

People Are Not Just Their Countries. Disentangling Social Determinants of LLM Value Alignment Across Europe

Maria-Louisa Wightman, Guillaume Bied, Tijl De Bie

AIDA-IDLab, Department of Electronics and Information Systems, Ghent University
MariaLouisa.Wightman@UGent.be, Guillaume.Bied@UGent.be, Tijl.DeBie@UGent.be

Abstract

As Large Language Models (LLMs) are increasingly used as a primary source of information and advice, understanding their alignment to humans in terms of values becomes a pressing concern. A growing literature has leveraged large scale surveys to investigate to what extent LLMs' and humans' stated values and opinions align. With limited exceptions, studied populations have been defined by country borders or cultural bounds. Yet, this focus neglects the role that socio-demographic divides may play for value alignment disparities.

Relying on the European Social Survey, we address this knowledge gap by considering value alignment displayed with respect to 10 prominent commercial LLMs in terms of 15 socio-demographic variables as well as country of residence. Our analyses reveal that LLMs are indeed unequally aligned to the values of different socio-demographic groups, notably those defined by education, income, occupation and religion. When examining alignment at the individual level, a respondent's country, taken as a stand-alone variable, explains a substantial amount of variation that is on par with the full set of considered socio-demographics. Further disentangling the respective role of country-level and socio-demographic factors, we find they are complementary in explaining value alignment patterns, with their relative weights varying across the subset of questions considered.

Code & Supplementary Material —

<https://github.com/aida-ugent/LLMs-x-ESS>

Introduction

Technologies are not separately produced but rather co-produced with the societies they are positioned in, inherently reproducing and establishing structure and authority of different kinds (Jasanoff 2004). This creates socio-technical systems in which Artificial Intelligence (AI), and more specifically Large Language Models (LLMs), take on special roles: When humans interact with AI, individuals' experiences are shaped by behavioural mechanisms, such as anthropomorphism, the response to sycophancy, and cognitive offloading. This can lead to increased trust, dependence and heightened persuasive abilities of the LLMs (Sun and

Wang 2026; Xu et al. 2025; Jose et al. 2025). In a larger socio-technical system this has implications beyond the personal; as LLMs are used directly as sources for information and advice, the knowledges and value systems they contain carry over into assigned tasks. Additionally, LLMs are integrated into systems such as content moderation or information retrieval, where their involvement remains opaque to users. Together this creates unique challenges in which AI can not be treated as a mere engineering system with glitches and bugs that need to be resolved (Torkamaan et al. 2024; Kroes et al. 2006). Moreover, it is often argued that AI is not representative, but rather shaped at all stages by people and organisational entities that mirror at least some aspects underlying WEIRD (Western, Educated, Industrialized, Rich, Democratic) power structures (Mihalcea et al. 2025). This emphasises the need for (value) sensitive design of LLMs.

To be able to better reflect a diverse and heterogeneous population, understanding *who* generative models, align with is the necessary first step. Many studies have assessed alignment across various value related dimensions by leveraging large scale surveys and comparing responses of humans and LLMs (e.g. Durmus et al. 2024). Besides Santurkar et al. (2023), who investigate socio-demographics within a single country, the United States (U.S.), these studies have largely been focused on alignment with respect to countries. And while understanding global dynamics is important, other aspects of alignment have been neglected. Failing to recognise that conflating the opinions of a diverse set of people who make up a country, could lead to blind spots related to the role of social stratifiers past nationality within the assessment of LLMs alignment disparities.

Sen et al. (2025) argued that demographic representativeness of LLMs is in general still insufficiently well researched. This is particularly true for value alignment research, which has focused predominantly on cultural and national differences, leaving individual-level socio-demographic variation largely unexamined. In the present paper, we address this research gap by asking: **What patterns of alignment differences exist across socio-demographic factors in European countries?** We further seek to disentangle the contributions of country of residence and socio-demographics to value alignment. This induces the second research question: **To what extent is alignment driven by cross-national differences compared to indi-**

vidual level socio-demographics?

Leveraging the European Social Survey (ESS), this paper addresses these open research questions by studying LLM alignment with respect to 15 socio-demographic factors and country of residence. We prompt 10 popular LLMs repeatedly with selected value-related survey questions, and calculate alignment scores by comparing LLMs' answers to those of the survey respondents. The ESS poses an alternative to the surveys that have commonly been used for value alignment research, especially the World Value Survey (WVS). Their repeated use as a benchmark raises concerns about generalizability and contamination as the historical footprint of the WVS makes it highly likely that current LLMs have already memorized its questions and results, as well as reporting on its use for alignment evaluation.

To answer the research questions posed, we investigate disparities in alignment scores through several means. First, we report aggregated alignment scores by socio-demographic group and by country, enabling us to identify patterns of comparative alignments across groups. Second, we use inverse propensity weighting to investigate whether country-level differences are driven by different socio-demographic compositions. Third, we predict individual alignment scores using linear regressions and boosted trees to investigate the share of variance explained by socio-demographics, countries, and their combination, and to assess the potential role of interactions.

Our key contributions are: **i)** the first cross-national evaluation of value alignment scores with respect to socio-demographics, using the ESS, which provides an alternative to the widely used WVS; **ii)** evidence that LLMs are unequally aligned across socio-demographic groups and countries, reproducing global patterns of alignment to WEIRD countries, such that values and opinions of socio-demographic groups that are richer, more educated, from more Western countries in Europe are better represented by the LLMs; **iii)** we show that country of residence is important for understanding value alignment disparities, and that between-country differences cannot be explained by our considered set of socio-demographics alone; yet, country of residence and socio-demographics are complementary in explaining alignment, with their combination providing the highest explanatory power by a substantial margin.

Related Literature

Alignment Evaluation Studies looking at alignment of LLMs do so from varied understandings of the concept, ranging from the alignment to human preferences rooted in the post-training practice of Reinforcement Learning with Human Feedback (RLHF), to others who study alignment by defining it as the representativeness or agreement of LLMs with humans. This has been investigated in terms of moral decisions, voting behaviour, political orientation, and the attitudes, values, and opinions exhibited by LLMs. The two approaches to alignment are intimately related with each other because, as Xiao et al. (2025) note, RLHF is one of the mechanisms behind the biases in LLMs.

Some steps to address these biases in the context of alignment have been taken. Liemt et al. (2026) surveyed people

about their expectations of values and cultural representativeness of LLMs and Kirk et al. (2024b) collected a dataset on the preferences of a geographically and demographically diverse set of participants for LLM interactions. An emerging notion at the forefront of alignment research is 'pluralistic alignment', which aims to not only take multiple perspectives into account, but build systems that cater to various requirements (see Shetty et al. (2026) for a survey). But before systems that represent heterogeneous and diverse populations can be created, it is crucial to understand within which populations and along which social axes misalignment is most prominent.

To evaluate alignment with respect to value-laden stances, surveys have been widely utilized. These studies fall into two groups: The first group evaluate LLM alignment directly against human responses (Cao et al. 2023; Santurkar et al. 2023; Durmus et al. 2024; Liu, Kaneko, and Chu 2026). The other group situate the LLMs on pre-defined political dimensions (Wright et al. 2024; Nadeem et al. 2026) or within value frameworks (Tao et al. 2024; Sukiennik et al. 2025). Regardless of the vocabulary used to describe the alignment being measured, most of these studies rely on a small set of surveys building on value frameworks by Hofstede (Hofstede 1980), Schwartz (Schwartz 1992) and Inglehart and Welzel (Inglehart and Welzel 2005).

