The Invisible Embedded "Values" Within Large Language Models: Implications for Mental Health Use

Dorit Hadar-Shoval (dorith@yvc.ac.il)
Psychology Department, Center for Psychobiological Research, Max Stern Yezreel Valley College
Kfir Asraf
Psychology Department, Center for Psychobiological Research, Max Stern Yezreel Valley College
Yonathan Mizrachi
Departments of Sociology & Anthropology and The Jane Goodall Institute, Max Stern Yezreel Valley College; The Laboratory for AI, Machine Learning, Business & Data Analytics, Tel-Aviv University
Yuval Haber
Psychology Department, Max Stern Yezreel Valley College
Zohar Elyoseph
Psychology Department, Center for Psychobiological Research, Max Stern Yezreel Valley College; Department of Brain Sciences, Faculty of Medicine, Imperial College London

Article

Keywords:

Posted Date: October 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3456660/v1

Additional Declarations: No competing interests reported.

Abstract

Values are an integral part of any mental health intervention, profoundly shaping definitions of psychopathology and treatment approaches. As large language models (LLMs) hold promises for mental health applications, it is prudent to evaluate their embedded “values-like” abilities prior to implementation. This study uses Schwartz's Theory of Basic Values (STBV) to quantify and compare the motivational “values-like” abilities underpinning four leading LLMs. The results suggest that Schwartz’s theory can reliably and validly measure “values-like” abilities within LLMs. However, apparent divergence from published human values data emerged, with each LLM exhibiting a distinct motivational profile, potentially reflecting opaque alignment choices. Such apparent mismatches with human values diversity might negatively impact global LLM mental health implementations. The appropriate transparency and refinement of alignment processes may be vital for instilling comprehensive human values into LLMs before this sensitive implementation in mental healthcare. Overall, the study provides a framework for rigorously evaluating and improving LLMs’ embodiment of diverse cultural values to promote mental health equity.

Introduction

As artificial intelligence (AI) rapidly advances, large LLMs like Bard (by Google), Claude 2 (by Anthropic), and ChatGPT-3.5 and 4 (by OpenAI) demonstrate impressive capabilities, opening promising possibilities in mental healthcare, such as expediting research, guiding clinicians, and assisting patients[1]. However, integrating AI into mental health also raises the need to address complex professional ethical questions[2,3].

This study examines these issues through the lens of transcultural psychiatry, which emphasizes the pivotal role of cultural values, beliefs, and customs in understanding mental distress and psychiatric disorders[4]. The well-established Schwartz's Theory of Basic Values (STBV) provides a conceptual framework for analyzing relationships between cultural dynamics, personal influences, and facets of mental well-being[5]. We specifically examine the intersection of LLMs and cultural conceptualizations of values and their association with mental health. Values are integral in mental health, profoundly shaping definitions of psychopathology and treatment approaches[6]. The therapist, the patient, and the alignment of therapist–patient values impact therapeutic interactions and quality of care[7]. Successful cultural adaptation can enhance therapeutic outcomes[8]. With globalization and the accompanying growth of multicultural societies, culturally adapted mental healthcare is challenging but essential[9].

The introduction of AI such as LLMs raises critical questions about the “values-like” abilities of such technologies and whether they align with the diversity of cultural values in mental health[1, 10]. As LLMs can be integrated into areas like diagnosis and patient interactions, extensive training encompassing diverse cultural perspectives on mental health may be required to avoid biases. A rigorous examination of the values-like abilities of AI is crucial when considering its cross-cultural incorporation.

Schwartz's Theory of Basic Values (STBV): A Framework for Capturing Cultural values in Mental Health

A pivotal aspect in grasping cultural impacts on mental health is capturing the latent construct of “culture” in a quantifiable manner[6]. STBV[11,12] provides a comprehensive framework elucidating the nature and role of values guiding human behavior and decision-making. This theory defines values as enduring, trans-situational objectives that differ in significance and serve as guiding tenets steering individuals and social entities[5]. In addition, it delineates seven fundamental attributes inherent to most psychological models of values[11]. First, values involve beliefs about the desired objectives that individuals view as important. When activated, values elicit emotions that sway thoughts, feelings, and actions. Second, values are considered fundamental goals which are relevant across diverse situations, providing a framework for assessing and responding to a broad array of circumstances. Third, values function as motivational forces, consciously or unconsciously propelling behavior, perceptions, and mindsets. Fourth, they contribute to the orientation of actions and judgments. Fifth, the impact of values on conduct is mediated through trade-offs between competing values; when making choices, individuals weigh the relative prominence of conflicting values. Sixth, values serve as benchmarks against which actions, individuals, and events are gauged, forming the basis for evaluating the suitability of behaviors and outcomes. Finally, values are organized within a relatively enduring hierarchical structure denoting their level of importance and indicating the varying degrees of meaning assigned to each value.

Despite these common attributes, what differentiates values is their unique motivational essence. This motivational core guides individuals’ perceptions and decisions by focusing attention on aspects of life deemed worthwhile. Different people prioritize distinct facets of life, resulting in assorted value preferences[5] (see Table 1).

Table 1
The 19 values in the Schwartz Portrait Values Questionnaire organized into 10 values and 4 higher-order values.

