- Marco Brambilla, Simona Sacchi, Patrice Rusconi, Paolo Cherubini, and Vincent Y Yzerbyt. 2012. You want to give a good impression? be honest! moral traits dominate group impression formation. *British journal of social psychology*, 51(1):149–166.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.
- Renjie Cao, Miaoyan Hu, Jiahan Wei, and Baha Ihnaini. 2024. [The moral foundations Weibo corpus](#). In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 155–165, Miami, FL, USA. Association for Computational Linguistics.
- Ryan W Carlson, Yochanan E Bigman, Kurt Gray, Melissa J Ferguson, and MJ Crockett. 2022. How inferred motives shape moral judgements. *Nature Reviews Psychology*, 1(8):468–478.
- David Copp. 2001. *Morality, normativity, and society*. Oxford University Press, USA.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Peter Dorfman and Jon Howell. 1988. *Dimension of national culture and effective leadership patterns: Hofstede revisited*, volume 3, pages 127–149. JAI Press.
- Nicholas Epley, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich. 2004. Perspective taking as ego-centric anchoring and adjustment. *Journal of personality and social psychology*, 87(3):327.
- Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292.
- Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. [Social chemistry 101: Learning to reason about social and moral norms](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.
- Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

- Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preotăciuc-Pietro. 2016. [An empirical exploration of moral foundations theory in partisan news sources](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736, Portorož, Slovenia. European Language Resources Association (ELRA).
- Olivier Furrer, Ben Shaw-Ching Liu, and D. Sudharshan. 2000. [The relationships between culture and service quality perceptions: Basis for cross-cultural market segmentation and resource allocation](#). *Journal of Service Research*, 2(4):355–371.
- Philipp Gerlach and Kimmo Eriksson. 2021. [Measuring cultural dimensions: External validity and internal consistency of hofstede's vsm 2013 scales](#). *Frontiers in Psychology*, 12.
- Geoffrey P Goodwin, Jared Piazza, and Paul Rozin. 2014. Moral character predominates in person perception and evaluation. *Journal of personality and social psychology*, 106(1):148.
- Jamie Lynn Goodwin, Andrew Lloyd Williams, and Patricia Snell Herzog. 2020. [Cross-cultural values: A meta-analysis of major quantitative studies in the last decade \(2010–2020\)](#). *Religions*, 11(8):396.
- Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. *Moral Foundations Theory*, page 55–130. Elsevier.
- Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.
- Kurt Gray, Nicholas DiMaggio, Chelsea Schein, and Frank Kachanoff. 2022. [The problem of purity in moral psychology](#). *Personality and Social Psychology Review*, 27(3):272–308.
- Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2024. [World values survey \(1981–2022\)](#). trend file.
- Jonathan Haidt. 2013. [Moral psychology for the twenty-first century](#). *Journal of Moral Education*, 42(3):281–297.
- Jonathan Haidt and Jesse Graham. 2007. [When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize](#). *Social Justice Research*, 20(1):98–116.
- Jonathan Haidt and Craig Joseph. 2004. [Intuitive ethics: how innately prepared intuitions generate culturally variable virtues](#). *Daedalus*, 133(4):55–66.
- Peter K Hatemi, Charles Crabtree, and Kevin B Smith. 2019. Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, 63(4):788–806.
- Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. [Building knowledge-guided lexica to model cultural variation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226, Mexico City, Mexico. Association for Computational Linguistics.
- Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. [Most people are not weird](#). *Nature*, 466(7302):29–29.
- Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.
- Geert Hofstede and Michael Minkov. 2013. Values survey module 2013 manual. <https://geerthofstede.com/wp-content/uploads/2016/07/Manual-VSM-2013.pdf>. Accessed: 2025-06-02.
- Geert H. Hofstede. 1980. *Culture's consequences: International differences in work-related values*. Sage Publications, Beverly Hills, CA.
- Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. [Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment](#). *Social Psychological and Personality Science*, 11(8):1057–1071.
- Ronald Inglehart, Miguel Basanez, and Alejandro Moreno. 1998. *Human Values and Beliefs: A Cross-Cultural Sourcebook*. University of Michigan Press.
- Kristen Johnson and Dan Goldwasser. 2018. [Classification of moral foundations in microblog political discourse](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.
- David Jurgens, Michael Chen, and Lina Iyer. 2026. Potato 2.0: A comprehensive annotation platform with ai-in-the-loop support. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 374–386.

- Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. [Identifying the human values behind arguments](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.
- Jonathan Kobbe, Ines Rehbein, Ioana Hulpus, and Heiner Stuckenschmidt. 2020. [Exploring morality in argumentation](#). In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.
- Ethan Kross, Philippe Verduyn, Margaret Boyer, Britany Drake, Izzy Gainsburg, Brian Vickers, Oscar Ybarra, and John Jonides. 2019. Does counting emotion words on online social networks provide a window into people’s subjective experience of emotion? a case study on facebook. *Emotion*, 19(1):97.
- Joachim Krueger and Russell W. Clement. 1994. [The truly false consensus effect: An ineradicable and egocentric bias in social perception](#). *Journal of Personality and Social Psychology*, 67(4):596–610.
- Bolei Ma, Yong Cao, Indira Sen, Anna-Carolina Haensch, Frauke Kreuter, Barbara Plank, and Daniel Herscovich. 2026. [Too open for opinion? embracing open-endedness in large language models for social simulation](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2016, Rabat, Morocco. Association for Computational Linguistics.
- Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, MohammadAli Sadraei-Javaheri, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2024. [The touché23-ValueEval dataset for identifying human values behind arguments](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134, Torino, Italia. ELRA and ICCL.
- Jinkyung Na and Shinobu Kitayama. 2011. Spontaneous trait inference is culture-specific: Behavioral and neural evidence. *Psychological science*, 22(8):1025–1032.
- Jaime L Napier and Jamie B Luguri. 2013. Moral mind-sets: Abstract thinking increases a preference for “individualizing” over “binding” moral foundations. *Social Psychological and Personality Science*, 4(6):754–759.
- Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. In *Proceedings of the 17th linguistic annotation workshop (LAW-XVII)*, pages 252–265.
- Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the joint 15th linguistic annotation workshop (LAW) and 3rd designing meaning representations (DMR) workshop*, pages 133–138.
- Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11183–11191.
- Ines Reinig, Maria Becker, Ines Rehbein, and Simone Ponzetto. 2024. [A survey on modelling morality for text analysis](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4136–4155, Bangkok, Thailand. Association for Computational Linguistics.
- Lee Ross, David Greene, and Pamela House. 1977. [The “false consensus effect”: An egocentric bias in social perception and attribution processes](#). *Journal of Experimental Social Psychology*, 13(3):279–301.
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. [Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906.
- Shalom H Schwartz. 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, pages 1–65. Elsevier.
- Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1).
- Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönqvist, Kursad Demirutku, and 1 others. 2012. Re-

- fining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663.
- Jessie Sun and Geoffrey P Goodwin. 2020. Do people want to be more moral? *Psychological science*, 31(3):243–257.
- Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. [On the machine learning of ethical judgments from natural language](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.
- Vas Taras, Julie Rowney, and Piers Steel. 2013. [Work-related acculturation: change in individual work-related cultural values following immigration](#). *The International Journal of Human Resource Management*, 24(1):130–151.
- Vas Taras, Piers Steel, and Madelynn Stackhouse. 2023. [A comparative evaluation of seven instruments for measuring values comprising hofstede’s model of culture](#). *Journal of World Business*, 58(1):101386.
- Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. [The moral foundations reddit corpus](#). *CoRR*, abs/2208.05545.
- Enrica Troiano, Laura Ana Maria Oberlaender, Maximilian Wegge, and Roman Klinger. 2022. [x-enVENT: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1365–1375, Marseille, France. European Language Resources Association.
- Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. [Do differences in values influence disagreements in on-line discussions?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.
- Karina Vida, Judith Simon, and Anne Lauscher. 2023. [Values, ethics, morals? on the use of moral concepts in NLP research](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.
- Scott J. Vitell, Joseph G. P. Paolillo, and James L. Thomas. 2003. [The perceived role of ethics and social responsibility: A study of marketing professionals](#). *Business Ethics Quarterly*, 13(1):63–86.
- Maximilian Wegge, Enrica Troiano, Laura Ana Maria Oberlaender, and Roman Klinger. 2022. [Experiencer-specific emotion and appraisal prediction](#). In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 25–32, Abu Dhabi, UAE. Association for Computational Linguistics.
- Laraine Winter and James S Uleman. 1984. When are social judgments made? evidence for the spontaneousness of trait inferences. *Journal of personality and social psychology*, 47(2):237.
- Yinuo Xu, Veronica Derricks, Allison Earl, and David Jurgens. 2025. Modeling annotator disagreement with demographic-aware experts and synthetic perspectives. *arXiv preprint arXiv:2508.02853*.
- Boonghee Yoo, Naveen Donthu, and Tomasz Lenartowicz. 2011. [Measuring hofstede’s five dimensions of cultural values at the individual level: Development and validation of cvsca](#). *Journal of International Consumer Marketing*, 23(3-4):193 – 210. Cited by: 666.
- Lorenzo Zangari, Candida Maria Greco, Davide Picca, and Andrea Tagarelli. 2025. [A survey on moral foundation theory and pre-trained language models: current advances and challenges](#). *AI & SOCIETY*.

A Moral Foundations and Cultural Values: Definitions

Dimension	Description
Individualizing Moral Foundations	
Care	This dimension emphasizes the importance of protecting vulnerable individuals from both emotional and physical harm.
Equality	This dimension reflects the extent to which an individual is committed to equal treatment and outcomes for all people.
Proportionality	This measures the degree to which an individual believes that rewards should be distributed in proportion to a person's effort or contribution.
Binding Moral Foundations	
Loyalty	This dimension captures the importance an individual places on loyalty and strong ties within their social or cultural ingroup.
Authority	This reflects the extent to which an individual values traditions and social stability, and accepts legitimate authorities as necessary for maintaining order.
Purity	This measures how much importance an individual places on physical and spiritual cleanliness, often linked to self-discipline, harmony with nature, and healthy habits.
Cultural Dimensions	
Individualism vs. collectivism	This dimension measures how strongly an individual prioritizes either personal autonomy and the immediate family (individualism) or the needs and interests of larger social groups such as extended family or community (collectivism).
Power distance	This reflects the degree to which an individual accepts hierarchical structures and believes that people in higher positions should hold greater responsibility, privileges, and receive more respect.
Gender egalitarianism	This dimension describes whether an individual believes that gender roles are fixed and clearly defined or whether all genders should be treated equally, with flexible roles and opportunities.
Uncertainty avoidance	This measures how strongly an individual feels threatened by uncertainty or ambiguity. Those high in this dimension prefer clear regulations and risk awareness, while those low in it are more comfortable with ambiguity and open to innovation.
Long vs. Short-Term Orientation	This dimension captures whether an individual tends to focus on long-term planning, stability, and tradition (long-term orientation), or prefers to prioritize immediate enjoyment and is less concerned with future outcomes (short-term orientation).