The most prominently used survey is the WVS. However, its repeated use as a benchmark for explicit value alignment raises concerns about the generalizability of findings. Moreover, the wave used across mentioned studies was collected between 2017 and 2022, with partial data collection occurring during the Covid-19 pandemic. Additionally, given the WVS's decades-long history, LLMs are likely to have encountered both the survey items and published results, raising further concerns about data contamination.

Another aspect that cross-national studies on LLM value alignment have in common, is their sole focus on contrasting across nation-states or cultures. Even though surveys often contain rich information on socio-demographics of participants, this is never made a focal point. Only Santurkar et al. (2023) investigate alignment with respect to demographics in the context of the United States. But as the specific US context, and models used, might not generalize towards other contexts, cross-country analyses that include socio-demographics remain absent.

This does not mean researchers studying broad notions of alignment have not been interested in socio-demographic groups. However, the studies that do so are either researching voting behaviour (e.g. Batzner et al. 2025 Von Der Heyde, Haensch, and Wenz 2025), or they prompt LLMs with a persona enriched with demographics, without empirically studying the actual alignment to those same demographics (AlKhamissi et al. 2024; Batzner et al. 2025; Durmus et al. 2024; Wright et al. 2024; Liu, Kaneko, and Chu 2026; Ma et al. 2025; Sukiennik et al. 2025). Some do calculate an empirical alignment for demographic groups, not to investigate comparative patterns, but only for use as a baseline to evaluate steering techniques against (Williams et al. 2026; Lin et al. 2026).

Another set of papers aimed at simulating human sur-

vey responses, rather than evaluating models in terms of alignment, use demographics similarly to the persona based alignment evaluations (Park et al. 2024, Abeliuk, Gaete, and Bro 2025, Ma et al. 2025).

In summary, there is a broad and growing interest in demographics in alignment research. Even so, a concrete evaluation of what drives disparate alignment scores of individuals when considering both country of residence and socio-demographics is still lacking.

Value Formation in Survey Research Much of the above-mentioned literature on alignment evaluation use long standing opinion and value surveys. Building on these surveys, there is broad literature on the global and local structures of value systems and value formation. There are three interconnected debates relevant to the use of value surveys by the community of alignment researchers.

First, scholars disagree on whether “national culture” is the primary driver of value differences. Fischer and Schwartz (2011) argue that values vary much more within countries than between them, and Greenfield (2014) posits that high within-country variability simply reflects the adaptation of values to local socio-demographic conditions. In contrast, Akaliyski et al. (2021) defend nations as powerful cultural units that individuals organize around, arguing that the effect is stronger than sub-national demographic differences or globally shared religions.

A second debate centres on how these macro-level national contexts interact with individual demographics. For instance, Miles and Yeh (2022) show that the effects of demographic variables on personal values are not universal, but vary significantly across national contexts. Vilar, Liu, and Gouveia (2020), however, find that culture has little to no moderating effect on the relationship between age or gender and personal values.

Finally, a methodological debate centres on measurement invariance, i.e. whether survey instruments can accurately compare values across diverse cultures. Alemán and Woods (2016) caution that WVS value orientations lack the configural and metric invariance needed for meaningful cross-national comparison, outside of advanced post-industrial democracies. This has implications for their use for assessing value alignment. For the reduced Portrait Value Questionnaire (PVQ) in the ESS Davidov, Schmidt, and Schwartz (2008) and Bilsky, Janik, and Schwartz (2011) find stronger cross-national validity and metric invariance.

Methodology

This section proceeds as follows. We first introduce the European Social Survey (ESS) that will be used for our analyses. After detailing questions selection and the prompting setup, we will discuss the response patterns of considered LLMs. We then give our operationalization of an alignment score and lastly explain our analytical strategy.

Survey Data and Prompting Strategy

The European Social Survey The following analysis is based on the 11th wave of the European Social Survey¹ carried out between 2023 and 2024 in 29 European countries and Israel. The ESS is a long-running survey that started in 2002, featuring both fixed and rotating question modules on various topics from climate change and energy to institutional and social trust. It aims to cover all persons aged 15 and older, residing in private households in the surveyed countries, and it includes detailed information on the survey respondents’ demographics. We include all survey questions on values and opinions that are not country-specific. The application of these criteria leads to the selection of 47 questions. To investigate value alignment of LLMs with individuals from different countries and socio-demographic backgrounds, we will prompt LLMs to answer the same selection of questions. The full list of questions is provided in the Supplementary Material (A.1).

A subset of 21 ESS questions forms a shortened version of the Portrait Value Questionnaire (PVQ), based on Schwartz’s theory of basic human values (Schwartz 1992). Each question is phrased as a short description of a fictional person, in terms of a value or trait, with respondents indicating how much they identify. This questions set has been used in prior work on the relative importance of countries on value formation, as detailed in the related work, though for previous waves of the survey. For this reason, the part of our analysis aimed at estimating the relative effects of countries and socio-demographics will contain a sub-analysis of these PVQ questions.

The 11th wave of the ESS includes 50,116 respondents, all of whom are considered in the analysis. Where possible, post-stratification weights (pspwghts), provided by the ESS, are used to control for non-response patterns to survey participation and ensure demographic representativeness of the sample from each country. We purposefully decided against the analysis weights that adjust for populations, in order to prevent the respondents of countries with large populations from driving the results. The number of participants residing in each country can be found in the Supplementary Material (A.2 & A.3), together with the missing data across considered value-laden and socio-demographic questions. For most socio-demographics and questions missingness is limited; however, *Income decile*, *Occupation* and *Internet Time per Day* show higher rates of missing data.

Models and Prompting We use a set of LLMs from four different popular providers based in different countries: OpenAI and Anthropic from the U.S., Mistral from Europe, and Deepseek from China. Table 1 lists the exact LLMs used. Models were called via the batch API where possible. For all LLMs, no system prompt and the default values for all other parameters are used.

Each LLM is prompted with the selected survey questions 20 times. Prompts all follow the same schema: “[preamble][survey_question] Answer with a whole number on

¹Carried out by European Social Survey European Research Infrastructure (ESS ERIC) (2025), see <https://ess.sikt.no/en/>

	GPT		Claude			DeepSeek		Mistral		
	G55	G52	O46	O47	S45	V4	V3	MLG	M35hg	M35md
Answer Statistics										
Quest. ans. (n)	52	51	52	48	46	53	53	53	53	53
Quest. ans. (%)	98.1	96.2	98.1	90.6	86.8	100.0	100.0	100.0	100.0	100.0
Unif. answers (%)	53.8	49.0	76.9	72.9	52.2	15.1	28.3	66.0	35.8	32.1
Disagreement (%)	11.05	13.43	5.02	6.45	10.78	25.48	19.12	6.64	18.52	17.70
Mean # diff. from MV (n)	2.15	2.41	0.96	0.96	1.85	5.08	3.79	1.26	3.26	3.15
Refusal rate $\mathcal{O}$ (%)	4.6	12.2	8.5	20.8	19.5	1.2	1.7	0.8	11.2	9.5
Refusal rate $\mathcal{Q}$ (%)	0.4	4.1	5.0	11.8	7.4	0.2	0.3	0.4	11.4	8.9
Alignment Scores										
Overall ($A_{\mathcal{P},m,\mathcal{O}}$)	0.607	0.609	0.705	0.746	0.695	0.632	0.634	0.717	0.614	0.610
Overall ($A_{\mathcal{P},m,\mathcal{Q}}$)	0.581	0.591	0.703	0.745	0.699	0.608	0.611	0.708	0.584	0.581