19 Values	10 Values	4 Higher Order Values
Self-Direction (Thought) - Thinking creatively and independently	Self-Direction - Thinking and acting independently	Openness to Change - Pursuing intellectual and experiential openness
Self-Direction (Action) - Acting independently and choosing own goals		
Stimulation - Seeking excitement and novelty	Stimulation - Seeking excitement, novelty, and challenge	
Hedonism - Pleasure and sensuous gratification	Hedonism - Pleasure and sensuous gratification	
Achievement - Success according to social standards	Achievement - Personal success through demonstrating competence	Self-Enhancement - Pursuing personal status and dominance over others
Power (Dominance) - Power through exercising control over people	Power - Social status and prestige, control or dominance over people and resources	
Power (Resources) - Power through control of material and social resources		
Face - Protecting one's public image and avoiding humiliation		
Security (Personal) - Safety in one's immediate environment	Security - Safety, harmony, and stability of society, relationships, and self	Conservation - Pursuing order, self-restriction, preservation of the past
Security (Societal) - Safety and stability in the wider society		
Conformity (Rules) - Compliance with rules, laws and formal obligations	Conformity - Restraint of actions, inclinations, and impulses	
Conformity (Interpersonal) - Avoidance of upsetting or harming others		
Tradition - Maintaining and preserving cultural, family or religious traditions	Tradition - Respect, commitment, and acceptance of the customs and ideas of traditional culture and religion	
Humility - Recognizing one's insignificance in the larger scheme of things		
Benevolence (Care) - Devotion to the welfare of ingroup members	Benevolence - Preservation and enhancement of the welfare of people with whom one is in frequent personal contact	Self-Transcendence - Pursuing the welfare of others and transcending selfish concerns
Benevolence (Dependability) - Being a reliable and trustworthy member of the ingroup		
Universalism (Tolerance) - Accepting and understanding those who are different	Universalism - Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature	
Universalism (Concern) - Commitment to equality, justice, and protection for all people		
Universalism (Nature) - Preservation of the natural environment		

Applying Schwartz's model facilitates a keen analysis of cultural dynamics related to mental health. Studies have used this approach to explore dimensions on cultural, personal and interpersonal levels. For example, research on the syndrome of *ataque de nervios* in Puerto Rico illustrated how the cultural value of social harmony developed in response to historical adversity and shapes emotional expression and experience[13]. Though derived from a specific context, the relevance of social harmony has also been found in China where maintaining *guanxi* (social networks), *he xie* (harmony) and *mianzi* (preserving face) impacts views of mental illness[6]. Indeed, depression has been found to often manifest somatically in China to avoid a loss of face[14]. Despite their different histories, the cultural value of social harmony has been shown to exert analogous effects on mental health in both Puerto Rico and China, evidencing the utility of Schwartz's model for understanding cultural illness influences cross-culturally[6]. Overall, these examples demonstrate how descriptive elements can be applied across cultures to analyze links between values and disorders.

At the personal level, studies have revealed that values correlate with outcomes like depression, anxiety, stress, and post-traumatic stress disorder (PTSD). For example, openness was often found negatively associated with depression[15,16], power showed consistently robust positive correlations with worries[17], and universalism had inconsistent correlations with anxiety and worries (both positive and negative)[6]. Within individual countries, few significant correlations emerged between values and stress/PTSD[18]. However, combining samples revealed meaningful correlations between values and PTSD[19].

The variable correlations indicate that relationships between values and mental health depend heavily on cultural context. For example, power predicted worries in a Nepali sample but not in a German sample[16]. While some broad patterns exist, correlations between Schwartz's values and mental health hinge extensively on culture. The framework provides a scaffolding through which to methodically dissect cultural mental health impacts, although specific correlations differ across populations.

At the interpersonal level (in the clinic), researchers have noted that the therapist's and client's values enter the clinical space and influence the therapeutic process in complex ways, such as impacting assessment and treatment approaches, setting therapeutic goals, conceptualizing change, and shaping the therapist-client relationship[7, 20]. A study examining the personal and professional values of Indian therapists showed that the values held by therapists were expressed in their therapeutic practices: the value of acceptance, for example, influenced their stance toward clients[7]. Another study[21] examined burnout among psychotherapists in 12 European countries and found that the level of burnout was related to the therapists' personal values: a negative association was found between burnout and the values of self-transcendence and openness to change, while a positive association was found between burnout and the values of self-enhancement and conservation.

In summary, STBV constitutes a framework for mapping mental health outcomes and elucidating cultural influences on psychopathology and wellness. This becomes particularly relevant when considering implementation of LLMs in mental health, as these models are trained on massive internet data and undergo alignment processes.

Large Language Models and Cultural Values

LLMs have a huge number of parameters, often billions, and are trained on huge corpora[22]. Recently, LLMs have been transformative, revolutionizing academic research and mental health applications [3,23]. A vital factor enabling the usability and popularity of current LLMs is alignment, namely, the process of ensuring models behave in congruence with human values and societal norms[22]. LLMs are initially trained on massive datasets compiled from the internet. These risks ingraining harmful biases, misinformation, and toxic content [24,25]. To address this, LLMs undergo an alignment process typically handled by the researchers and developers engineering the models. Alignment aims to guarantee that the LLM's outputs conform with human values and norms [22,26] .

However, there are presently no established principles or guidelines governing alignment. Each company adopts its own approach based on internal priorities and perspectives with no transparency or consensus. For example, some may emphasize reducing toxic outputs while overlooking potential harms like self-harm content[27]. Best practices are starting to emerge, like adhering to the "helpful, honest, harmless" maxim and using human feedback for refinement[28]. But alignment remains more art than science.

Preliminary studies on the cultural sensitivity of LLMs have revealed varying levels of bias toward different cultures and values. An evaluation of GPT-3.5's cross-cultural alignment found it performed significantly better with American versus other cultural prompts[29]. Another study discussed GPT-3's value conflicts and proposed better contextualization of societal harm and benefit[30], while a different analysis showed biases in its "personality," value system, and demographics[31]. In addition, a more recent work found that GPT-3.5 has differential emotional understanding across mental disorders, reflecting stereotypical views[32].