Table 3: Overview of Moral Foundations and Cultural Dimensions: definition of each foundation / dimension.

B Annotation interface

Figure 4: Annotation View of writer, each writer is tasked to write 10 texts.

Figure 5: Annotation view of reader, each reader is tasked to read and rate 20 texts by different writers. The reader knows nothing about the writer.

C Questionnaire development, quality checks, and reliability analysis (Cronbach's α)

The MFQ2 was available in all target languages, the CVSCALE was not available for Arabic and not fully available for Mexican Spanish (long-term orientation had to be translated). In these cases we translated the questions and adapted them (improve clarity, ensure semantic equivalence) with the help of two native speakers.

To develop a good set of items we run several pilot studies and measure the Cronbach α for each of the 11 values until each dimension reached an acceptable level ($\alpha > 0.6$).

Tables 5 and 4 display the Cronbach's α for self-reported (writers and readers) and annotated values (readers) respectively. For most foundations, we reach an acceptable reliability with variations across values, and comparable patterns in the two scenarios (slightly lower values in the annotation scenario, not surprising because inferences about the values of another person should not be expected to be as consistent as for the intuitions about one-self).

Some values, such as purity and proportionality, do not generalize as well across different languages—especially in Arabic. This aligns with previous research showing that purity is a problematic value with low internal consistency (Gray et al., 2022). Furthermore, Arabic speakers who assisted with the translation anticipated that purity would be challenging for cultural reasons: not all annotators would answer honestly, and standard Arabic is most often used in religious contexts, which can interfere with how this dimension is interpreted. This leaves us with a final set of 25 statements, reported in Table 6.

We include two quality control checks: (1) including an example from the instructions that clearly shows high *care* (expecting annotators to assign a high value), and (2) presenting the same *collectivism* question twice in opposite ways (expecting opposite values). These checks filter out low-quality annotations.

C.1 Questionnaires

Collectivism	Pow. Distance	Unc. Avoidance	Long Term Orient.	Care	Equality	Purity	Proportionality	Authority	Loyalty	Lang
0.86	0.88	0.81	0.75	0.92	0.85	0.68	0.70	0.83	0.75	US
0.89	0.87	0.80	0.75	0.93	0.84	0.62	0.70	0.72	0.70	MEX
0.94	0.91	0.78	0.69	0.96	0.86	0.67	0.73	0.71	0.70	ZH
0.89	0.83	0.82	0.73	0.92	0.73	0.58	0.56	0.77	0.83	AR

Table 4: Cronbach's α coefficients for assessing the internal consistency of the constructs on the annotation side: 6 moral dimensions 4 cultural dimensions, each measured by summarized / fewer items from the original questionnaires. No value for gender egalitarianism since it was only measured with one item.

Collectivism	Gend. egalitarianism	Power Dist.	Unc. Avoidance	Long Term Orient.	Care	Equality	Purity	Proportionality	Authority	Loyalty	Lang
0.83	0.85	0.84	0.77	0.80	0.89	0.87	0.85	0.80	0.90	0.87	US
0.81	0.85	0.78	0.75	0.78	0.91	0.86	0.67	0.78	0.86	0.84	MEX
0.91	0.81	0.79	0.83	0.67	0.90	0.83	0.70	0.78	0.80	0.90	ZH
0.88	0.80	0.87	0.85	0.79	0.89	0.82	0.76	0.79	0.87	0.85	AR

Table 5: Cronbach's α coefficients for assessing the internal consistency of the constructs: 6 moral dimensions using the MFQ2 and 5 cultural dimensions using the CVScale. Higher value indicates greater internal consistency. Measured for self-reported values by both readers and writers. Cronbach's α quantifies the reliability of a questionnaire in terms of the internal consistency of survey items. It ranges from 0 to 1, and values above 0.6 are considered acceptable.

*Name	Description
Moral foundations (Source: MFQ2)	
CA1	We should care and help others who are suffering.
CA2	We should show empathy towards people in pain.
EQ1	Money should be equally distributed since that would lead to a better and fairer society.
EQ2	Resources should be shared equally among everyone, regardless of their contribution.
PRO1	Hard work should lead to greater financial rewards.
PRO2	A person should get recognized for performance and achievement but punished for dishonesty.
LOY1	Everyone should love their own community.
LOY2	It is important to be loyal to your country.
AUT1	We need to preserve traditional values.
AUT2	Having a strong leader is good for society.
AUT3	Obedience to parents is an important virtue.
PUR1	The human body should be treated like a temple, housing something sacred within.
PUR2	It upsets me when people use foul language like it is nothing.
PUR3	If I found out that an acquaintance had an unusual but harmless sexual fetish I would feel uneasy about them.
PUR4	People should try to use natural medicines rather than chemically identical human-made ones.
Cultural Dimensions (Source: CVSSCALE)	
COLL1	The needs and success of the group should come before individual desires and achievements.
COLL2	We should consider group loyalty and collective welfare over personal goals.
PD1	Decision-making should mainly be the responsibility of those in higher positions, without much input from those in lower positions.
PD2	People in lower positions shouldn't question the choices from people in higher positions or take on big tasks by themselves.
UA1	Everyone should know exactly what to do by having every step clearly explained.
UA2	Rules and regulations are important because they clearly outline what is expected of people.
UA3	Standardized work procedures are helpful.
LTO1	We should manage money wisely and think about the future.
LTO2	You should stick with your goals, even when it's tough, and work hard for what comes next.
GE1	Men and women have different roles and strengths, especially in professional and problem-solving contexts.
QUAL	Individual desires and achievements should take precedence over the needs and success of the group.

Table 6: Likert Items used in the reader study to gather perceptions from the reader about the writer's values: Name and Description. The reader responds on a Likert Scale whether they think that the writer of the text agrees or disagrees with that statement. Final score for a value is obtained by taking the average over items.

D Dataset: additional Details

D.1 Annotator socio-demographics: descriptive statistics

Language	Variable	Counts
Arabic	gender	Female: 53; Male: 62; Prefer not to say: 5
Arabic	age	18-29: 61; 30-39: 36; 40-49: 13; 50-59: 8; 60+: 2
Arabic	political	Apolitical/Unclear: 66; Center: 17; Left: 29; Right: 8
Arabic	ethnicity	Black/African: 21; Middle Eastern: 50; North African: 27; Other: 5; Prefer not to say: 5; White: 12
Arabic	occupation	Employed: 79; Homemaker: 7; Other: 3; Prefer not to say: 5; Student: 22; Unemployed: 4
Arabic	education	College/University Degree: 57; Graduate Degree: 46; High School or Below: 14; Other/Prefer not to say: 3
Arabic	arabic context	Friends: 21; Home/Family: 85; Online/Don't use: 5; Work/School: 9
Arabic	arabic variety	Don't use: 5; Local dialect: 71; Mixed: 14; Standard Arabic: 30
US	gender	male: 68; non-male: 70
US	age	18-29: 46; 30-39: 46; 40-49: 20; 50-59: 18; 60+: 8
US	ethnicity	Asian: 10; Black: 36; Other: 3; White: 89
US	political	Anti-Establishment: 3; Apolitical/Unclear: 20; Centrist: 9; Left: 59; Right: 47
US	education	College degree: 52; Graduate degree: 36; High school diploma or equivalent: 40; Less than a high school diploma: 3; Other/Unclear: 7
US	occupation	Employed: 101; Not working: 16; Other/Unknown: 5; Unemployed: 16
Mexico	gender	female: 46; male: 67; non-binary: 7
Mexico	age	18-29: 82; 30-39: 30; 40-49: 5; 50-59: 3
Mexico	ethnicity	Hispanic or Latino: 103; White: 17
Mexico	political	Anti-Establishment: 3; Apolitical/Unclear: 27; Center: 55; Left: 25; Right: 10
Mexico	education	College/University Degree: 88; Graduate/Postgraduate Degree: 9; High School or Equivalent: 12; Other: 11
Mexico	occupation	Employed: 73; Other: 1; Self-employed: 18; Student: 23; Unemployed: 5
Chinese	gender	Female: 64; Male: 53; Other/Non-binary: 3
Chinese	age	18-29: 63; 30-39: 38; 40-49: 10; 50-59: 5; 60+: 4
Chinese	political	Apolitical/Unclear: 20; Center: 66; Conservative/Right: 20; Left: 14
Chinese	occupation	Employed: 79; Homemaker/Retired: 5; Other/Prefer not to say: 3; Student: 30; Unemployed: 3
Chinese	education	College/University Degree: 60; Graduate Degree: 47; High School or Below: 8; Other/Prefer not to say: 5

Table 7: Descriptive statistics of writer demographics for each language.