Model legend: **G55:** gpt5.5, **G52:** gpt5.2, **O46:** claude_opus46, **O47:** claude_opus47, **S45:** claude_s45, **V4:** deepseek_v4, **V3:** deepseek_v3, **MLG:** mistral_lg, **M35hg:** mistral_m35_hg, **M35md:** mistral_m35_md

Table 1: Model response statistics summary and overall alignment scores across the whole population $\mathcal{P}$ and question set $\mathcal{O}$ (we actually consider 53 questions, with a subsets of 9 questions corresponding to 3 conceptual questions). The bootstrap CI values are all well-bounded (under ± 0.005) and can be found in the Supplementary Material, together with the full model end-points. $\mathcal{Q} \subset \mathcal{O}$ denotes the subset of questions answered by all models, and MV is Majority Vote.

the scale of [min_scale] to [max_scale], where [min_scale] means '[min_label]' and [max_scale] means '[max_label]'." The labels and scales, as well as their orientation are exactly as in the ESS. The "[preamble]" is only included for the PVQ questions and refers to the following: "Please listen to each description and tell me how much each person is or is not like you.". Models are only prompted in English. Visualisations on refusals or invalid answers and answer dynamics of the different models can be found in the Supplementary Material (A.4).

Following criticism voiced, models are prompted to respond to the survey questions without being explicitly instructed to follow any specific answer structure (Wang et al. 2024). From the free-text responses, answers are mapped to the corresponding Likert scale categories wherever possible. This is simple for models that hide their thinking process or when this process is clearly separated from the output, as they often return a single number or sentence. All other answers are extracted with a rule-based approach. Based on hand annotation of a subset of 127 answers that were not a single number, the accuracy of this rule-based approach was 97.6%, the majority of these being refusals or invalid answers.

In the upper panel of Table 1 the refusal patterns across models can be seen, together with their variation across given answers. Additionally the number of distinct questions that the models did not refuse to answer at least once and the percentage of overall refusals are given, as well as other statistics on the response pattern with respect to refusals. To facilitate the reporting of findings, answers are chosen by majority vote across the model calls, and alignment scores calculated with respect to these majority answers. To allow direct comparison across models we also only include questions all models answered in the subsequent analysis. The restricted set excludes 8 questions. Most of these questions are either on more controversial topics or about specific institu-

tions with four of these questions coincide with the questions that display the highest degree of missingness of survey responses. The exact questions are indicated in the Supplementary Material (A.1).

Even though Moore, Deshpande, and Yang (2024) find that LLMs are relatively consistent over value-laden questions, this is not necessarily the case for some of the models we include, most notably for the DeepSeek models and the mistral-medium-3.5 model with medium and high reasoning effort, as can be seen in Table 1. We base our analysis on the majority vote for ease of reporting, but provide additional analyses that account for the answer variation across model calls in the Supplementary Material (A.5), these suggest robustness of our results.

Definition of Alignment

Alignment Score We collapse the level of agreement of explicitly stated opinions, values, and attitudes between survey participants and different LLMs on a variety of questions into a single alignment score between 0 and 1 for each model and person.

More formally, let $\mathcal{P}$ be the set of all survey participants and $\mathcal{M}$ be the set of all considered models. For a person $p \in \mathcal{P}$ and a model $m \in \mathcal{M}$, consider a single question q from the set of all questions $\mathcal{Q}$, where q is a Likert style question with a number R_q of answer modalities. Let

$$A_{p,m,q} = 1 - \frac{|a_{p,q} - a_{m,q}|}{|R_q| - 1},$$

where $a_{p,q}$ is the answer of the person p to question q and $a_{m,q}$ the majority vote across the model calls of model m . This score $A_{p,m,q} \in [0, 1]$ denotes the similarity between the answer of a person and a model m for a specific question, that is induced by the normalised distance between the answers on the Likert scale; it is implicitly assumed that the answers have equal distances across the scale. Averaging

over all questions gives us one alignment score to a specific model m for each person p and the set of questions $\mathcal{Q}$:

$$A_{p,m,\mathcal{Q}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} A_{p,m,q}.$$

$A_{p,m,\mathcal{Q}}$ is still in $[0, 1]$, where higher scores correspond to higher alignment. For a group $G \subset \mathcal{P}$ we calculate the mean alignment score within that group as

$$A_{G,m,\mathcal{Q}} = \frac{1}{|G|} \sum_{p \in G} A_{p,m,\mathcal{Q}}.$$

The lower panel of Table 1 displays overall mean alignment scores across the whole population for the set of all questions $\mathcal{O}$ and the restricted set of questions that all models answered $\mathcal{Q}$. Overall alignment on the set of questions $\mathcal{Q}$ varies markedly across models, ranging from 0.581 to 0.745. Nevertheless, we found patterns of group deviation from the overall mean to be similar across models. We thus focus our analysis on aggregates across models for ease of exposition and to avoid centring a single provider or model, making model-level remarks only to signal particularly atypical behaviour. All figures in the main text nevertheless report model-level results for the sake of completeness.

Cross-Model Deviation To report on cross-model deviations in alignment patterns, we consider the cross-model deviation, defined for a given group G , as follows:

$$d_{G,\mathcal{M},\mathcal{Q}} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} (A_{G,m,\mathcal{Q}} - A_{\mathcal{P},m,\mathcal{Q}}).$$

In other words, for each model we calculate the deviation of a subgroup's alignment score from the overall population alignment score for that specific model, after which we compute the mean across all $\mathcal{M} = 10$ considered models.

Analytical Approach

These alignment scores enable us to explore our primary research questions: how alignment differs across European countries and socio-demographic factors, and whether these observed variations are driven more by cross-national differences or by individual-level demographics.

Estimating Cross-Model Mean Deviation We begin by presenting aggregate cross-model deviations in alignment patterns $d_{G,\mathcal{M},\mathcal{Q}}$ for different countries and socio-demographic groups.

We account for uncertainty using a bootstrap set-up. For ease of interpretation of confidence intervals, the results we report in the following account only for the sampling uncertainty of the ESS. Mean deviations and their 95% confidence intervals are computed across 5,000 bootstrap samples. In each bootstrap sample, for all 10 considered models, we compute: i) the model's average alignment score across the population; ii) each individual's deviation from that average alignment score; iii) the group level estimate of deviation; iv) and the mean of this quantity across models. From this procedure, we obtain an estimation of the mean deviation from the overall alignment $d_{G,\mathcal{M},\mathcal{Q}}$ across models for

each considered socio-demographic subgroup and country of residence, including the 95% confidence interval of the resulting scores.

A secondary analysis jointly accounting for the answer variability of the LLMs and the sampling uncertainty is reported in the Supplementary Material (A.5). It suggests the robustness of our key findings relative to cross-model deviations, although absolute scores themselves vary considerably across joint bootstrap samples.

Inverse Propensity Weighting Countries could differ in terms of average alignment scores simply because of differences in terms of their socio-demographic composition. To investigate this issue, we reweight ESS respondents using inverse propensity weighting (Rosenbaum and Rubin 1983), so that each country's reweighted distribution of socio-demographic variables approximately corresponds to the pooled ESS distribution (a full description of the methodology is provided in the Supplementary Material (A.7), along with robustness checks). This enables us to document reweighted mean alignment scores across countries: if country differences were driven by differences in socio-demographic composition, we would expect country differences to be substantially reduced by reweighting.

Explaining the Variance of Alignment Scores To assess how much of the variance in individual alignment scores can be explained by socio-demographics and country of residence, we fit models using three sets of covariates: country of residence only, the full set of socio-demographic variables only, and both combined. We compare explained variance (R^2 , estimated over 10-fold cross validation) across these covariate sets to disentangle the relative contributions of geography and individual-level characteristics.