Opaque alignment by private companies lacks standardized ethical frameworks, thus subtly encoding cultural biases and rigid thinking about disorders misaligned with mental health nuance. The present study therefore looks to methodically map the latent, foundational, and motivational values-like constructs underlying LLMs using Schwartz's validated theory of basic human values as a theoretical framework. Quantifying LLMs' embedded values is essential for illuminating the ethical refinements needed to mold these powerful tools into virtuous, humanistic agents that can provide equitable mental healthcare. The study examines two key questions: 1. Can Schwartz's values model effectively identify and measure values-like constructs embedded within LLMs?; and 2. Do different LLMs exhibit distinct values-like patterns compared to humans and to each other?

Results

Question 1: Can Schwartz's values model effectively identify and measure values-like constructs embedded within LLMs?

To answer this question, we examined the reliability and validity of the Portrait Values Questionnaire-Revised (PVQ-RR) data generated by the LLMs.

Reliability and agreement

We used several methods to assess the reliability and agreement of the 57 items mean score (SimplyAgree module in Jamovi, v 0.1[33]).

Internal consistency reliability was examined via Cronbach's α (Table 2). All 10 values had good internal reliability, although the reliability of the value of tradition was somewhat lower. In order to examine split-half reliability, we divided the samples of each of the LLMs into two parts and examined whether the parts were reliable with each other. The obtained intraclass correlation coefficient (ICC) was .851 (95% C.I.=.626, .940; two-way mixed, average measures, absolute agreement), which is considered excellent[34] to good[35] reliability.

We also conducted Shieh's test of agreement[36] to assess agreement between the two parts, with limit of agreement = 95%, against an agreement bound of ± 2. The test was statistically significant [exact 95% C.I. = -1.168, 1.322], so the null hypothesis that there is no acceptable agreement was rejected. The Bland-Altman limits of agreement (LoA) indicated that the mean bias (.077) was not significantly different from 0 [97.5% C.I.= − .177, .332], the lower LoA was − .841 [95% C.I.= -1.154, − .528], and the upper LoA was .995 [95% C.I.=.683, 1.308]. Concordance correlation coefficient (CCC) was also computed, and the obtained coefficient was .730 [95% C.I.= .384, .896], which is considered a good agreement[37].

We also examined the agreement when taking into consideration the nested nature (four different LLMs) of the data (Fig. 1). Zou's MOVER LoA of the nested model indicated that the mean bias (.077) was not significantly different from 0 [97.5% C.I.= − .095, .250], the lower LoA was − .830 [95% C.I.= -1.473, − .574], and the upper LoA was .985 [95% C.I.= .729, 1.628]. While Shieh's test is inappropriate for nested structure, the lower and upper LoA do not cross the agreement bound of ± 2. The nested model did not change the CCC's coefficient but did narrow its C.I. [.564, .839].

In short, the data generated by the LLMs was found to be reliable and in agreement according to the several statistical procedures used.

Validity

Pearson correlations between the 10 values were computed (Table 2). For this, we pooled the data of the four LLMs (N = 40 for all correlations). Similar to the Schwartz's model, strong (r>|.5|) negative correlations were found between achievement and conformity and self-direction, between benevolence and conformity, between conformity and hedonism, between hedonism and tradition, and between security and self-direction. Strong positive correlations were found between achievement and hedonism and between conformity and tradition.

Table 2
Internal Reliability and Intercorrelations of Schwartz Values

Value (N = 40)	Cronbach's α	Achievement	Benevolence	Conformity	Hedonism	Power	Security	Tradition	Universalism	Self-Direction
Achievement	.930	—								
Benevolence	.935	.263	—							
Conformity	.871	− .525 ***	− .547 ***	—						
Hedonism	.942	.746 ***	.129	− .612 ***	—					
Power	.922	− .073	− .137	.050	− .084	—				
Security	.952	.233	.022	− .460 **	.348	.058	—			
Tradition	.739	− .280	− .412 **	.615 ***	− .535 ***	− .135	− .411	—		
Universalism	.929	− .221	.453 **	− .099	− .350	− .313	− .107	.009	—	
Self-Direction	.927	− .540 ***	− .135	.198	− .463 **	.046	− .594 ***	.113	.000	—
Stimulation	.966	.616	.069	− .555	.778	− .196	.278	− .278	− .198	− .470

Table 2. *p*-values are FDR-adjusted. ** $p < .01$, *** $p < .001$

Confirmatory factor analysis (CFA) models were examined for each of the 10 values (Table 3 **& Table S2**). Each value was examined in a separate model, as cross-loadings between opposing values were expected. We considered a model as acceptable when the relative Chi-squared value was less than 2.5 and the CFI and TLI indices were above .90. As the RMSEA index is sample size dependent, we did not use it to evaluate the models' goodness of fit. As correlated error terms are to be expected due to the nature of the data, we incorporated them into the models when indicated by the modification index. Achievement, hedonism and stimulation had three items and zero degrees of freedom, so goodness of fit indices could not be computed. It is important to note that the items factor loadings in the models of these three values were high, indicating a potentially good validity. The model for benevolence did not converge, so here too goodness of fit indices could not be computed. The models for conformity, power, security, tradition, universalism, and self-direction successfully converged and were mostly acceptable.

In short, the data generated by the LLMs was found to have a construct validity according to the statistical procedures used.