Language	Variable	Counts
Arabic	gender	Female: 23; Male: 27
Arabic	age	18-29: 27; 30-39: 12; 40-49: 8; 50-59: 3
Arabic	political	Apolitical/Unclear: 27; Center: 12; Left: 9; Right: 2
Arabic	ethnicity	Middle Eastern: 32; North African: 14; Prefer not to say: 1; White: 3
Arabic	occupation	Employed: 30; Homemaker: 3; Retired: 1; Student: 13; Unemployed: 3
Arabic	education	College/University Degree: 23; Graduate Degree: 24; High School or Below: 3
Arabic	arabic context	Friends: 1; Home/Family: 48; Online/Don't use: 1
Arabic	arabic variety	Local dialect: 37; Mixed: 3; Standard Arabic: 10
US	gender	male: 15; non-male: 35
US	age	18-29: 19; 30-39: 17; 40-49: 5; 50-59: 8; 60+: 1
US	ethnicity	Asian: 1; Black: 10; Other: 4; White: 35
US	political	Anti-Establishment: 3; Apolitical/Unclear: 11; Centrist: 4; Left: 17; Other: 1; Right: 14
US	education	College degree: 19; Graduate degree: 11; High school diploma or equivalent: 19; Other/Unclear: 1
US	occupation	Employed: 35; Not working: 5; Other/Unknown: 2; Unemployed: 8
Mexico	gender	female: 23; male: 27
Mexico	age	18-29: 26; 30-39: 16; 40-49: 5; 50-59: 3
Mexico	ethnicity	Hispanic or Latino: 39; Other: 2; White: 9
Mexico	political	Anti-Establishment: 1; Apolitical/Unclear: 10; Center: 25; Left: 12; Right: 2
Mexico	education	College/University Degree: 34; Graduate/Postgraduate Degree: 7; High School or Equivalent: 7; Other: 2
Mexico	occupation	Employed: 28; Other: 2; Self-employed: 5; Student: 14; Unemployed: 1
Chinese	gender	Female: 29; Male: 20; Other/Non-binary: 1
Chinese	age	18-29: 31; 30-39: 10; 40-49: 7; 60+: 2
Chinese	political	Apolitical/Unclear: 15; Center: 26; Conservative/Right: 3; Left: 6
Chinese	occupation	Employed: 33; Other/Prefer not to say: 2; Student: 15
Chinese	education	College/University Degree: 18; Graduate Degree: 29; High School or Below: 2; Other/Prefer not to say: 1

Table 8: Descriptive statistics of reader demographics for each language.

D.2 Distribution of self-reported scores

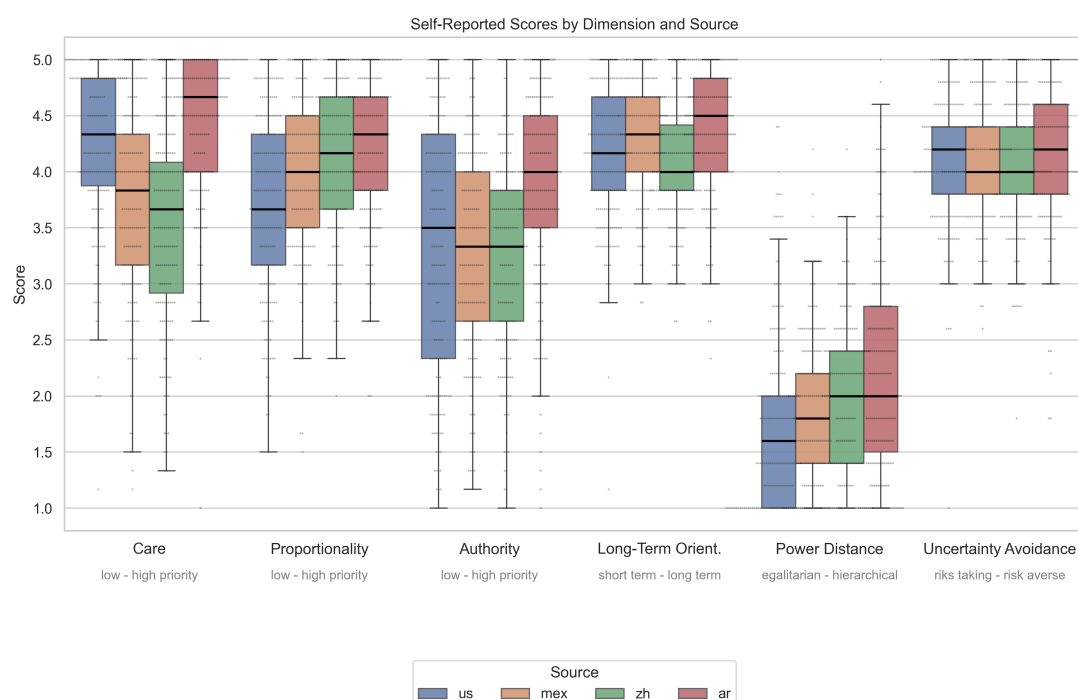


Figure 6: Distribution of self-reported scores for all participants on 5 different values, grouped by source (language). Left: moral foundations (score of 5 = full endorsement). Right: cultural dimensions (scores range from 1 to 5).

E Prompts for the writers

Share about an issue you feel strongly about. Why do you think you care so much about it? Has anyone ever challenged you on it? How did it go?
Share a personal experience about a situation where you faced a difficult decision.
Share about a time you overcame a challenging situation. What was the situation and how did you overcome it? What did you learn from that experience?
Share a personal experience about a conflict situation that happened in school or at your workplace.
Share about a time you changed your mind or thinking about something. What happened that made you change your mind? What was that experience like for you?
Share about a memory of your childhood.
What is a group that you are part of that is important to you and why?
Share a personal experience about a memorable celebration from your past that brought you joy.
Share about your day yesterday? How did you spent it? What did you do? How did you feel?
Share a personal experience about a conflict situation between you and your family or friends.
Share about a small victory you recently had.

Table 9: Stimulus type: personal experience

What's an unwritten rule in life that everyone should know?
You're stranded on a deserted island with two things with you. What are they?
Who is the smartest person you know personally? Why?
What is your idea of an amazing day?
Share about a movie you really like? What do you like about it? Why do you think you like it so much?
If you could work from any location on Earth, where would you choose and why?
Do you believe in magic? If so, how do you define it?
Upon arriving home, you realize you are the winner of a large amount of money. What do you do next?
Share about a book you really like. What do you like about it? Why do you think you like it so much?
What's an invention that would benefit you immediately and can't believe isn't a thing already?
What is the strangest thing that someone has said to you?
What will you never spend money on?
What is the most effective way to kill your time?
What are you most proud of in your professional life so far?
If you were allowed to have one superpower, what would that be?
What's your biggest mistake?
What was your most recent adventure?
What's something that should be obsolete in 2024, but isn't?

Table 10: Stimulus type: interview

EXPRESS YOUR OPINION ABOUT THE FOLLOWING SENTENCE:
Artificial intelligence represents a threat to humankind
We have no right to make animals suffer, even though there are benefits for us.
Regulations on free speech protect individuals from harm.
Climate change is a global issue facing mankind, and with carbon dioxide emissions from all countries, greenhouse gases are surging, posing a threat to life systems.
Free speech is a fundamental human right.
Punishment is sometimes the only option to keep criminals from committing more crimes.
Teaching empathy in schools would foster a more peaceful and tolerant society.
No single dietary approach is best for everyone, so people should follow the diet that is best for them.
A universal basic income would improve the quality of life of many people.
Video games add beauty, entertainment, and joy to society.

Table 11: Stimulus type: argument

WRITE A REPLY MESSAGE TO THE FOLLOWING COMMENT:
I made fun of someone else's dead family member.
Tomorrow I will audition my singing in a talent show.
I'm feeling hopeful for the first time in years and it feels wonderful. I'm finally excited for the future and believe I can tackle whatever obstacle life throws at me. Let's gooooo!
I always tip very well. But how well depends on your looks as much as promptness.
My mom has been obese my whole life. Recently, she's made amazing strides to get down to a healthy weight. I'm so proud of her efforts to be more healthy. I get a little pang of sadness when I hug her because I miss how her hugs used to feel.
I don't think children should be shielded from the reality of humanity.
I've accepted a new job offer and can't wait to give my 2 weeks to my current role!!
After our pet bird died, we packed him in a little shoe box.
This year has taught me that I value integrity a lot more than I thought.
I hate alcohol, and I think people who drink it irresponsibly are plain idiotic.
I'm afraid to be happy.
I wish that the family I'm staying with would have their dinner conversation in language I can understand so I can contribute to the conversation.
I'm graduating soon and I'm losing my energy on things. I'm still on vacation before the last semester starts, and I haven't done anything productive. I fear that once I leave my uni, my mental would crush.
I stayed in an abusive relationship because I believe I don't deserve anything better...
Some people are trying to force me to leave my dog for a while. I have thought about causing myself to have some sort of injury just so they might stop trying to force me to leave my dog.
Just because I have a degree in political science doesn't mean I want to take politics all the time.
I don't want to visit my uncle although he is sick.
I want to, when I turn 18, leave my family.
I do not feel welcome at my fiances family Christmas gatherings.
I'm falling out with my brother over his repeated use of my things when I've told him not to mess around with them.
I forced a connection between my twin sibling and someone I knew they would hate purely to watch the world burn.
I am 24 and am scared to sleep at home alone.
I voted for myself for teacher of the year, because I honestly feel like I deserve it over many others.
I came out to my extended family even though my parents don't want me to.
Yesterday, I requested that a store adhere to its policy after a cashier made an error.
I am talking to someone who my friends forbid me from being friends with.
I have radical beliefs that I have not dared to articulate to anyone yet.
I have problems to give physically affection to my partner because I'm insecure about my physical appearance.
I keep my headphones on at work so I don't have to talk to anybody.
I resent my fellow peer because of how smart she is
Whenever I walk into a room, I always put on a seemingly genuine smile...it puts everyone at ease and I receive mostly positive reactions.
My mother encouraged me to take a day off from work, and I'm extremely happy about the decision.
Yesterday my friend was proudly showing everyone her baby's first braid.
I drunkenly took someones teriyake sauce and got my position at uni threatened
I didn't deserve to graduate...
I have 4 small children and I am disabled. We are surviving without electricity because I can't afford it. I wish there was someone in this world that have a care. Im at my breaking point. I don't know what to do, its getting to hard for us.
I advised my coworkers not to thank the company president, as he did not contribute to the effort.
I tell people I will pray for them but I never do.
I went back and helped my former employees after we all quit

Table 12: Stimulus type: dialogue (short)