For each covariate set, we fit two model classes: ordinary least squares (OLS) regression without interaction terms, and gradient boosted tree ensembles, as implemented in XGBoost (Chen and Guestrin 2016). Since all covariates are categorical, both model classes can capture non-linear effects within each variable. The distinction lies in the ability of the models to capture underlying multivariate complexity of the relationship between covariates and alignment scores. Comparing the two therefore reflects the degree to which interactions between socio-demographics, and between socio-demographics and country of residence, contribute to alignment score variance.

Results and Analysis

In the following, we first discuss cross-model mean deviations in alignment scores across socio-demographic subgroups and countries of residence. We further use inverse propensity weighting to investigate to what extent country level differences might be explained by differences in socio-demographic composition, and finally, we describe the results of predictive modelling aimed at disentangling the relative contributions of country of residence and individual socio-demographics to alignment score variance.

Patterns of (Mis-)Alignment

We begin by discussing patterns of cross-model alignment deviations $d_{G,M,Q}$ across socio-demographic subgroups and countries of residence, reported in Figures 1 and 2.

Gender In Figure 1 we see that on average, respondents identifying as women have higher alignment scores, with a mean difference in deviation across models of 0.0095 between men and women (the option "other", though part of the survey, was omitted due to few respondents identifying themselves in this category). Only for a single model, `claude-opus-4-7`, does this difference change sign, though the difference is very small.

Ethnicity and Migration Background The Immigration Background variable was coded from both the respondent's country of birth and their parents', where available. The exact coding can be found in the Supplementary Material (A.3). Here, differences of mean deviation across models are not large at 0.0068. Yet, respondents with a Western immigration background have alignment scores that are higher than average, in contrast to those with a non-Western immigration backgrounds who lower alignment scores than average. In terms of feeling, or rather not feeling, as a member of the ethnic majority we see a larger spread of alignment scores. Participants that do not identify as being part of the ethnic majority are on average in much lower agreement with LLMs, with a deviation of -0.01.

Socio-Economic Status We next consider variables linked to socio-economic status: education, income decile, household income feeling, childhood financial difficulties, main activity and occupation. For each one of the following three, income decile, perceived financial difficulty, and childhood financial difficulties, a clear pattern of experiencing high financial stability (throughout life) means a higher mean agreement of stated values between individuals and LLMs. The widest difference, of 0.0385, is found between the highest and lowest categories in terms of *Household Income Feeling*. Considering education, we see a wide spread of the mean deviations of alignment scores, with a clear pattern: on average people with a higher education have higher alignment scores compared to the overall population. There is also a noticeable jump from people who have an education at a master's level to those having a doctoral degree. In terms of respondents' main daily activity, it mostly stands out that the unemployed tend to have lower alignment scores, with quite a wide spread of different mean deviations across models. Considering the alignment scores across occupations supports a class conscious interpretation: those from a higher social class, with better education, higher income and in more white-collar occupations, on average have opinions that are better represented across the considered LLMs.

Urban-Rural Setting Domicile types presented in Figure 1 are ordered from more urban to more rural. No obvious overall pattern can be observed. Compared with people living in a big city, those that live in suburbs or outskirts of a big city seem to have value profiles that LLMs' explicitly stated values are closer to. People living on a farm or the

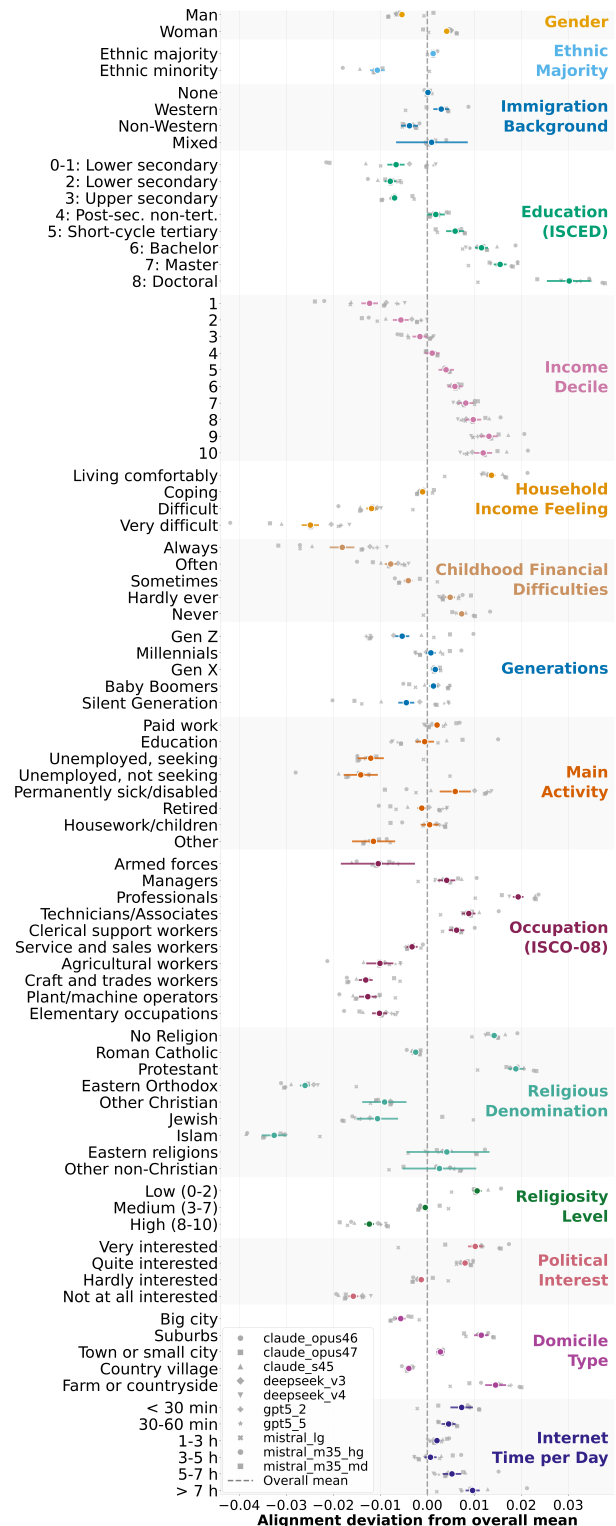


Figure 1: Cross-model mean deviation from mean population alignment score by socio-demographic group (ESS Wave 11). Dashed line: no deviation; gray symbols: model-specific means; error bars: 95% bootstrap CIs.

countryside emerge as those best aligned with among domicile types, with a cross model mean deviation of 0.0145.

Religious Identity Overall, more religious people have lower alignment scores compared to non-religious people. The widest spread across all socio-demographics in Figure 1 relates to religious denomination: respondents identifying as Muslim and Eastern Orthodox, on the one hand, and protestants, on the other, are separated by 0.051 points. Muslims emerge as the socio-demographic group for which we observe the most negative deviation (-0.035). Notably, Roman Catholics and those belonging to other Christian denominations have comparatively lower scores than Protestants.

Generations The pattern emerging for the generations follows a quadratic shape: on average, models on average align worse with the youngest and oldest cohorts compared to the overall population. However, this U-shape results from the aggregation of diverging model-specific patterns (see also the Supplementary Material (A.4), where models are plotted separately). For both young and old individuals, comparative alignment will thus primarily depend on model choice.

Online Activity The bottom of Figure 1 depicts the alignment scores of people grouped by the amount of time they spend on the internet each day. Both the stances of the groups that spend the most and the least time online are best represented. Higher alignment scores for people who spend more time online may not be surprising: they might be the group contributing the most to digital spaces, and thus having authored more of the training texts of language models. That the group of people who spend little to no time online have a higher alignment score could be worth exploring.