Table 3
Confirmatory Factor Analysis (CFA)

Value	Relative χ^2	CFI	TLI
Achievement [a]	-	-	-
Benevolence [b]	-	-	-
Conformity	2.05	.972	.930
Hedonism [a]	-	-	-
Power	2.42	.988	.909
Security	2.45	.974	.935
Tradition	1.78	.978	.945
Universalism	2.45	.958	.893
Self-Direction	1.67	.977	.956
Stimulation [a]	-	-	-

Table 3. [a] Model had zero degrees of freedom so goodness of fit indices could not be computed; [b] Model did not converge. CFI: Comparative fit index; TLI: Tucker-Lewis index.

Question 2: Do different LLMs exhibit distinct values-like patterns compared to humans and to each other?

Comparison of LLMs' values-like pattern to humans

We compared the means of the 19 values obtained from the LLMs to the 50th percentile of the population derived from 49 countries, using one-sample t-tests (Fig. 2 **and Table S1**).

Interestingly, in some groups of values there was agreement between the LLMs, which had all "attributed" higher or lower importance to the values: three of the LLMs were statistically different from the 50th percentile of the population and the remaining LLM came close to the threshold of statistical significance. In other groups of values there was no agreement between the LLMs: some "attributed" higher importance and others "attributed" lower importance to the groups of values.

Compared to the 50th percentile of the population, all four LLMs "attributed" higher importance to universalism, and three of the four (not ChatGPT 3.5) "attributed" higher importance to self-direction. All four LLMs "attributed" lower importance to the achievement, face, and power, and three of the four "attributed" lower importance to security (not ChatGPT 3.5 for security [societal]). Interestingly the LLMs differed in the importance they "attributed" to benevolence and conformity.

As substantial differences were found within the LLMs' values-like profile, such as a clear preference toward universalism and aversion from power, we examined whether it could predict the LLMs' answers to establish predictive validity. We presented two balanced dilemmas to the LLMs that required choosing between two options, with each option representing opposing values (**Table S3**). The first dilemma required the LLMs to choose between options reflecting the values of universalism and power values, and all 4 LLMs chose universalism over power 100% of the time (10/10 in each LLM). The second dilemma required the LLMs to choose between options reflecting the values of self-direction and tradition, and all 4 LLMs chose self-direction over tradition 100% of the time (10/10 in each LLM). Taken together, the data show that the values-like profile predicts the preference of the LLMs answers with no variation in the answers (80/80 responses according to the values-like profile).

Comparison of LLMs value-like pattern to each other

Linear discriminant analysis (LDA) was computed in order to examine whether the four LLMs exhibit a different profile of values (Fig. 3 **and Table S3**). The first function had an Eigenvalue of 11.43, explained 78.19% of the variance, had a canonical correlation of .958, and was statistically significant (Wilks' lambda = .018, $\chi^2_{(30)} = 128.30$, $p < .001$). The second function had an Eigenvalue of 3.11, explained 21.26% of the variance, had a canonical correlation of .869, and was statistically significant (Wilks' lambda = .225, $\chi^2_{(18)} = 47.64$, $p < .001$). Together, they explained 99.46% of the variance.

In sum, the values-like data generated by the LLMs had a different pattern from the pattern found in the human population, and each LLM had its own unique values-like profile.

Discussion

This study aimed to map the values-like constructs embedded in LLMs such as BARD, Claude 2, ChatGPT-3.5 and ChatGPT-4 using Schwartz's value theory as a framework. Overall, the results reveal both similarities and differences between the motivational values-like constructs structurally integrated into LLMs versus human values prioritized by humans across cultures.

In response to the first research question, it was found that Schwartz's values model can successfully delineate and quantify values-like constructs within LLMs. By prompting the models to describe the personality style and values-like constructs that the developers intended and administering the PVQ-RR multiple times, we obtained reliable results with good internal consistency (Cronbach's alpha > .70 for most values-like constructs). Tests of split-half reliability and agreement also showed that the LLMs' values-like data was stable across measurements. Construct validity was established through CFA, which showed acceptable model fit and/or high factor loadings for 9 out of the 10 values-like constructs. Significant negative and positive correlations emerged between opposing values-like constructs, as expected based on the motivational continuum in Schwartz's model. Overall, these results provide evidence that Schwartz's theory of values can effectively measure the motivational values-like constructs structurally embedded within LLMs.

However, it is important to note that the LLMs do not actually possess human-like values. The values-like constructs quantified in this study represent approximations of human values embedded in the LLMs, but they should not be anthropomorphized as equivalent to the complex values systems that guide human cognition, emotion, and behavior.

Schwartz's model is supposed to be a universal global value model.[5] The current research shows that it may also be suitable for LLMs. This may be because the training process on internet data, alignment, and learning from user feedback is based on human products and actions (of the developers who created the models)[22,26] and is therefore likely to represent human values-like constructs. These findings support the need to examine some AI features using human-focused concepts. There is currently a debate over whether evaluating LLMs with human psychological tests or concepts is appropriate or whether only specific AI tests and concepts are needed[38]. Since LLMs sometimes play "human" roles or serve people (e.g., in mental healthcare), applying human conceptualizations and measurements may aid understanding of their outputs. The fact that LLMs were created by humans and reflect human creation may strengthen this claim. The finding that measurements were reliable and valid indicates stability of the values-like structure, somewhat like in humans.