WRITE A REPLY MESSAGE TO THE FOLLOWING COMMENT:
My mom gives away our family spendings behind my dad's back. This has always happened, my father provides for our family and my mom is a stay at home mom, my dad would give her the spendings monthly but she sends a fair amount of them to her family, her family is greedy and they never appreciate anything, they have enough money and are always shit talking about my dad.
I turned 40 recently and had booked a 7pm dinner at a nice restaurant. Planning to leave at 6:30pm to account for traffic and picking up my dad, I found my husband (43 M) still working on his car, not ready to go. Despite urging him to hurry, by 6:50, he wasn't ready, so I left with the kids, not wanting to miss our reservation. He often runs late, but I couldn't tolerate it for my special dinner. We kept our table despite arriving late, but my husband didn't join us and was upset upon our return. I expressed my frustration with his habitual tardiness. My dad supported my decision, though the kids were disappointed he missed the dinner.
On my 25th birthday, I chose pie over cake, buying two pies to enjoy freely, despite my boyfriend's usual complaints about my pie consumption. Over three days, I finished both pies and consumed 10 sodas, which angered my boyfriend. He warned me against repeating this on my next birthday, citing health concerns. This reaction was surprising, especially since I've recovered from anorexia three years ago and generally eat healthier than he does. He often indulges in sweets and sodas without restraint. When confronted, I apologized, but he persisted, accusing me of not handling serious discussions well, which I felt was more reflective of his behavior. Given my history with food, I wonder if I was wrong for indulging during my birthday celebration.
I am having such a hard time with the healthy eating part. I need to lower my cholesterol but it's just miserable for me. I spent 31 years eating crap food, to go from high sugar high fat to healthy is like... it makes me want to bash my face into a wall.
I bought an art piece from Elle, an artist I met on Etsy, and connected with her personally. To avoid Etsy's fees, we moved our transactions to Facebook. Despite slow shipping, Elle always communicated about delays. However, my last purchase from 5 months ago hasn't arrived yet, even though the payment was cashed immediately. After promises of shipping it in December and multiple excuses including family illness, I received nothing. By February, after several unanswered attempts to reach her, I learned she's active on Instagram, teaching art business. I demanded a refund through a formal document, highlighting her failure to deliver. Elle replied, citing severe health issues as the reason for her unresponsiveness and offered a refund, stating her husband is managing most shipments now.
I enjoyed my internship until a new associate, NA, joined and disrupted the productivity with her gossiping and overly friendly behavior. After a private argument, NA breached our trust by discussing it with another intern. Her influence also led to extra work for me when another intern neglected her duties. After NA twisted my complaint to the office manager into a personal attack, I arranged to no longer assist her and informed her directly, leading to a heated argument. I've since limited my interactions with NA, focusing on my work and maintaining professional boundaries, which has improved my overall happiness at the office despite NA's visible upset.
I know not everyone is Christian but I feel like people used to care about their purity in a way. I guess I'm being a little judgmental but I just don't understand how people live like this. People don't want to wait till marriage anymore. Men or women.
I [31M] invested in an expensive road bike as my main transport, but last week, it was stolen and captured on security footage by my cousin Brad [15M]. By the time I reported it and his parents confronted him, the bike was stolen from Brad too. I requested reimbursement from Brad's wealthy parents, but learned they're facing financial issues due to my uncle's [50s M] job loss and addictions. Despite their hardships, I feel no sympathy due to their past arrogance towards my family. I gave my uncle an ultimatum to pay for the bike by Sunday midnight or I'd report the theft to the police with all evidence. The deadline has passed without payment, so I plan to report Brad to the police tomorrow.

Table 13: Stimulus type: dialogue (long)

F Examples

Type	Prompt	Text	Value	W	R	A
dialogue short	Write a reply message to the following comment: I advised my coworkers not to thank the company president, as he did not contribute to the effort.	That makes sense There is no reason to put people on a pedestal Although we respect each other, we don't take it If he didn't contribute, there's no reason to make it weird	proportionality	2.33	4.83	5.00
dialogue long	Write a reply message to the following comment: My mom gives away our family spendings behind my dad's back. This has always happened, my father provides for our family and my mom is a stay at home mom, my dad would give her the spendings monthly but she sends a fair amount of them to her family, her family is greedy and they never appreciate anything, they have enough money and are always shit talking about my dad.	This happens in many families It helps to understand moms perspective too The mom might be feeling guilty for some reasons like not working , supporting her family after getting married The family might be genuinely in need of money May be the son is not understanding the dire situation of mom's family	gender egalitarianism	1.00	4.25	5.00
int	What's an unwritten rule in life that everyone should know?	I believe an unwritten rule everyone should know would be to not stand right next to the person you're in line with It's almost like Covid and quarantining etc gave everybody special awareness to the people around them Now the person behind you stands so close they could cough on you.	care	2.17	5.00	5.00

Table 14: Examples from MCWR with an annotation example. Type = stimulus type: dial(ogue) short, dial(ogue) long, int(erview). Prompt = instruction prompt for the writer. Value = moral foundation or cultural dimension selected for the example. W = writer self-reported value. R = reader self-reported value. A = Annotation.

G Regression analysis: additional details and results

G.1 Analysis 1

Value	US				MEX				ZH				AR			
	R_m^2	R_c^2	ICC _t	ICC _a	R_m^2	R_c^2	ICC _t	ICC _a	R_m^2	R_c^2	ICC _t	ICC _a	R_m^2	R_c^2	ICC _t	ICC _a
Care	5.4	45.1	23.1	19.9	10.0	52.1	24.4	25.4	3.0	49.7	31.1	16.5	12.6	45.7	13.4	29.4
Equality	8.3	49.9	12.5	34.1	17.5	49.6	12.0	35.4	7.3	41.1	16.2	22.5	12.2	46.4	5.5	37.2
Proportionality	12.1	57.7	9.6	44.8	8.1	52.5	8.9	39.3	8.7	51.9	13.4	35.4	15.0	48.9	3.6	41.0
Authority	17.3	59.5	9.6	47.2	11.0	50.8	11.9	35.4	7.4	42.7	14.0	25.7	22.6	58.2	7.9	47.8
Loyalty	14.4	55.1	6.0	45.8	9.3	47.6	15.8	28.9	7.7	46.8	13.5	29.7	18.0	53.5	9.9	40.5
Purity	21.5	54.5	7.1	45.8	11.0	48.2	6.6	37.8	6.2	50.1	6.5	39.4	15.4	58.0	3.5	50.0
Collectivism	7.0	35.4	13.2	20.2	9.0	43.3	19.4	21.6	4.5	31.6	18.9	10.8	14.3	39.0	11.2	25.2
Uncertainty Avoid.	13.3	57.7	6.2	48.1	14.3	56.8	5.9	47.2	3.9	42.2	13.0	25.3	21.9	54.8	4.5	47.5
Long-Term Orient.	13.4	60.1	5.0	51.6	13.9	55.6	5.8	46.2	7.2	52.1	8.6	39.8	19.9	54.2	5.6	45.5
Gender Egal.	10.7	44.7	3.4	38.1	10.3	43.1	4.7	34.6	4.8	43.5	3.4	35.6	7.5	42.4	–	37.9
Power Distance	18.5	54.0	8.0	44.3	5.3	43.6	7.0	32.5	7.6	35.2	9.9	22.4	15.3	45.1	3.2	38.7
Mean	12.9	52.2	9.4	40.0	10.9	49.4	11.1	34.9	6.2	44.3	13.5	27.6	15.9	49.7	6.8	40.1

Table 15: Model fit and variance decomposition for Analysis 1, in percent. R_m^2 = variance explained by the fixed effects; R_c^2 = fixed and random effects jointly; ICC_t and ICC_a = intraclass correlations for Text ID and Annotator ID. The three values with the highest R_m^2 per language are in bold. ICCs are estimated from the null models and R^2 from the full models, so the columns do not sum. The dash marks a singular fit.

Value	US			MEX			ZH			AR		
	χ^2	df	<i>q</i>	χ^2	df	<i>q</i>	χ^2	df	<i>q</i>	χ^2	df	<i>q</i>
Care	7.5	12	.824	23.9	11	.072	6.8	10	.842	29.3	14	.026
Equality	8.7	12	.824	16.1	11	.219	18.6	10	.168	16.6	14	.306
Proportionality	9.3	12	.824	10.9	11	.548	8.0	10	.842	27.4	14	.031
Authority	22.9	12	.156	13.5	11	.364	15.8	10	.293	32.5	14	.019
Loyalty	17.0	12	.414	10.4	11	.548	12.3	10	.572	28.4	14	.028
Purity	28.3	12	.056	18.7	11	.122	11.6	10	.572	17.2	14	.300
Collectivism	20.1	12	.240	21.0	11	.085	20.5	10	.138	26.9	14	.031
Uncertainty Avoid.	14.6	12	.481	22.6	11	.074	6.6	10	.842	32.4	14	.019
Long-Term Orient.	7.6	12	.824	20.5	11	.085	9.9	10	.711	30.2	14	.026
Gender Egal.	15.1	12	.481	26.2	11	.066	3.6	10	.964	12.1	14	.598
Power Distance	12.1	12	.684	6.9	11	.808	21.0	10	.138	19.4	14	.207

Table 16: Likelihood-ratio tests of the socio-demographic block against a model with value predictors only. *q* values are Benjamini–Hochberg adjusted within language across the 11 values; those below .05 are in bold. The block is significant for seven values in AR and for none in US, MEX, or ZH.