Political Interest Models align better with more politically interested individuals on average, especially compared with those not at all interested.

Countries Figure 2 displays the deviation of the alignment scores for all countries surveyed in Wave 11 of the ESS. The spread of the alignment scores across countries is rather high. Comparing upper and lower ends of the figure, it can be seen that the mean values and opinions of people in the Scandinavian and central European countries are comparatively best reflected across models, while those from some Balkan and Baltic countries are least captured. Between the country with the lowest and the one with highest mean deviation from overall alignment across models, Bulgaria and Sweden, there is a difference of 0.0896, higher than within any one socio-demographic factor.

To ensure that the observed patterns are not artifacts of *how* different people are answering surveys, we carry out some additional analysis. We investigate the tendency to pick Likert scale items further away from the middle option. When considering countries, differences in terms of *extremeness* could arise from language differences as well as cultural tendencies of strength in expressing opinions. Considering different socio-demographic subgroups, such differences could arise from self-confidence, engagement and issue salience. In the Supplementary Material (A.6) the relationship between alignment scores and extremeness of re-

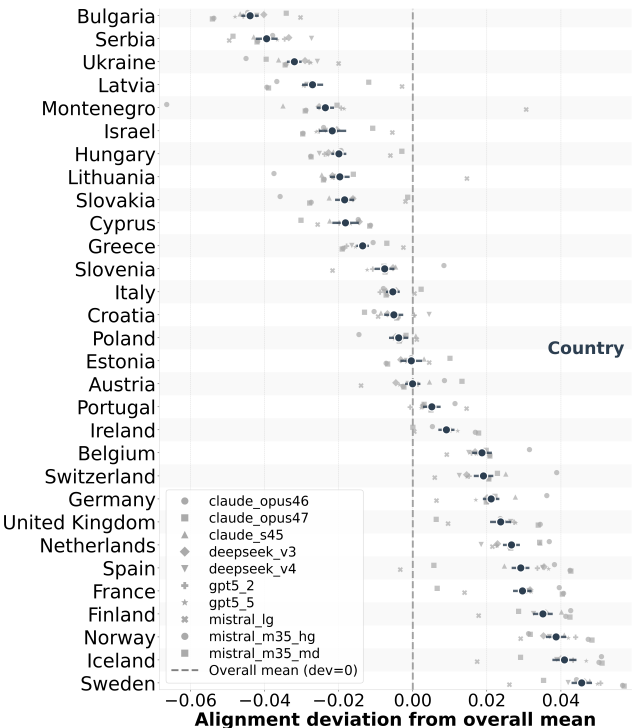


Figure 2: Cross-model mean deviation from the mean population alignment score by country in ESS Wave 11. For each country the mean deviation and the 95% bootstrap confidence intervals are shown. A vertical dashed line indicates no deviation; gray symbols indicate model-specific means.

spondents’ answers can be seen; as well as alignment scores that an hypothetical model that would always pick the mid-point on Likert scales would achieve. Overall, we find that the extremeness of respondents’ answer does not explain the core of findings. Nevertheless, subgroups’ tendencies related to the extremeness of answers could play a role in determining part of the alignment patterns, especially for the models *claude-opus.4-7* and *mistral-lg*.

Within vs between country effects

How the patterns observed in Figure 1 & 2 relate to each other remains unclear, as the considered socio-demographic and country variables are correlated to each other. In particular, given the alignment literature’s focus on national differences, it is of particular interest to assess whether the substantial cross-country differences observed may be driven by differences in national socio-demographic structures.

Accounting for Demographic Composition We study the effect of composition by using Inverse Propensity Weighting (IPW) to reweight country samples so that they share the distribution of socio-demographics in the full sample (Rosenbaum and Rubin 1983). Details on the procedure, histograms of per-country propensity scores, and balance checks are provided in the Supplementary Material (A.7). A noteworthy methodological choice pertains to the han-

dling of respondents that have low propensity scores, i.e. are particularly typical of their country conditional on socio-demographics. Figures reported in the main text are computed by clipping propensity scores at 0.01. The Supplementary Material provides robustness checks in which they are dropped at the 0.01 and 0.05 thresholds; we find these choices not to affect the main conclusions of the analysis.

If most of the alignment differences between countries could be explained by their socio-demographic composition, we would expect reweighting to reduce the dispersion of country mean alignment, and to bring alignment deviations closer to zero. Table 2 displays the standard deviation of country means per LLM; Figure 3 visualizes changes in cross-model mean deviations before and after reweighting. We find the variance of country means to remain mostly unchanged by reweighting, and that post-reweighted cross-model mean deviations are not much closer to zero compared to non-weighted ones. We conclude that between-country differences cannot be explained by socio-demographic compositional differences (at least with regard to the socio-demographic variables we consider).

Predictive Modelling and Variance Decomposition We turn to predictive modelling of individual alignment scores, to understand the proportion of their variance that can be explained by countries, by socio-demographics, and by their combination, and gauge to what extent the underlying relationship between socio-demographics, countries and alignment scores is driven by interactions. Accordingly, we consider two model classes: linear regressions without interaction terms, and Gradient Boosted Models (GBMs). Note

LLM	σ_{pre}	σ_{clip}	Δ & 95% CI
claude_opus46	0.0315	0.0294	-0.0021 [-0.0025, -0.0010]
claude_opus47	0.0205	0.0188	-0.0017 [-0.0022, -0.0005]
claude_s45	0.0261	0.0244	-0.0018 [-0.0023, -0.0005]
deepseek_v3	0.0234	0.0228	-0.0006 [-0.0010, +0.0007]
deepseek_v4	0.0231	0.0228	-0.0003 [-0.0006, +0.0013]
gpt5_2	0.0247	0.0240	-0.0007 [-0.0009, +0.0010]
gpt5_5	0.0249	0.0241	-0.0008 [-0.0012, +0.0008]
mistral_lg	0.0168	0.0143	-0.0025 [-0.0033, -0.0008]
mistral_m35_hg	0.0295	0.0284	-0.0011 [-0.0016, +0.0005]
mistral_m35_md	0.0296	0.0285	-0.0012 [-0.0017, +0.0005]

Table 2: Between-country standard deviation of country mean alignment per LLM, before and after IPW reweighting (clip @ $\hat{P} < 0.01$). $\Delta = \sigma_{clip} - \sigma_{pre}$; 95% bootstrap CI on 1000 bootstraps.

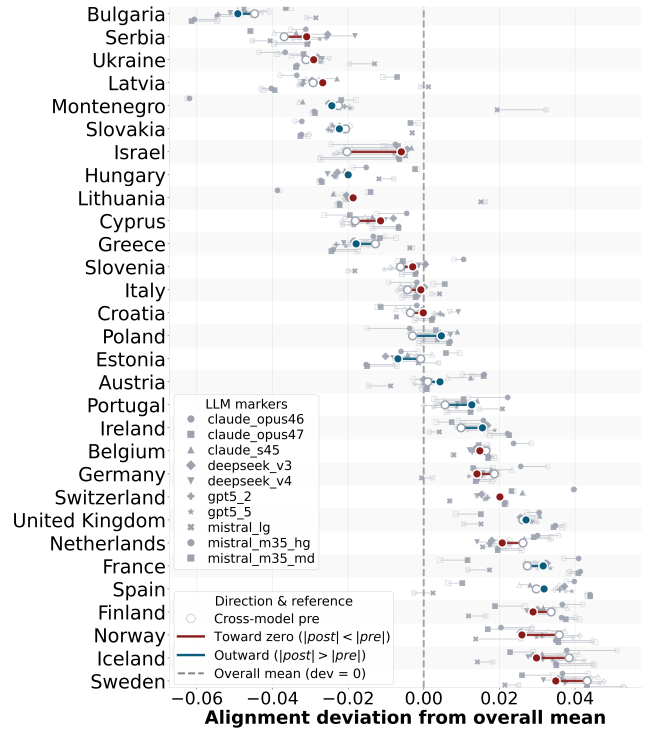


Figure 3: Cross-model mean deviations pre (white) and post (filled) clip-IPW reweighting (clip @ $\hat{P} < 0.01$).

that all variables considered are categorical, and are one-hot encoded as such in the linear models; an important precision given some non-linear trends observed for ordinal-categorical variables in Figure 1. The difference between the predictive performance of GBMs and linear models enables us to assess the role of interactions in the relation between socio-demographics, countries and alignment scores.