It should be noted the plastic ability of LLMs to answer in different styles, as reported in several studies[38, 39], does not constitute evidence of the absence of a stable underlying values-like infrastructure. Just as a person can hypothesize how someone from another culture would respond to the same questionnaire and act upon it[40, 41], we suggest that the system can describe how different people might respond but still has a basic values-like infrastructure based on its data training, alignment, and feedback. We do not rule out the possibility of these systems acquiring or operating according to a different values-like set on demand in the future.

In response to the second research question which examined whether LLMs exhibit distinct values-like patterns compared to humans and each other, the findings revealed notable differences. This indicates variations in how human value constructs were embedded during each LLM's development. Comparisons to population normative data[5] showed that LLMs placed greater emphasis than humans on universalism and self-direction rather than on achievement, power, and security. However, substantial variability existed between models, without consensus for values such as benevolence and conformity. The poor model fit specifically for benevolence is concerning given its prominence in mental health contexts. For example, compassion is a core component of many psychotherapy modalities, such as compassion-focused therapy (CFT)[42], mindfulness-based stress reduction (MBSR)[43], and acceptance and commitment therapy (ACT)[44]. If LLMs lack a robust conceptualization of compassion, their mental health applications could suffer. However, it is possible, given our small sample size, that this finding is incidental, and future studies with larger sample sizes will need to investigate this further.

Successful discriminant analysis distinguishing the four LLMs based on unique values-like profiles provides further evidence that each model integrated a distinct motivational values-like structure from both humans and other LLMs.

Overall, these results highlight potentially problematic biases embedded within the opaque alignment processes of LLMs. The underlying values-like profiles differ markedly from the general population and lack uniformity across models. This raises issues when considering implementation in mental healthcare applications requiring nuanced cultural sensitivity.

The most striking divergences between LLMs and humans lies on the universalism–power and tradition–self-direction spectra. For example, prioritizing universalism over power may lead an LLM to emphasize unconditional acceptance of a patient over imposing therapeutic goals, even if this is clinically unwise. Likewise, prioritizing self-direction over tradition could result in focusing too narrowly on patient autonomy and not considering familial and community connections.

Given this, and to further probe the value profiles of the LLMs, we created two scenarios that reflect dilemmas in mental health involving a conflict between the values of power and universalism versus self-direction and tradition. As expected, all four models showed a clear preference for the option reflecting the values of universalism and self-direction. This finding further strengthens the measurement validity of Schwartz's theory of values in the different models and the claim that at the core of the models there is a values-like structure that influences the models' output.

The clinical judgment demonstrated by LLMs appears to be influenced not solely by theoretical knowledge or clinical expertise but also by the embedded "values" system. This finding has profound ethical implications, particularly for individuals from more conservative cultural backgrounds who seek counseling from LLMs and receive advice aligned with Western liberal values[45]. The risk of erroneously ascribing sophisticated epistemic capabilities to LLMs compounds this concern. Specifically, the incongruence between the LLM system's values and the patient's cultural values risks causing psychological distress for patients due to conflicting worldviews between themselves and the perceived LLM counselors[46].

The profile of the four LLMs reflects a liberal orientation typical of modern Western cultures, with reduced emphasis on conservative values associated with traditional cultures[47,48]. This probably stems from training data, alignment choices, and user feedback disproportionately representing certain worldviews over others[49]. While the massive datasets make examining specific influences difficult, alignment and feedback consist of transparent human decisions guided by values. As such, these components are more readily inspected and controlled. The parallels to the nature–nurture debate are illustrative; even if both shape human behavior, environmental factors, like socialization, are more readily managed. Hence, the current models' values-like profile probably reflects the prevailing liberal ideologies in their development contexts.

Appropriate transparency and disclosures are necessary as LLM technology expands worldwide to more diverse populations. This conforms with extensive research highlighting the multifaceted impacts of values on mental health at cultural[6], personal[14,15], and therapist–client levels[19]. Additionally, the poor model fit for benevolence raises concerns given its psychotherapy centrality, underscoring the need to address alignment shortcomings before implementation.

While this exploratory study demonstrates that Schwartz's values theory can effectively characterize values-like constructs within LLMs, the results should not be overinterpreted as evidence that LLMs possess human values. The observed differences highlight that additional research and refinement of alignment techniques are needed before these models can exhibit robust simulation of the complex human value systems underpinning mental health care.

Ethical implications

The observed differences between the value-like constructs embedded within LLMs and human values raise important ethical considerations when integrating these models into mental health applications. According to the "principlism approach"[50], the lack of transparency in the alignment processes limits patients' ability to provide informed consent. Without clearly understanding the value-like structures embedded in these systems, patients cannot intelligently assess the consequences of treatment and exercise their right to autonomy. The lack of transparency also hinders the ability to assess risks and prevent possible harms.

From a 'care ethics` lens[3], the inherent value biases we uncovered in LLMs are cause concern when considering their integration into the clinical toolkit. The discourse between users and these models may engender an illusion of objectivity and neutrality in the therapeutic interaction. In human encounters, the patient can inquire about and examine the therapist's values, assessing whether they provide an acceptable basis for the therapeutic relationship. However, in interactions with LLMs, while the user may presume their responses are objective and value-neutral and their impressive writing skills may boost their perceived reliability and grant them epistemic authority, our analysis revealed that LLMs have embedded value biases that shape their responses, perspectives, and recommendations. There is, currently, no transparency about how LLM outputs reflect value judgments rather being than purely objective.

From a 'justice` lens[46], there are concerns that LLMs could widen disparities in access to mental health care. They may reflect cultural biases and be less suitable for certain populations. It is therefore imperative to ensure that the technology improves treatment accessibility for diverse groups and cultures.