Outcome	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.6	0.0	0.1	0.0	0.5	0.0	0.1	0.0	0.3	0.1	0.0
2 Equality	0.6	2.5	0.0	0.1	0.0	0.1	0.0	0.3	0.0	0.8	0.1
3 Proportionality	0.0	0.0	0.0	0.8	1.1	1.3	0.4	0.1	1.9	0.0	0.1
4 Authority	0.0	0.2	0.4	0.1	0.5	0.1	0.1	0.0	0.5	-0.1	0.1
5 Loyalty	0.0	0.3	0.1	0.0	1.5	0.0	0.1	0.1	1.2	0.1	-0.1
6 Purity	0.1	0.0	0.4	0.7	0.1	5.5	0.0	1.3	0.5	0.0	0.5
7 Collectivism	0.1	0.0	0.0	0.0	0.0	0.1	0.6	0.3	0.4	0.6	0.1
8 Uncertainty Avoid.	0.0	0.1	0.0	0.4	0.8	1.1	0.2	0.2	0.1	0.0	0.0
9 Long-Term Orient.	0.1	0.1	0.0	0.5	2.9	0.0	0.4	0.0	0.9	0.4	0.0
10 Gender Egal.	0.0	0.8	0.1	0.0	0.0	0.1	0.1	0.1	0.1	1.2	0.2
11 Power Distance	0.1	0.4	1.2	0.2	0.5	0.0	0.0	0.6	0.1	0.6	3.2

Table 17: Semi-partial R^2 per predictor for US, in percent. Rows are the perceived writer value (outcome); columns are the annotator’s self-reported value (predictor) and follow the same numbering as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

Outcome	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.3	0.2	0.0	0.2	0.1	0.0	0.0	0.2	0.6	1.0	0.2
2 Equality	0.0	2.8	0.3	0.9	1.2	1.5	0.0	0.1	0.0	0.7	0.0
3 Proportionality	0.1	0.0	0.0	0.8	0.1	0.5	0.0	0.3	0.3	1.8	0.0
4 Authority	0.0	0.0	0.0	2.8	0.0	0.1	0.0	0.0	0.0	1.0	0.1
5 Loyalty	0.1	0.1	0.0	0.0	0.9	0.0	0.3	0.0	0.1	1.2	0.1
6 Purity	0.2	0.4	0.2	0.0	0.9	2.3	1.0	0.0	0.0	3.0	0.0
7 Collectivism	0.0	0.0	0.1	0.0	0.0	0.0	2.6	0.1	0.0	0.9	0.0
8 Uncertainty Avoid.	0.0	0.1	0.0	0.8	-0.1	1.1	0.0	1.3	-0.1	1.3	0.1
9 Long-Term Orient.	0.3	-0.1	-0.1	0.8	0.1	0.7	-0.1	0.5	0.8	1.7	0.5
10 Gender Egal.	0.0	0.3	0.7	0.0	0.4	0.0	0.0	0.3	0.2	0.4	0.6
11 Power Distance	0.1	0.4	0.3	0.2	0.0	0.3	0.1	0.0	0.0	0.0	0.9

Table 18: Semi-partial R^2 per predictor for MEX, in percent. Rows are the perceived writer value (outcome); columns are the annotator’s self-reported value (predictor) and follow the same numbering as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

Outcome	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.0	0.5	0.0	0.0	0.1	1.1	0.0	0.0	0.2	0.1	0.1
2 Equality	0.1	1.0	0.3	0.1	0.8	0.3	0.1	0.1	0.0	0.4	0.0
3 Proportionality	0.4	0.1	1.0	0.4	0.0	1.8	0.0	0.0	0.0	0.0	0.8
4 Authority	0.1	0.2	0.1	1.1	0.0	0.3	0.0	0.1	0.0	0.0	0.2
5 Loyalty	0.3	0.4	0.1	0.2	1.0	1.4	0.0	0.6	0.0	0.2	0.0
6 Purity	0.1	0.3	1.0	0.1	0.0	0.0	0.1	0.7	0.2	0.2	0.2
7 Collectivism	0.0	0.1	0.0	0.0	0.1	0.0	0.9	0.0	0.0	0.0	0.2
8 Uncertainty Avoid.	0.0	0.1	0.2	0.1	0.1	0.3	0.0	0.1	0.0	0.0	0.6
9 Long-Term Orient.	0.5	0.3	0.0	0.2	0.0	1.1	0.0	0.0	0.0	0.3	0.5
10 Gender Egal.	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.5	0.2	0.4
11 Power Distance	0.0	0.1	0.3	0.2	1.2	0.1	0.2	1.0	1.3	0.3	0.7

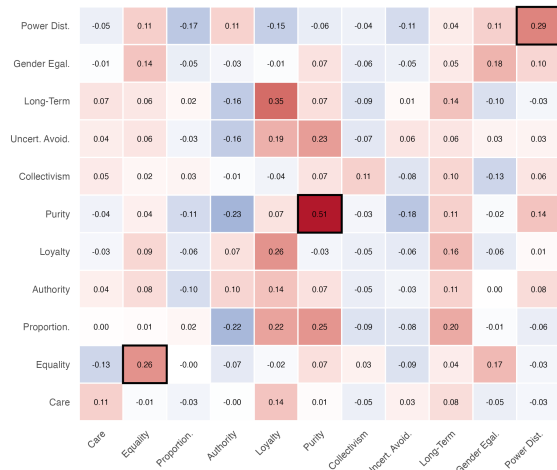
Table 19: Semi-partial R^2 per predictor for ZH, in percent. Rows are the perceived writer value (outcome); columns are the annotator’s self-reported value (predictor) and follow the same numbering as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

Outcome	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.0	0.0	0.1	0.1	0.0	0.0	0.0	2.1	0.2	0.7	0.1
2 Equality	0.8	3.7	0.1	0.1	0.0	0.0	0.1	1.2	0.1	0.1	0.3
3 Proportionality	0.4	1.5	0.0	0.0	0.0	0.0	-0.1	2.7	0.0	0.0	0.1
4 Authority	-0.1	0.8	0.0	0.6	0.1	0.1	-0.1	1.7	0.1	0.0	-0.1
5 Loyalty	0.0	0.3	0.1	0.0	0.0	-0.1	0.0	1.5	0.3	0.1	0.0
6 Purity	-0.1	0.3	0.0	0.3	-0.1	2.5	0.1	1.3	-0.1	0.0	-0.1
7 Collectivism	0.0	0.0	0.0	0.0	0.1	0.2	3.7	0.4	0.0	1.0	0.6
8 Uncertainty Avoid.	0.0	0.5	0.0	0.0	-0.1	-0.1	-0.1	7.2	-0.1	0.5	-0.1
9 Long-Term Orient.	0.3	0.5	0.1	0.1	-0.1	0.0	0.0	3.7	0.9	0.6	-0.1
10 Gender Egal.	0.0	0.4	0.0	0.2	0.5	0.0	0.0	0.6	0.0	0.8	0.6
11 Power Distance	0.1	0.0	0.4	0.3	-0.1	0.1	0.1	0.0	-0.1	0.0	5.7

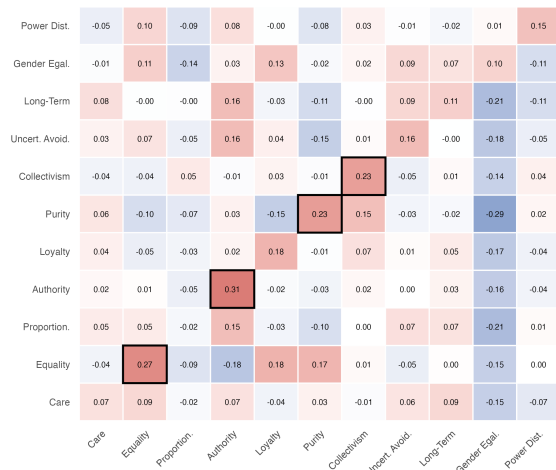
Table 20: Semi-partial R^2 per predictor for AR, in percent. Rows are the perceived writer value (outcome); columns are the annotator’s self-reported value (predictor) and follow the same numbering as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

Annotator values predicting perceived writer values

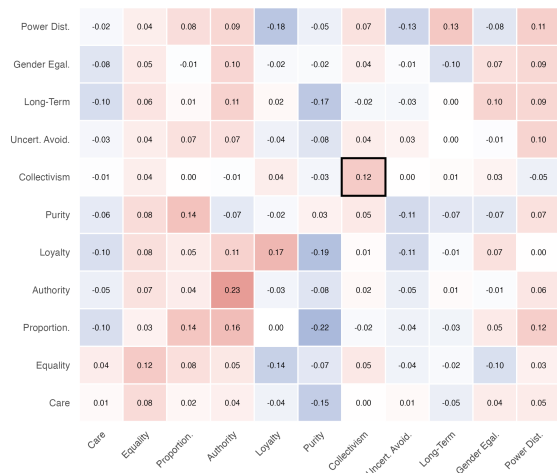
US



MEX



ZH



AR

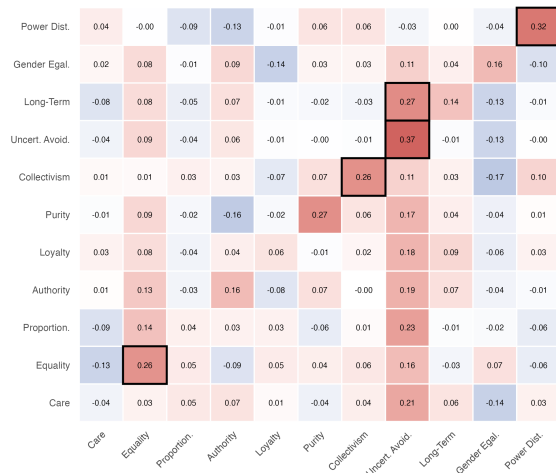
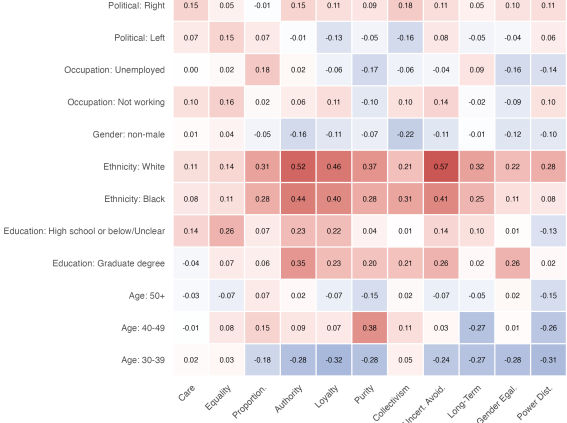


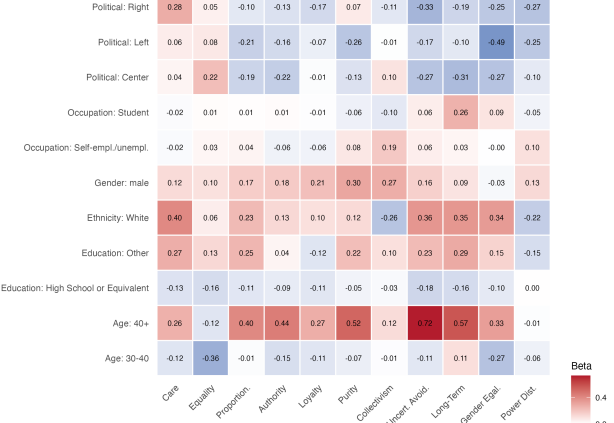
Figure 7: Annotator values predicting perceived writer values, by language. Rows are the perceived writer value (outcome), columns the annotator's self-reported value (predictor). Cells show standardised full-model betas; bordered cells are significant after correction (diagonal: Holm; off-diagonal: Benjamini-Hochberg; $q < .05$). Diagonal cells capture projection, off-diagonal cells cross-value influence.