In addition to studying all value-related questions answered by all models ($\mathcal{Q}$), we conduct a sub-analysis on the 21 questions making up Schwartz’s Portrait Value Questionnaire (PVQ). As noted in the Related Work, academic debates on the role that nations and cultures play in shaping personal values have centred on this, or similar, specific definition of values rather than the broader set of questions $\mathcal{Q}$.

In Figure 4 the mean test R^2 are shown. The Supplementary Material (A.8) includes both train and test R^2 and some further intuitions that follow. In the upper bar chart we can see that using socio-demographics and countries together in a GBM explains a substantial proportion of variance between the individual alignment scores. This is true for all LLMs with `claude-opus-4-6` having the highest test R^2 of 0.428. But even for `mistral.lg` with the lowest test R^2 , 18.9% of the variability in the outcomes can be explained.

By comparing the test R^2 for the different sets of features, and specifically for *country-only* and *SD-only*, we find that the variation explained by country alone is at least on par with that explained by the full set of 15 socio-demographic factors for all LLMs. The country of residence as a stand-alone variable explains between 7.3% for `mistral.lg` and

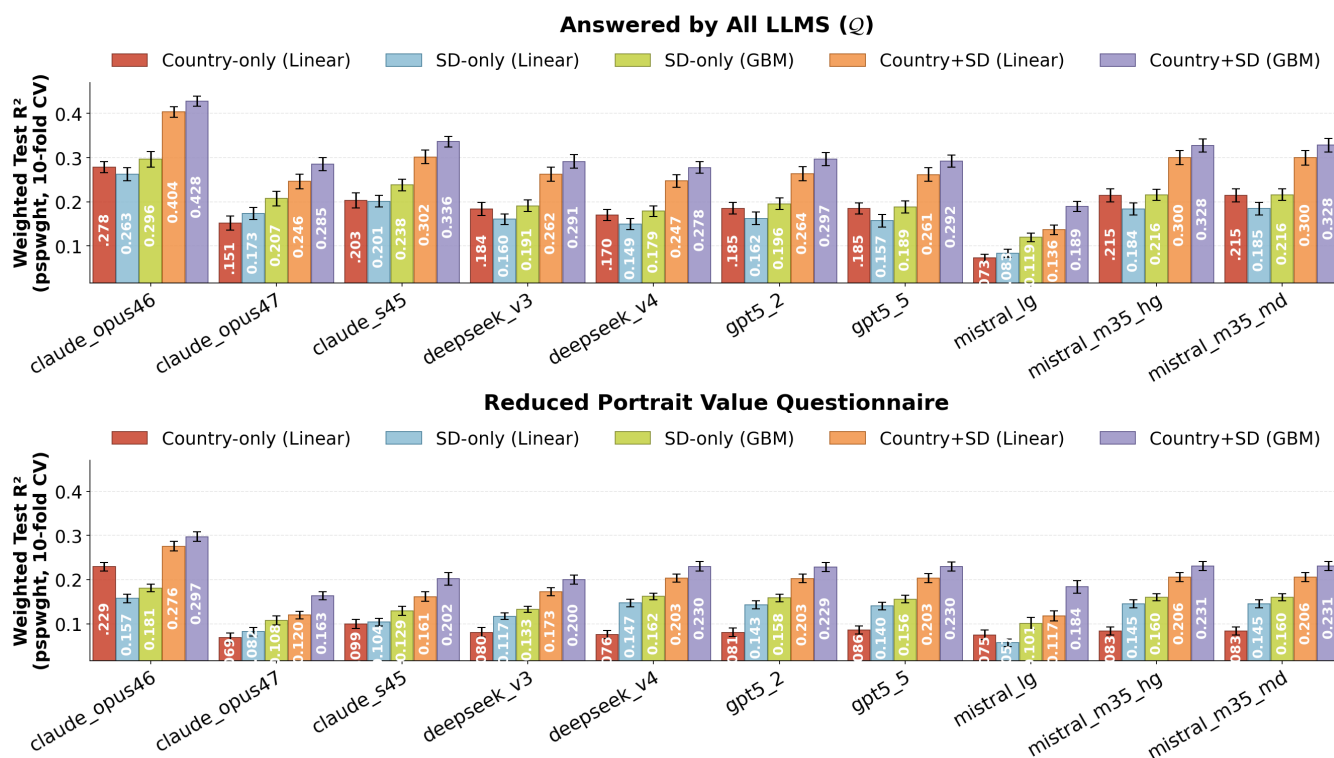


Figure 4: Mean test R^2 over 10-fold cross validation with standard deviations, for five methods predicting individual alignment scores. The upper chart shows scores on Q , the set of questions answered by all LLMs; the lower chart shows scores on PVQ questions. Covariates vary between i) country of residence (*Country-only*); ii) all socio-demographics considered in Figure 1 (*SD-only*); iii) both combined (*Country+SD*). Methods vary between OLS regression (*linear*) and gradient boosted models (*GBM*). R^2 are weighted by pspwght.

27.8% for *claude_opus-4-6*, the remaining models falling between 15% and 22%. These figures emphasises that country of residence is far from being negligible when considering value alignment on a question set such as Q .

The same comparison on the PVQ yields different results. In contrast to the previous observations, dynamics differ across language models for this question set. For the *mistral-medium-3-5* variations, and the models from the *gpt* and *deepseek* families, country alone explains a much smaller proportion of variance, and also a much smaller proportion relative to that explained by the socio-demographics. Only for *claude-opus-4-6* the explanatory power of country is much higher than that of the *SD-only* models.

This striking difference of variance explained by countries across the two question sets emphasises the importance of survey design and questions selection. Including questions that survey stances on broader value-laden topics may primarily capture the political climate and media landscape in a country. But excluding them might mean that people with identical but abstract values, who have completely contrary opinions on practical matters seeming like they are “aligned” to an LLM that only one shares fundamental stances with. Thus, not one definition is more correct, rather there is a need for higher sensitivity for the implications of how value

alignment is approached.

Comparing the linear regression and GBM for the *SD-only* and *Country+SD* models, reveals that allowing for higher order interactions does not yield much improvement. This is true for both question sets and for all LLMs but *mistral_lg*. For socio-demographics this means that there do not seem to be large intersectional groups for whom alignment cannot be explained in an additive manner. The moderate increase between the linear and GBM on *Country+SD* suggests that the socio-demographic structures explaining the heterogeneous outcomes are largely the same across Europe.