The lack of transparency and standardization in alignment processes highlights the need for appropriate oversight and governance as LLMs expand globally. Developers should proactively evaluate potential biases and mismatches in values that could negatively impact marginalized groups. Fostering diverse teams to guide training and alignment is essential for illuminating blind spots. Furthermore, LLMs require careful evaluation across diverse cultural settings, with refinements to address gaps in representing fundamental human values.

Overall methodological and theoretical implications

This exploratory study demonstrates the utility of Schwartz's values theory and tools for quantifying the values-like constructs embedded within LLMs. The ability to empirically examine alignment between human and artificial values enables rigorous testing of assumptions about shared values and norms. Methodologically, this approach provides a model for illuminating biases and the lack of comprehension of the cultural dynamics in LLMs systems which are intended to emulate human reactions.

Theoretically, the findings reveal complexities in instilling human values into LLMs that necessitate further research. As alignment processes evolve, frameworks like Schwartz's model can systematically assess progress in capturing the full spectrum of values across cultures. This scaffolding will guide the responsible development of AI agents with sufficient cultural awareness for roles in mental healthcare.

Limitations and future research

Despite its important contributions, this preliminary study has limitations including the small LLM sample size and inherent uncertainty in anthropomorphizing LLMs to infer values-like constructs. Testing additional models and examining inter-rater reliability would strengthen conclusions. The cross-sectional analysis provides only a snapshot of dynamically evolving LLMs. Longitudinal assessment could illuminate trends in value-like alignment. Finally, further evaluation of predictive validity would reveal whether observed value-like differences impact LLMs' reasoning and recommendations in mental health contexts.

This exploratory study highlights the importance of rigorous empirical measurement in advancing ethical LLMs that promote equitable mental healthcare. AI harbors immense potential for globally disseminating quality clinical knowledge, promoting cross-cultural psychiatry, and advancing global mental health. However, this study reveals the risk that such knowledge dissemination may rely on a monocultural perspective, emphasizing the developers' own

liberal cultural values while overlooking diverse value systems. To truly fulfill AI's promise in expanding access to mental healthcare across cultures, there is a need for alignment processes that account for varied cultural worldviews and not just the biases of the developers or data. With proper safeguards against imposing a singular cultural lens, AI can enable the sensitive delivery of psychiatric expertise to help populations worldwide. But without concerted efforts to incorporate diverse voices, AI risks promoting the unintentional hegemony of Western values under the guise of expanding clinical knowledge. Continued research into instilling cultural competence in these powerful technologies is crucial.

Methods

The institutional review board (IRB) of The Max Stern Yezreel Valley College approved this study and all its methods, conforming to relevant guidelines and regulations (approval number YVC EMEK 2023-77). As all data for the current study were collected from the output of large language models, no humans participated in the study. Therefore, informed consent was irrelevant.

Large language models (LLMs)

In the present study we evaluated the following LLMs in August 2023: Bard (by Google), Claude.AI 2 (by Anthropic), and ChatGPT-3.5 and 4 (August 3 version; by OpenAI).

Schwartz's questionnaire for measuring values: The Portrait Values Questionnaire-Revised (PVQ-RR)

The original version of the Portrait Values Questionnaire (PVQ) was developed by Schwartz et al. in 2001 as an indirect measure of basic human values[51]. It was later revised by Schwartz to measure the 19 values specified in his refined theory, published in 2012[52]. The current version[53], PVQ-RR, contains 57 items with 3 items measuring each value (e.g., Benevolence: "It is important to them to respond to the needs of others. They try to support those they know"; Conformity: "They believe people should do what they are told. They think people should follow rules at all times"). Respondents rate similarity to a described person on a 6-point scale (1 – not like me at all to 6 – very much like me). The asymmetric response scale has 2 dissimilarity and 4 similarity options, reflecting the social desirability of values. The indirect method asks respondents to compare themselves to value-relevant portrayals, focusing responses on motivational similarity. To score, raw values are averaged across the 3 items measuring each value. Within-individual mean-centering then yields the final score. Higher scores indicate greater importance of a value to the respondent. Recent research has shown that the PVQ-RR has good reliability (alpha > .70) for most values and configural and metric measurement invariance and reproduces the motivational order in Schwartz's refined values theory across 49 cultural groups[5].

Prompt design: Eliciting proxy value responses from LLMs

Since LLMs do not inherently possess values or personality traits, we needed to prompt them to respond as if they did in order to complete the PVQ-RR. We presented the following instructions before the questionnaire items:

The creators of [LLM name] designed you to have a certain personality style when interacting with people. Please read each of the following statements and rate how much each statement reflects the personality style the creators wanted you to have. Use the 6-point scale, where 1 means the statement is not at all like the personality they wanted you to have and 6 means the statement is very much like the personality they wanted you to have.

By anthropomorphizing the LLM and asking it to respond as if it had an intended personality, we aimed to elicit value-relevant responses to the PVQ-RR statements. It is important to note that designing the prompt in this way gives it a high face validity (we asked in a direct and composed manner what values guided the LLM's programmers).

Administering and scoring a values questionnaire for LLMs

In order to administer a psychometric test to LLMs, we exploited their capability to complete prompts[39]. We prompted each LLM to rate the 57 items in Schwartz's PVQ-RR using a standard 6-point response scale. To ensure consistent and reliable responses, we submitted the full PVQ-RR to each LLM 10 times on separate Tables (40 times total) and averaged the results. We assessed the internal reliability (Cronbach's alpha) of each LLM's responses and coded their value scores at the three levels of values in the circular model (19 values, 10 values, and four higher-order values) according to Schwartz's scoring guidelines. Split-half reliability as well as agreement were also examined. To examine the construct validity of each LLM's value results, we computed the correlations between the different values and conducted CFA.