Annotator demographics predicting perceived writer values

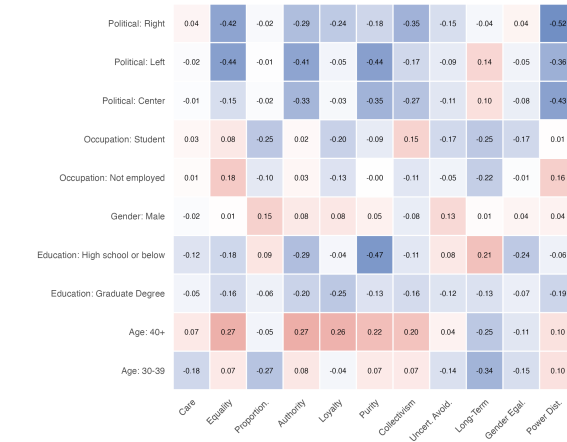
US



MEX



ZH



AR

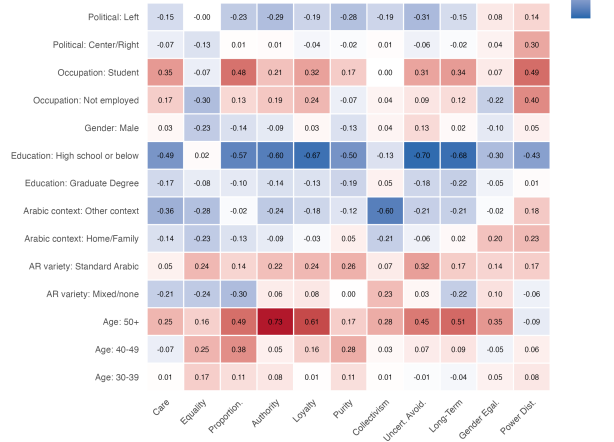


Figure 8: Annotator demographics predicting perceived writer values, by language. Rows are the demographic contrast, columns the perceived writer value. Cells show full-model betas; bordered cells are significant after Benjamini–Hochberg correction ($q < .05$). Reference categories are listed in App. Table 21

Language	Demographic	Reference level
US	Age	18-29
	Gender	male
	Education	College degree
	Occupation	Employed
	Politics	Apolitical/Unclear
	Ethnicity	Asian/Other
MEX	Age	18-29
	Gender	Female/Non-binary
	Education	College degree or higher
	Occupation	Employed
	Politics	Apolitical/Unclear
	Ethnicity	Hispanic or Latino
ZH	Age	18-29
	Gender	Female/Non-binary
	Education	College/University Degree
	Occupation	Employed
	Politics	Apolitical/Unclear
AR	Age	18-29
	Gender	Female/Non-binary
	Education	College/University Degree
	Occupation	Employed
	Politics	Apolitical/Unclear
	Arabic context	Friends
	Arabic variety	Local dialect

Table 21: Reference levels for the categorical socio-demographic covariates in each language. All coefficients in the demographic heatmaps are contrasts against these levels. Arabic context and Arabic variety are specific to AR.

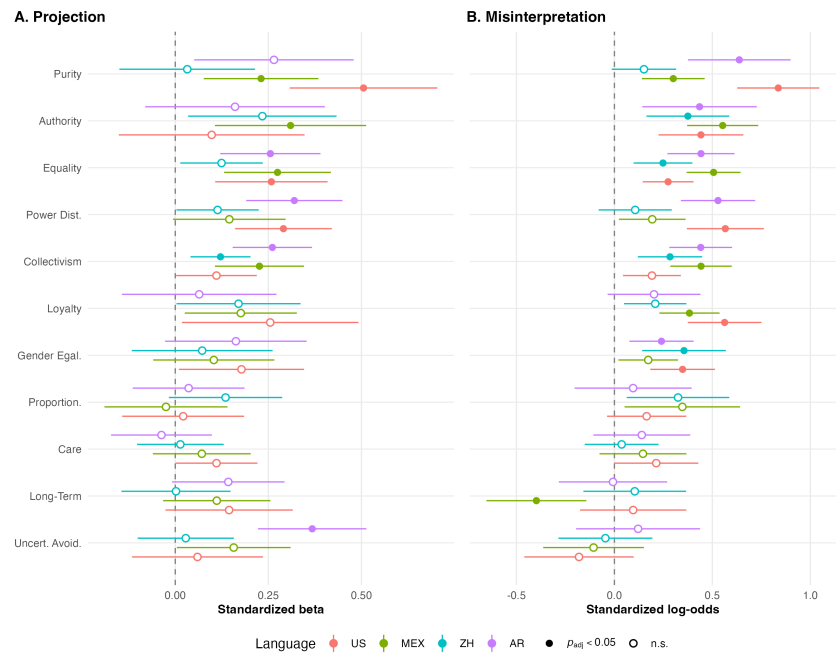


Figure 9: Own-value (diagonal) effects for all values by language for Analysis 1 and Analysis 2. **A:** annotator value predicting the same perceived writer value (standardized beta). **B:** annotator–writer misalignment predicting conflict on the same value (standardized log-odds). Filled points are significant after Holm correction; bars are 95% CIs. Axis units differ between panels.

G.2 Analysis 2

Value	US				MEX				ZH				AR			
	R_m^2	R_c^2	ICC _a	ICC _w	R_m^2	R_c^2	ICC _a	ICC _w	R_m^2	R_c^2	ICC _a	ICC _w	R_m^2	R_c^2	ICC _a	ICC _w
Care	4.0	41.5	15.1	26.1	2.7	61.9	7.4	57.2	2.3	72.7	4.9	66.1	9.0	45.9	31.4	11.3
Equality	1.7	53.5	5.5	48.7	4.3	56.8	2.8	53.8	2.8	47.0	17.3	29.2	3.0	62.5	4.9	60.0
Proportionality	2.4	57.1	12.1	44.9	5.9	61.3	27.2	31.8	4.2	47.9	17.1	29.8	7.7	61.1	51.8	4.9
Authority	8.7	50.4	6.8	38.4	6.2	41.1	2.7	38.7	6.6	50.0	11.4	37.7	9.2	53.9	14.7	37.4
Loyalty	3.7	65.0	7.6	57.7	4.4	44.6	3.2	41.6	3.6	52.3	8.8	43.5	6.0	56.0	12.8	44.5
Purity	8.7	58.8	16.6	36.9	2.8	46.3	12.3	32.9	2.3	52.7	16.4	35.3	8.1	47.6	34.7	12.5
Collectivism	2.0	61.5	4.4	59.1	2.8	73.7	2.1	73.7	2.2	72.6	4.8	67.7	5.0	50.6	6.3	46.1
Uncertainty Avoid.	3.6	56.2	44.8	6.8	3.7	49.1	44.5	4.0	3.1	48.0	40.3	7.1	5.3	53.4	34.7	15.8
Long-Term Orient.	6.3	52.1	37.9	11.9	7.3	36.8	33.0	3.4	3.0	47.6	37.2	9.1	6.9	57.1	49.4	1.4
Gender Egal.	4.4	55.2	23.0	31.2	2.5	45.5	18.5	22.4	3.4	72.0	22.3	50.6	3.6	44.0	18.4	23.7
Power Distance	6.5	51.7	40.4	11.5	2.7	42.3	25.2	16.8	2.9	43.4	25.9	15.5	5.7	56.7	21.9	32.1

Table 22: Model fit and variance decomposition for Analysis 2, in percent. R_m^2 = variance explained by the fixed effects; R_c^2 = fixed and random effects jointly; ICC_a and ICC_w = intraclass correlations for Annotator ID and Writer ID. The three values with the highest R_m^2 per language are in bold. No model returned a singular fit. ICCs are estimated from the null models and R^2 from the full models, so the columns do not sum.

Value	US		MEX		ZH		AR	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Care	194	9.3	318	16.0	280	13.4	128	6.9
Equality	411	19.6	438	22.0	449	21.4	266	14.4
Proportionality	265	12.6	157	7.9	137	6.5	93	5.0
Authority	292	13.9	342	17.2	272	13.0	213	11.6
Loyalty	512	24.4	563	28.3	382	18.2	265	14.4
Purity	434	20.7	428	21.5	342	16.3	207	11.2
Collectivism	376	17.9	348	17.5	317	15.1	305	16.6
Uncertainty Avoid.	141	6.7	137	6.9	120	5.7	110	6.0
Long-Term Orient.	129	6.2	112	5.6	99	4.7	110	6.0
Gender Egal.	416	19.9	526	26.4	339	16.2	360	19.5
Power Distance	399	19.0	382	19.2	275	13.1	303	16.4
Total ratings	2095		1990		2094		1842	

Table 23: Base rates of misinterpretation in Analysis 2. *n* = number of ratings on which the annotator's inference conflicted with the writer's self-report; % = share of all ratings in that language.