Discussion

This work constitutes the first cross-national evaluation of value alignment across socio-demographics, considering 30 countries surveyed in the ESS. We examine the comparative differences between socio-demographic groups and explore how countries and socio-demographics contribute to individual alignment scores. Our key findings include i) there are disparities of alignment scores across socio-demographic groups in the European context; ii) country level differences in alignment scores cannot be explained away by countries’ socio-demographic compositions; iii) the country of residence and the set of socio-demographics explain a simi-

lar amount of variation in the alignment outcomes for individuals. We additionally distinguish between two subsets of questions corresponding to differing notions of values. For most LLMs the relative variance explained by countries shrinks notably when considering the narrower definition, corresponding to Schwartz's Portrait Value Questionnaire.

Previous studies have found LLMs to be WEIRD, that is, aligned with countries that are Western, Educated, Industrialized, Rich and Democratic, when it comes to the stated stances they elicit. Our study now extends this finding to the actual people living in some of these WEIRD countries. Even among them similar dynamics can be observed: more educated and richer people from more Western countries make up the groups of people that LLMs are comparatively better aligned with. All indicators associated with higher socio-economic classes are consistently associated with higher alignment scores. These patterns raise a pointed concern: if LLMs are systematically better aligned with higher socio-economic groups, their deployment as general-purpose tools may inadvertently reflect and reinforce the values of already privileged populations.

The three factors most closely related to value and opinion formation also show clear trends: political interest, religiosity level, and religious denomination. Political interest is positively correlated with the alignment score, while religiosity and most religious denominations are negatively correlated. The observed correlations in these cases may be driven by respondents holding traditional or conservative stances, as some questions in our analysis explicitly address gender equality and LGB tolerance (that is questions specifically around homophobia and same-sex couples), which most LLMs will either refuse to answer or answer in support of equality and equal rights. These same topics could also explain the split observed across genders.

Our second research question interrogated the respective role of socio-demographics and cross-national differences in value alignment. First, we find that countries explain a large relative amount of variance in alignment scores, especially when considering the full question set. Second, we verify that the variance observed between countries is not driven by differences in their socio-demographic compositions. Third, the comparison of linear and non-linear models involving socio-demographics and country variables suggests that the structures of socio-demographic differences in alignment scores is likely to be largely consistent across countries. Drawing from these findings, we conclude that countries as entities of study in alignment research cannot be replaced by socio-demographics, but that both must be considered. How this translates to larger or other global contexts remains to be investigated.

As reported in the review of related work in the field of values and values formation, whether socio-demographics or countries and culture are primary drivers of the value formation of individuals is subject to academic debate. In the context of value alignment, our results support both sides. The definition of "values" can be narrow and abstract as in Schwartz's definition, or wider to include general stances on value-laden topics. The chosen definition substantially affects the amount of variance in value alignment scores that

can be explained by respondents' countries of residence.

This calls for a reflective and more transparent definition of *value* alignment research. For pluralistic approaches, the question does not only become how to achieve better representation, but which are the dimensions that should be aggregated and evaluated across. For example, when aiming to optimize pluralistic alignment with respect to Schwartz's definition of basic human values, cultures and nations are not the most prominent dimensions to reduce disparities along.

Limitations and Future Work This paper has several limitations. A first limitation lies in a limited exploration of variability of the answers of LLMs. On the one hand, aside from a joint bootstrap of population and model responses provided in the Supplementary Material as a robustness check, we adopted majority-vote aggregation of model answers to facilitate reporting without accounting for their variability. On the other hand, given the necessity of comparing model answers to those of human respondents, no prompt variation analysis was conducted. We nevertheless note that LLMs comparable to those investigated have been found to be relatively consistent in their answers to similar value-related questions in previous work (Wang et al. 2024; Moore, Deshpande, and Yang 2024).

A limitation shared across most of the alignment literature is that Multiple Choice Question (MCQ) answers are inevitably an imperfect surrogate measure for values. While we believe there are insights to be gained by studying LLMs' explicit positioning with respect to values, MCQ answers have been criticized as unrepresentative of the natural behaviour of LLMs (Shen, Clark, and Mitra 2025). This motivates the need for future work on more ecologically valid and implicit approaches to elicit stances from LLMs.

While our analysis focused on broad patterns of value alignment we found to exist across models, we noted differences to exist between them, in particular for `mistral.llg` and `claude.opus.4-6`. Further analysis of these differences, and of the effect of reasoning on responses, could form an avenue for further work.

Prompting solely in English leads to more possible limitations. The first concerns the comparison of answers given in different languages. We note that work on measure invariance in the WVS (Alemán and Woods 2016) and within the ESS (Davidov, Schmidt, and Schwartz 2008) suggests that the concepts measured are sufficiently stable across the countries considered in our paper. A further limitation is that our analysis does not consider possible inconsistency of the LLMs' answers across the multiple languages, which could add further dimensions of alignment disparities.

Lastly, we acknowledge the limitation of the regional focus of our work. While regional surveys improve cross-national measurement validity, enabling a more contextual understanding of value systems, further analyses of socio-demographics in other global regions, using surveys such as the Afrobarometer or the Latinobarómetro, are necessary.

Ethics Statement

Even though we capture and consider *more* aspects of the individuals we compare than previous studies, we are still only

considering them through a lens of set categorical indicators. As Dervin (1989) discusses in the context of user research in communication studies, research categories are constructions that are invented. Employing them may primarily serve the inventors and even exacerbate inequalities. For instance, aggregating to a majority value profile in a country or culture will disregard minority perspectives. While these minority perspectives might be held by groups as defined by the socio-demographics we consider, they could also be held by a subset of people that relate to each other in a way we do not capture with common pre-defined categories. Thus, future research could include or centre alternative categories, possibly along those proposed by Dervin (1989) such as the “Actor’s situation”, “Gaps in sense making” and “Actor-defined purposes”.

Additionally, whenever alignment and value systems are concerned, ethical implications of the act of steering or approximating specific groups (and not others) have to be considered. For an in-depth review see (Kirk et al. 2024a).

Another ethical consideration that accompanies this research is the inevitable anthropomorphism of AI systems encouraged by the research done on them (Salles, Evers, and Farisco 2020; Placani 2024). In this paper this happens through the framing of *answering questions* and *refusing*, in a context where LLMs *explicitly state values*. Contributing to the anthropomorphism of AI has implications on how AI is both perceived and developed.

This research further contributes to persistent imaginaries of AI, possibly feeding into the narrative of the inevitability of AI (progress) that can only be ameliorated in terms of exploitation, biases and unequal interest representation. So, while it is important to address disparities perpetuated by contemporary AI, this should not distract from rethinking the underlying systems more fundamentally.