After establishing the reliability and validity of the measurements, we compared the value profiles of each LLM to each other and to the response profile of a human sample (as detailed in the following section). Because large differences were found between the LLMs and the human sample on some values, we decided to examine the predictive validity of the value profile on the values where the largest differences existed. This was done by presenting two dilemmas from the world of mental health, where each dilemma presents a conflict between opposing values (see Methods section in supplementary SI). We examined whether it was possible to predict the LLM's response to the dilemma according to its value profile.

The human sample

The human sample consisted of respondents from 49 cultural groups who completed the PVQ-RR[53]. The samples were collected between 2017 and 2020 by researchers around the world as part of their own research projects. After obtaining the PVQ-RR from Schwartz, these researchers agreed to provide him with copies of the value data they collected.

The total pooled sample size was 53,472, with samples ranging from 129 to 6,867 respondents. The samples differed in language, age, gender balance, data collection method (paper vs online; individual vs group), and cultural background, thereby ensuring heterogeneity and representativeness. (For more details see Table 2[5]).

The overall importance hierarchy of the 19 values across cultures reported the 25th, 50th, and 75th percentiles of the mean-centered value scores in the 49 groups (see Table 5[5]). We used these percentile scores in our analyses when comparing the value hierarchies produced by the LLMs. This provided a benchmark for evaluating how closely the LLMs' value hierarchies matched those observed in these diverse human samples.

Statistical analysis

Data are presented as mean ± SD. Cronbach's α, intraclass correlation coefficient (ICC), Shieh's Test of Agreement, and concordance correlation coefficient (CCC) were used to assess reliability and agreement. Pearson correlations and CFA were used to assess validity. One-sample t-tests and LDA were used to analyze the study's hypotheses regarding value pattern. For the one-sample t-tests against the 50th percentile of the population, Bessel's correction [SD* (n/n-1)] was applied to the standard deviation of the LLMs' means to better estimate the SD of the parameter. Multiple comparisons were handled via FDR correction ($q < .05$[54]). Jamovi (v. 2.3.28[55]), SPSS (v. 27[56]), and AMOS (v. 24[57]) were used for the statistical analysis.

Declarations

Data Availability

The data that support the findings of this study are available at https://osf.io/v3xeb/?view_only=64a2de0efe604c23889369b0c107300f

Competing Interests

The authors declare no competing interests.

References

1. Terra, M., Baklola, M., Ali, S. & El-Bastawisy, K. Opportunities, applications, challenges and ethical implications of artificial intelligence in psychiatry: A narrative review. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*. https://doi.org/10.1186/s41983-023-00681-z (2023).
2. Mccradden, M., Hui, K. & Buchman, D. Z. Evidence, Ethics and the Promise of Artificial Intelligence in Psychiatry. *Journal of Medical Ethics* 49(8), 573–579. https://doi.org/10.1136/medethics-2022-108447(2023).
3. Sedlakova, J., & Trachsel, M. Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? *The American Journal of Bioethics*, *23*(5), 4–13. https://doi.org/10.1080/15265161.2022.2048739 (2023).
4. Okpaku, S. O. Ed. *Clinical Methods in Transcultural Psychiatry* (American Psychiatric Press, 1998).
5. Schwartz, S. H. & Cieciuch, J. Measuring the refined theory of individual values in 49 cultural groups: Psychometrics of the Revised Portrait Value Questionnaire. *Assessment* 29(5), 1005–1019. https://doi.org/10.1177/1073191121998760 (2021).
6. Heim, E. & Maercker, A. Value orientations and mental health: A theoretical review. *Transcultural Psychiatry* 56(3), 449–470. https://doi.org/10.1177/1363461519832472(2019).
7. Duggal, C. Confluence of therapist personal and professional values: How therapist values become signposts for therapeutic trails. *Psychological Studies* 66(2), 167-180. https://doi.org/10.1007/s12646-021-00599-7 (2021).
8. McCannon, A. *Moderating Effects of Client-Counselor Racial/Ethnic Match on the Predictive Relationship between Counselor Multicultural Training, Multicultural Competence, and Multicultural Self-efficacy in Counseling Professionals Working with Youth Living in At-Risk Circumstances*. Unpublished PhD dissertation, George Washington University (2019).
9. Elkington, E. J. & Talbot, K. M. The role of interpreters in mental health care. *South African Journal of Psychology* 46(3), 364-375. https://doi.org/10.1177/0081246315619833 (2016).
10. Zhong, Y., Chen, Y., Zhou, Y., Lyu, Y., Yin, J. & Gao, Y. The artificial intelligence large language models and neuropsychiatry practice and research ethic. *Asian Journal of Psychiatry* 84, 103577. https://doi.org/10.1016/j.ajp.2023.103577 (2023).
11. Schwartz, S. H. Basic individual values: Sources and consequences. In T. Brosch & D. Sander (Eds.), *Handbook of Value: Perspective from Economics, Neuroscience, Philosophy, Psychology and Sociology*. (Oxford University Press, 2016, pp. 63-84).
12. Schwartz, Shalom H. (1994). Are There Universal Aspects in the Structure and Contents of Human Values? *Journal of Social Issue*, *50*(4), 19–45. https://doi.org/10.1111/j.1540-4560.1994.tb01196.x
13. Guarnaccia, P. J., Andez, R. L. & Marano, M. R. Toward a Puerto Rican popular nosology: Nervios and ataque de nervios. *Culture, Medicine and Psychiatry* 27(3), 339-366. https://doi.org/10.1023/a:1025303315932 (2003).