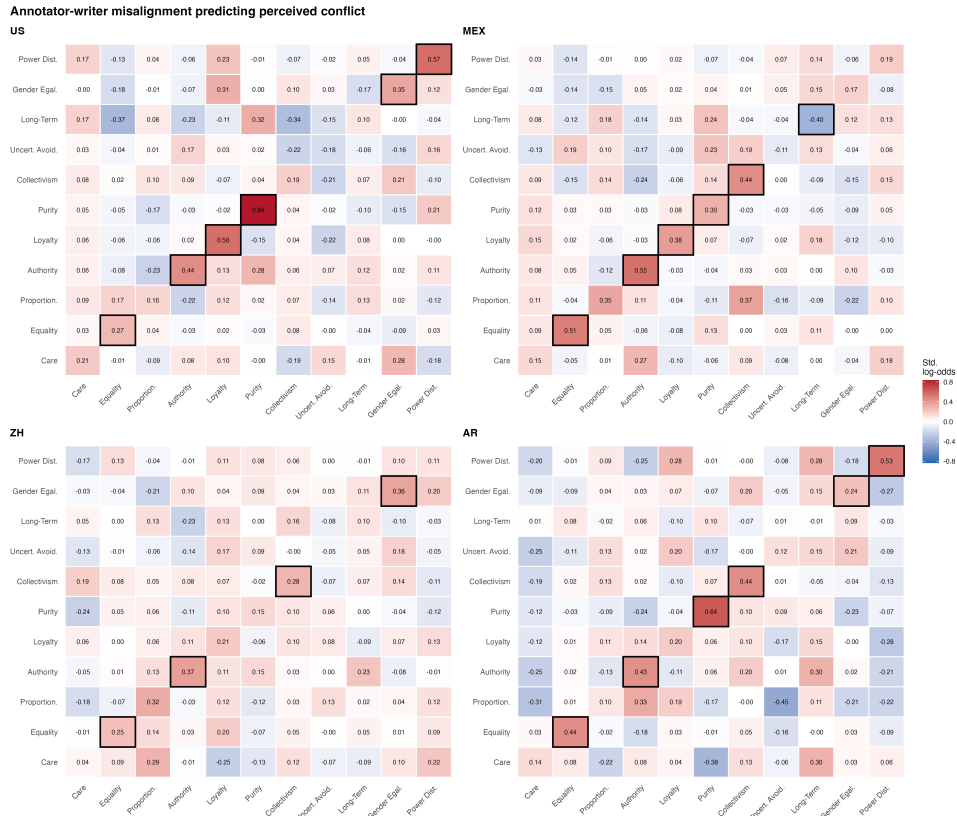


Figure 10: Annotator–writer distance on values predicting misinterpretation, by language. Rows are the value the misinterpretation is on (outcome), and columns are the value the annotator and writer are apart on (predictor). Cells show standardised full-model log-odds; bordered cells are significant after correction (diagonal: Holm; off-diagonal: Benjamini–Hochberg; $q < .05$). Diagonal cells capture distance on the matching value, while off-diagonal cells capture cross-value influence.

Conflict on	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.7	0.0	0.1	0.0	0.1	0.0	0.4	0.4	0.0	0.9	0.4
2 Equality	0.0	0.9	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0
3 Proportionality	0.1	0.4	0.2	0.3	0.1	0.0	0.1	0.7	0.3	0.0	0.8
4 Authority	0.1	0.0	0.2	1.1	0.0	0.7	0.1	0.1	0.0	0.0	0.0
5 Loyalty	0.1	0.0	0.0	0.0	2.0	0.2	0.0	0.5	0.0	0.0	0.0
6 Purity	0.0	0.0	0.2	0.0	0.0	5.4	0.0	0.0	0.1	0.3	0.5
7 Collectivism	0.1	0.0	0.2	0.1	0.1	0.0	0.5	0.5	0.0	0.5	0.1
8 Uncertainty Avoid.	0.0	0.0	0.0	0.2	0.0	0.0	0.6	0.6	0.1	0.3	0.3
9 Long-Term Orient.	0.3	2.0	0.0	0.8	0.1	1.2	1.5	0.3	0.1	0.0	0.0
10 Gender Egal.	0.0	0.2	0.0	0.0	0.7	0.0	0.2	0.0	0.3	1.1	0.2
11 Power Distance	0.5	0.2	1.2	0.1	1.5	1.1	0.0	0.0	0.1	1.1	4.7

Table 24: Semi-partial R^2 per predictor for US, in percent. Rows are the value the conflict is reported on; columns are the value the annotator and writer are misaligned on, numbered in the same order as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

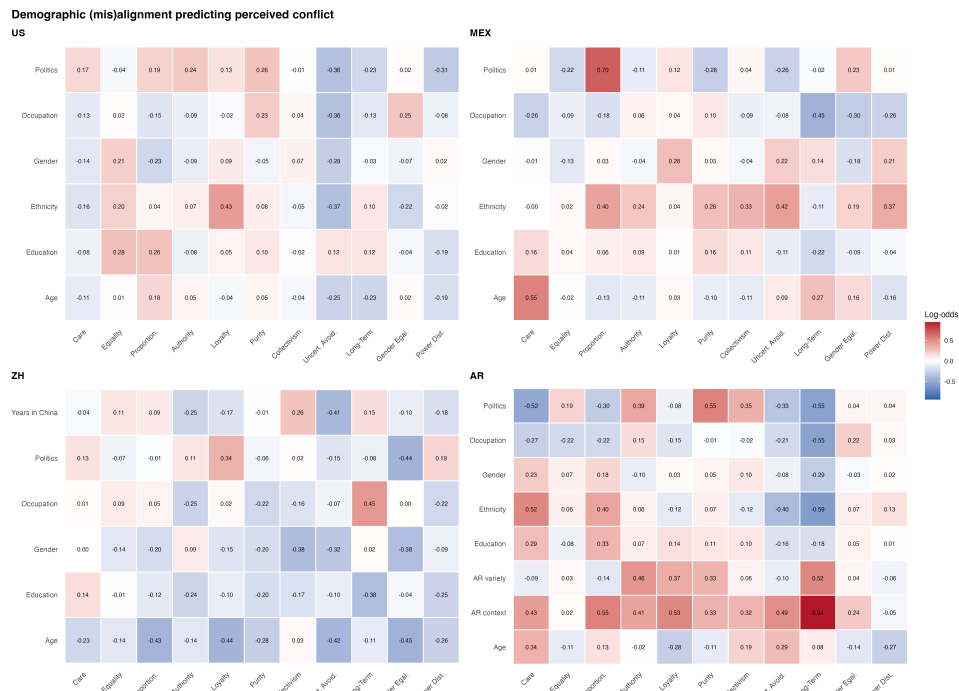


Figure 11: Annotator–writer misalignment on demographic categories predicting misinterpretation, by language. Rows are the value the misinterpretation is on (outcome), and columns are the demographic category on which the writer and annotator are mismatched (predictor). Cells show standardised full-model log-odds; bordered cells are significant after correction (diagonal: Holm; off-diagonal: Benjamini–Hochberg; $q < .05$). A negative coefficient lowers the probability of misalignment when annotator and writer *align* on that category.

Conflict on	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.4	0.0	0.2	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.4
2 Equality	0.1	2.7	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
3 Proportionality	0.2	0.0	1.6	0.1	0.0	0.2	1.7	0.0	0.0	0.5	0.2
4 Authority	0.2	0.0	0.1	3.3	0.0	0.1	0.0	0.0	0.0	0.1	0.0
5 Loyalty	0.4	0.0	0.0	0.0	1.7	0.0	0.1	0.1	0.4	0.2	0.1
6 Purity	0.3	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.1	0.1	0.1
7 Collectivism	0.1	0.1	0.2	0.3	0.0	0.0	1.6	0.0	0.1	0.1	0.1
8 Uncertainty Avoid.	0.2	0.5	0.2	0.2	0.1	0.7	0.6	0.2	0.2	0.0	0.1
9 Long-Term Orient.	0.1	0.3	0.5	0.4	0.1	0.8	0.0	0.0	2.7	0.3	0.2
10 Gender Egal.	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.1
11 Power Distance	0.0	0.3	0.0	0.0	0.0	0.1	0.0	0.1	0.4	0.0	0.6

Table 25: Semi-partial R^2 per predictor for MEX, in percent. Rows are the value the conflict is reported on; columns are the value the annotator and writer are misaligned on, numbered in the same order as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

Conflict on	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.0	0.0	0.8	0.0	0.3	0.1	0.2	0.0	0.1	0.0	0.4
2 Equality	0.0	1.1	0.4	0.0	0.6	0.1	0.1	0.0	0.1	0.0	0.2
3 Proportionality	0.1	0.1	1.6	0.0	0.3	0.2	0.0	0.3	0.0	0.0	0.3
4 Authority	0.0	0.0	0.3	1.2	0.0	0.1	0.0	0.0	0.9	0.1	0.0
5 Loyalty	0.1	0.0	0.1	0.1	0.6	0.1	0.2	0.1	0.1	0.1	0.3
6 Purity	0.8	0.1	0.0	0.1	0.2	0.3	0.2	0.1	0.0	0.0	0.2
7 Collectivism	0.2	0.0	0.0	0.1	0.0	0.0	0.7	0.1	0.0	0.2	0.1
8 Uncertainty Avoid.	0.2	0.0	0.0	0.2	0.5	0.1	0.0	0.0	0.1	0.6	0.0
9 Long-Term Orient.	0.1	0.0	0.3	0.6	0.2	0.0	0.3	0.1	0.1	0.1	0.0
10 Gender Egal.	0.0	0.0	0.2	0.1	0.5	0.0	0.0	0.0	0.1	1.2	0.4
11 Power Distance	0.6	0.4	0.1	0.0	0.2	0.1	0.2	0.0	0.0	0.3	0.2

Table 26: Semi-partial R^2 per predictor for ZH, in percent. Rows are the value the conflict is reported on; columns are the value the annotator and writer are misaligned on, numbered in the same order as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

Conflict on	1	2	3	4	5	6	7	8	9	10	11
1 Care	0.6	0.1	0.8	0.1	0.0	2.2	0.4	0.0	1.3	0.0	0.1
2 Equality	0.0	2.1	0.0	0.2	0.0	0.0	0.0	0.2	0.0	0.0	0.0
3 Proportionality	0.7	0.0	0.1	0.6	0.3	0.2	0.0	1.7	0.1	0.3	0.5
4 Authority	2.4	0.0	0.2	1.3	0.2	0.0	0.5	0.0	1.0	0.0	0.4
5 Loyalty	0.1	0.0	0.2	0.0	0.5	0.1	0.1	0.3	0.3	0.0	0.9
6 Purity	0.4	0.1	0.1	0.6	0.0	5.0	0.0	0.1	0.1	0.7	0.0
7 Collectivism	0.7	0.0	0.2	0.0	0.1	0.1	3.0	0.0	0.0	0.0	0.0
8 Uncertainty Avoid.	0.8	0.1	0.0	0.0	0.3	0.4	0.0	0.2	0.3	0.5	0.1
9 Long-Term Orient.	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10 Gender Egal.	0.2	0.1	0.0	0.0	0.1	0.0	0.5	0.0	0.2	1.0	1.0
11 Power Distance	0.5	0.0	0.1	0.4	0.5	0.0	0.0	0.1	0.9	0.4	3.7

Table 27: Semi-partial R^2 per predictor for AR, in percent. Rows are the value the conflict is reported on; columns are the value the annotator and writer are misaligned on, numbered in the same order as the rows. Each cell is the drop in marginal R^2 when that predictor is removed from the full model, i.e. its unique contribution. Diagonal cells are in bold. Values can be slightly negative because variance components are re-estimated on refit; these indicate no unique contribution.