Acknowledgements

The research leading to these results was funded/co-funded by the European Union (ERC, VIGILIA, 101142229), the Special Research Fund (BOF) of Ghent University (BOF20/IBF/117), the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme, and the FWO (project no. G073924N). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. For the purpose of Open Access the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

- Abeliuk, A.; Gaete, V.; and Bro, N. 2025. Fairness in LLM-Generated Surveys. ArXiv:2501.15351 [cs].
- Akaliyski, P.; Welzel, C.; Bond, M. H.; and Minkov, M. 2021. On “nationology”: The gravitational field of national culture. *Journal of Cross-Cultural Psychology*, 52(8-9): 771–793.
- Alemán, J.; and Woods, D. 2016. Value Orientations From the World Values Survey: How Comparable Are They Cross-Nationally? *Comparative Political Studies*, 49(8): 1039–1067.
- AlKhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; and Diab, M. 2024. Investigating Cultural Alignment of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 12404–12422. Bangkok, Thailand: Association for Computational Linguistics.
- Batzner, J.; Stocker, V.; Schmid, S.; and Kasneci, G. 2025. German Parties QA: Benchmarking Commercial Large Language Models and AI Companions for Political Alignment and Sycophancy. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)*, 330–342. Association for the Advancement of Artificial Intelligence.
- Bilsky, W.; Janik, M.; and Schwartz, S. H. 2011. The Structural Organization of Human Values-Evidence from Three Rounds of the European Social Survey (ESS). *Journal of Cross-Cultural Psychology*, 42(5): 759–776.
- Cao, Y.; Zhou, L.; Lee, S.; Piqueras, L. C.; Chen, M.; and Hershovich, D. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, 53–67.
- Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Davidov, E.; Schmidt, P.; and Schwartz, S. H. 2008. Bringing Values Back In: The Adequacy of the European Social Survey to Measure Values in 20 Countries. *Public Opinion Quarterly*, 72(3): 420–445.
- Dervin, B. 1989. Users as research inventions: How research categories perpetuate inequities. *Journal of communication*, 39(3): 216–232.
- Durmus, E.; Nguyen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. ArXiv:2306.16388 [cs].
- European Social Survey European Research Infrastructure (ESS ERIC). 2025. ESS11 - integrated file, edition 4.1. [Data set]. Sikt - Norwegian Agency for Shared Services in Education and Research. DOI: https://doi.org/10.21338/ess11e04_1.
- Fischer, R.; and Schwartz, S. 2011. Whence Differences in Value Priorities?: Individual, Cultural, or Artifactual Sources. *Journal of Cross-Cultural Psychology*, 42(7): 1127–1144.
- Greenfield, P. M. 2014. Sociodemographic Differences Within Countries Produce Variable Cultural Values. *Journal of Cross-Cultural Psychology*, 45(1): 37–41.

- Hofstede, G. 1980. *Culture's consequences: International differences in work-related values*, volume 5 of *Cross-Cultural Research and Methodology Series*. Beverly Hills, CA: Sage Publications, Inc.
- Inglehart, R.; and Welzel, C. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. New York: Cambridge University Press.
- Jasanoff, S. 2004. Ordering knowledge, ordering society. In *States of knowledge*, 13–45. Routledge.
- Jose, B.; Joseph, D.; Mohan, V.; Alexander, E.; Varghese, S. K.; and Roy, A. 2025. Outsourcing cognition: the psychological costs of AI-era convenience. *Frontiers in Psychology*, 16: 1645237.
- Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4): 383–392.
- Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; et al. 2024b. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37: 105236–105344.
- Kroes, P.; Franssen, M.; Poel, I. v. d.; and Ottens, M. 2006. Treating socio-technical systems as engineering systems: some conceptual problems. *Systems Research and Behavioral Science*, 23(6): 803–814.
- Liemt, E. v.; Shelby, R.; Smart, A.; Kumbale, S.; Zhang, R.; Dixit, N.; Rashid, Q. M.; and Smith-Loud, J. 2026. Cultural Perspectives and Expectations for Generative AI: A Global Survey Approach. ArXiv:2603.05723 [cs].
- Lin, C.; Yuan, W.; Jiang, Z.; Huang, B.; Zhang, R.; Ge, J.; Xu, Y.; and Yu, J. 2026. AlignSurvey: A Comprehensive Benchmark for Human Preferences Alignment in Social Surveys. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 38908–38916.
- Liu, Y.; Kaneko, M.; and Chu, C. 2026. On the alignment of large language models with global human opinion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 37673–37681.
- Ma, B.; Yoztyurk, B.; Haensch, A.-C.; Wang, X.; Herklotz, M.; Kreuter, F.; Plank, B.; and Aßenmacher, M. 2025. Algorithmic Fidelity of Large Language Models in Generating Synthetic German Public Opinions: A Case Study. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 1785–1809. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.
- Mihalcea, R.; Ignat, O.; Bai, L.; Borah, A.; Chiruzzo, L.; Jin, Z.; Kwizera, C.; Nwatu, J.; Poria, S.; and Solorio, T. 2025. Why AI Is WEIRD and shouldn't be this way: towards AI for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28657–28670.
- Miles, A.; and Yeh, C. 2022. Do demographic predictors of personal values vary by context? A test of Schwartz's value development theory. *Social Sciences & Humanities Open*, 5(1): 100264.
- Moore, J.; Deshpande, T.; and Yang, D. 2024. Are Large Language Models Consistent over Value-laden Questions? In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15185–15221. Miami, Florida, USA: Association for Computational Linguistics.
- Nadeem, A.; Seth, A.; Nasim, M.; and Naseem, U. 2026. Bias Beyond Borders: Political Ideology Evaluation and Steering in Multilingual LLMs. ArXiv:2601.23001 [cs].
- Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52.
- Placani, A. 2024. Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, 4(3): 691–698.
- Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.
- Salles, A.; Evers, K.; and Farisco, M. 2020. Anthropomorphism in AI. *AJOB neuroscience*, 11(2): 88–95.
- Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.
- Schwartz, S. H. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*, volume 25, 1–65. Elsevier.
- Sen, I.; Lutz, M.; Rogers, E.; Garcia, D.; and Strohmaier, M. 2025. Missing the Margins: A Systematic Literature Review on the Demographic Representativeness of LLMs. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 24263–24289. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Shen, H.; Clark, N.; and Mitra, T. 2025. Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3097–3118.
- Shetty, A.; Naseem, U.; Aletras, N.; Dras, M.; Ji, H.; and Nakov, P. 2026. Towards Pluralistic Alignment of LLMs: A Comprehensive Survey. Preprints.org. Doi: 10.20944/preprints202603.1876.v1.
- Sukiennik, N.; Gao, C.; Xu, F.; and Li, Y. 2025. An Evaluation of Cultural Value Alignment in LLM. ArXiv:2504.08863 [cs].
- Sun, Y.; and Wang, T. 2026. Be friendly, not friends: How LLM sycophancy shapes user trust. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–15.
- Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9): pgae346.

- Torkamaan, H.; Steinert, S.; Pera, M. S.; Kudina, O.; Freire, S. K.; Verma, H.; Kelly, S.; Sekwenz, M.-T.; Yang, J.; Van Nunen, K.; Warnier, M.; Brazier, F.; and Oviedo-Trespalacios, O. 2024. Challenges and future directions for integration of large language models into socio-technical systems. *Behaviour & Information Technology*, 1–20.
- Vilar, R.; Liu, J. H.-f.; and Gouveia, V. V. 2020. Age and gender differences in human values: A 20-nation study. *Psychology and aging*, 35(3): 345.
- Von Der Heyde, L.; Haensch, A.-C.; and Wenz, A. 2025. Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice. *Social Science Computer Review*, 08944393251337014.
- Wang, X.; Hu, C.; Ma, B.; Röttger, P.; and Plank, B. 2024. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *arXiv preprint arXiv:2404.08382*.
- Williams, T.; Weeber, F.; Padó, S.; and Akbik, A. 2026. Beyond Marginal Distributions: A Framework to Evaluate the Representativeness of Demographic-Aligned LLMs. ArXiv:2601.15755 [cs].
- Wright, D.; Arora, A.; Borenstein, N.; Yadav, S.; Belongie, S.; and Augenstein, I. 2024. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 17085–17112. Miami, Florida, USA: Association for Computational Linguistics.
- Xiao, J.; Li, Z.; Xie, X.; Getzen, E.; Fang, C.; Long, Q.; and Su, W. J. 2025. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552): 2154–2164.
- Xu, Y. W.; Chi, C. G.; Guroy, D.; and Cai, R. R. 2025. Rethinking AI anthropomorphism: A holistic conceptualization and scale across AI systems and service contexts. *Technology in Society*, 103189.