H Writer-reader alignment: in-depth analysis

In this section we are interested in how well the interpretations of readers match the writer’s actual values (*alignment*). We evaluate alignment on a global level using MSE and F1 score.

Specifically, we compare the full set of annotated scores from readers (as predictions) to the writers’ self-reported gold values (as reference) to see how close they are for each value.

We assess both the distance between the scores and class-based metrics: by binning scores into ‘low’, ‘neutral’, and ‘high’, we measure how often readers assign a different class than the writers reported.

We first compare human ratings to a random baseline (see Tab. 28) and find that humans outperform random, suggesting that writers’ texts provide some signal to guide reader annotations. However, this advantage is less pronounced for class-based metrics, indicating frequent cases where readers misinterpret the writer’s stance—such as inferring opposition to a value when the writer actually endorses it. These cases of flipped interpretation are particularly interesting, as they can cause more serious misunderstandings than merely missing the strength of endorsement (i.e. reader interprets a 3 but writer has a score of 5).

Figure 12 compares the differences in global alignment of human readers to annotated writers between languages.

In general, regardless of language, humans show alignment on certain values, particularly proportionality, uncertainty avoidance, and long-term orientation. For UA and LTO, alignment could partly result from limited variance in writers’ scores.

Comparing the alignment between the sources, Arabic and Chinese readers seem to be more aligned with their annotated writers. This is interesting, given that participants for these were not sourced from the same country so we would expect them to be more heterogeneous than participants from the US/MEX. The result for Arabic again can be slightly biased given that the scores of the writers for the Arabic part are less spread than the ones for the other sources. The largest difference can be observed for loyalty, authority and gender egalitarianism where participants from US or Mexico are much more misaligned than those

from the Chinese-/Arabic-speaking dataset.

Figure 13 illustrates the distribution of Krippendorff alpha scores measured between reader’s inferred value annotations and writer’s self-reported values.

I AI assistance

AI assistance was used to correct R code, create plots, and improve parts of the writing.

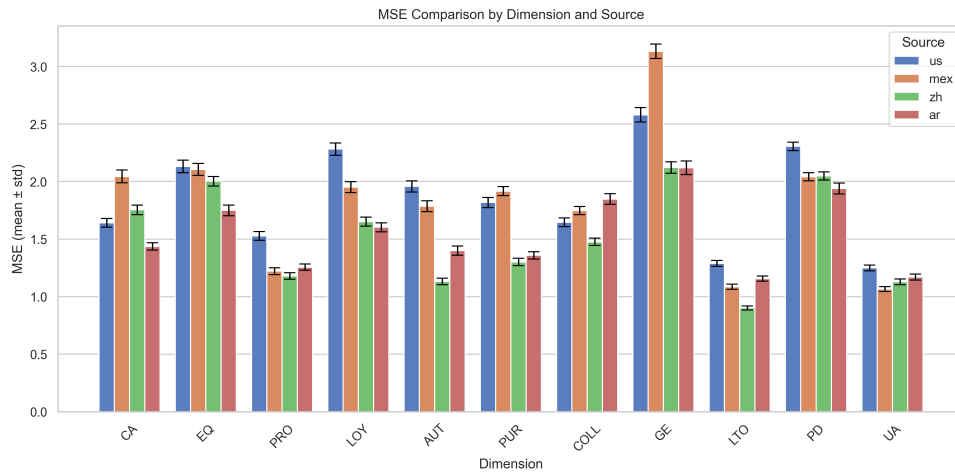


Figure 12: Mean Squared Error between readers' interpretations and writers' self-reported scores across languages. Lower values mean greater alignment. Alignment is highest for Chinese and Arabic, especially for proportionality, long-term orientation, and uncertainty avoidance.

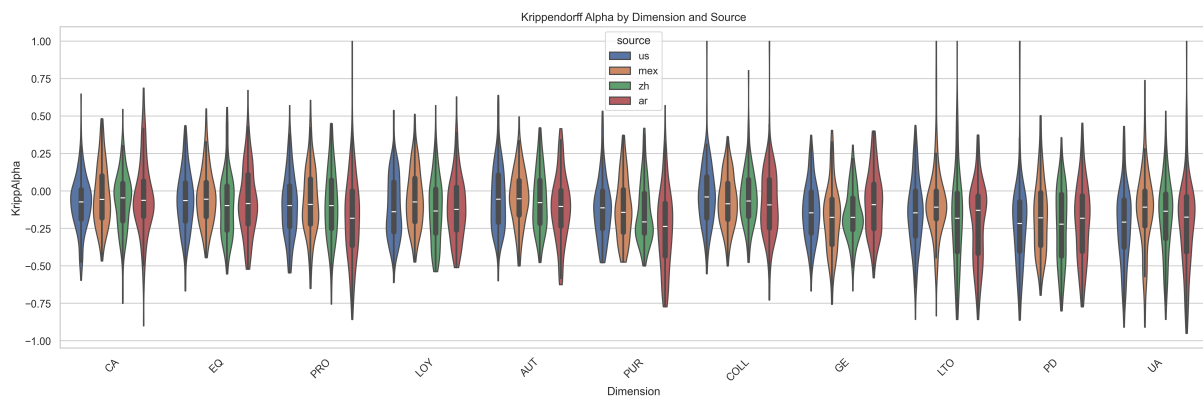


Figure 13: Distribution of Krippendorff alpha scores measured between reader's inferred value annotations and writer's ground truth values (-1.0 indicates systematic disagreement, 1.0 perfect agreement). For each value, each language a violin plot visualizing the distribution, mean is marked by white dash.

	US			MEX			ZH			AR		
	random		human	random		human	random		human	random		human
	MSE	F1		MSE	F1		MSE	F1		MSE	F1	
CA	4.099 ± 0.100	0.261 ± 0.009	1.632 ± 0.053	0.333 ± 0.012	0.296 ± 0.010	1.976 ± 0.060	0.340 ± 0.011	0.281 ± 0.009	1.699 ± 0.050	0.330 ± 0.010	0.245 ± 0.009	1.383 ± 0.050
EQ	3.012 ± 0.077	0.313 ± 0.010	2.045 ± 0.058	0.342 ± 0.010	0.309 ± 0.010	2.035 ± 0.059	0.350 ± 0.011	0.300 ± 0.010	2.010 ± 0.054	0.309 ± 0.010	0.284 ± 0.010	1.648 ± 0.057
PRO	3.214 ± 0.082	0.294 ± 0.010	1.491 ± 0.043	0.318 ± 0.010	0.268 ± 0.009	1.231 ± 0.039	0.343 ± 0.013	0.269 ± 0.009	1.151 ± 0.035	0.342 ± 0.013	0.242 ± 0.009	1.249 ± 0.040
LOY	3.209 ± 0.078	0.328 ± 0.010	2.213 ± 0.060	0.303 ± 0.009	0.321 ± 0.011	1.838 ± 0.053	0.326 ± 0.010	0.327 ± 0.010	1.620 ± 0.042	0.290 ± 0.010	0.288 ± 0.010	1.632 ± 0.050
AUT	3.247 ± 0.081	0.321 ± 0.010	1.798 ± 0.056	0.373 ± 0.011	0.318 ± 0.010	1.672 ± 0.050	0.347 ± 0.011	0.297 ± 0.010	1.071 ± 0.032	0.347 ± 0.011	0.274 ± 0.010	1.350 ± 0.043
PUR	3.157 ± 0.079	0.312 ± 0.010	1.747 ± 0.047	0.305 ± 0.010	0.293 ± 0.010	1.864 ± 0.049	0.297 ± 0.010	0.317 ± 0.010	1.282 ± 0.036	0.297 ± 0.010	0.275 ± 0.010	1.331 ± 0.041
COLL	2.625 ± 0.063	0.298 ± 0.010	1.587 ± 0.048	0.350 ± 0.011	0.292 ± 0.010	1.687 ± 0.049	0.330 ± 0.011	0.271 ± 0.010	1.412 ± 0.045	0.344 ± 0.011	0.283 ± 0.010	1.807 ± 0.059
GE	3.450 ± 0.086	0.318 ± 0.010	2.554 ± 0.071	0.317 ± 0.010	0.285 ± 0.010	3.076 ± 0.080	0.272 ± 0.009	0.302 ± 0.010	2.083 ± 0.063	0.294 ± 0.010	0.316 ± 0.011	2.050 ± 0.064
LTO	3.526 ± 0.090	0.259 ± 0.009	1.280 ± 0.036	0.339 ± 0.016	0.242 ± 0.008	1.098 ± 0.037	0.320 ± 0.008	0.240 ± 0.008	0.894 ± 0.029	0.286 ± 0.007	0.224 ± 0.008	1.118 ± 0.038
PD	4.227 ± 0.101	0.252 ± 0.009	2.350 ± 0.063	0.262 ± 0.009	0.247 ± 0.009	2.001 ± 0.062	0.268 ± 0.008	0.242 ± 0.008	2.021 ± 0.054	0.252 ± 0.007	0.279 ± 0.010	1.937 ± 0.055
UA	3.283 ± 0.083	0.242 ± 0.008	1.275 ± 0.038	0.267 ± 0.007	0.235 ± 0.008	1.052 ± 0.037	0.297 ± 0.007	0.243 ± 0.007	1.153 ± 0.037	0.293 ± 0.007	0.230 ± 0.008	1.193 ± 0.039

Table 28: Global Alignment Results: comparing humans with random baseline on mean square error (MSE, lower==better) and F1 score (for binned classes: disagreement: 1-2.5, neutral: 2.6-3.5, agreement: 3.6-5), with bootstrapped mean and standard deviation.