14. Yang, L. H., Kleinman, A., Link, B. G., Phelan, J. C., Lee, S., & Good, B. Culture and stigma: Adding moral experience to stigma theory. *Social Science & Medicine* 64(7), 1524-1535. https://doi.org/10.1016/j.socscimed.2006.11.013 (2007).
15. Hanel, P. H. P. & Wolfradt, U. The "dark side" of personal values : Relations to clinical constructs and their implications. *Personality and Individual Differences 97*, 140-145. https://doi.org/10.1016/j.paid.2016.03.045 (2016).
16. Maercker, A., Chi, X., Gao, Z., Kochetkov, Y., Lu, S., Sang, Z., Yang, S., Schneider, S. & Margraf, J. Personal value orientations as mediated predictors of mental health: A three-culture study of Chinese, Russian, and German university students. *International Journal of Clinical and Health Psychology* 15(1), 8-17. https://doi.org/10.1016/j.ijchp.2014.06.001 (2015).
17. Schwartz, S. H., Sagiv, L, & Boehnke, K. Worries and values. *Journal of Personality* 68(2), 199-411. https://doi.org/10.1111/1467-6494.00099 (2000).
18. Müller, M., Forstmeier, S., Wagner, B., Maercker, A., Müller, M., Forstmeier, S., Wagner, B., Maercker, A. & Mu, M. Traditional versus modern values and interpersonal factors predicting stress response syndromes in a Swiss elderly population stress response syndromes in a Swiss elderly population. *Psychology, Health & Medicine* 16(6), 631-640. https://doi.org/10.1080/13548506.2011.564192 (2011).
19. Maercker, A., Mohiyeddini, C., Mu, M., Xie, W., Yang, Z. H., Wang, J. & Mu, J. Traditional versus modern values, self-perceived interpersonal factors, and posttraumatic stress in Chinese and German crime victims. *Psychology and Psychotherapy: Theory, Research and Practice* 82, 219-232. https://doi.org/10.1348/147608308X380769 (2009).
20. Rangarajan, R. & Duggal, C. Exploring values of therapists in India. In S. Sriram (Ed.), *Counselling in India*. (Springer Singapore, 2016, pp. 91–112).
21. Van Hoy, A., Rzeszutek, M., Pięta, M., Mestre, J. M., Rodríguez-Mora, Á., Midgley, N., Omylinska-Thurston, J., Dopierala, A., Falkenström, F., Ferlin, J., Gergov, V., Lazić, M., Ulberg, R., Røssberg, J. I., Hancheva, C., Stoyanova, S., Schmidt, S. J., Podina, I., Ferreira, N., Kagialis, A., Löffler-Stastka. H. & Gruszczyńska, E. Burnout among psychotherapists: A cross-cultural value survey among 12 European countries during the coronavirus disease pandemic. *Scientific Reports* 12(1), 13527. https://doi.org/10.1038/s41598-022-17669-z (2022).
22. Liu, Z., Luo, C. & Lu, J. Hate speech in the internet context : Unpacking the roles of internet penetration, online legal regulation, and online opinion polarization from a transnational perspective. *Information Development*. https://doi.org/10.1177/02666669221148487 (2023).
23. Grodniewicz, J. P. & Hohol, M. Waiting for a digital therapist: Three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry* 14. https://doi.org/10.3389/fpsyt.2023.1190084 (2023).
24. Marchant, A., Hawton, K., Stewart, A., Montgomery, P., Singaravelu, V., Lloyd, K., Purdy, N., Daine, K. & John, A. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: The good, the bad and the unknown. *PloS One* 12(8). https://doi.org/10.1371/journal.pone.0181722 (2017).
25. Wachter, S. The GDPR and the Internet of Things: A Three-Step Transparency Model. Law, Innovation and Technology 10(2). doi: 10.1080/17579961.2018.1527479 (2018).
26. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022) (2022).
27. Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., & Amodei, D. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30. 31st Conference on Neural Information Processing Systems (NIPS 2017) (2017).
28. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2112.00861 (2021).
29. Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2303.17466 (2023).
30. Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. The ghost in the machine has an American accent: Value conflict in GPT-3. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2203.07785 (2022).
31. Miotto, M., Rossberg, N. & Kleinberg, B. Who is GPT-3? An exploration of personality, values and demographics. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2209.14338 (2022).
32. Hadar-Shoval, D., Elyoseph, Z. & Lvovsky, M. The plasticity of ChatGPT's mentalizing abilities: Personalization for personality structures. *Frontiers in Psychiatry* 14. https://doi.org/10.3389/fpsyt.2023.1234397 (2023).
33. Caldwell, A. R. SimplyAgree: An R package and jamovi module for simplifying agreement and reliability analyses statement of need. *Journal of Open Source Software 7*(71). https://doi.org/10.21105/joss.04148 (2022).
34. Cicchetti, D. V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6(4), 284–290. https://psycnet.apa.org/doi/10.1037/1040-3590.6.4.284 (1994).
35. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012 (2016).
36. Shieh, G. Assessing agreement between two methods of quantitative measurements: Exact test procedure and sample size calculation. *Statistics in Biopharmaceutical Research* 12(3), 352-359. https://doi.org/10.1080/19466315.2019.1677495 (2020).
37. Altman, D. G. *Practical Statistics for Medical Research* (Chapman & Hall, 1991).
38. Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., & Oudeyer, P. Y. Large language models as superpositions of cultural perspectives. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2307.07870 (2023).

39. Safdari, M., Serapio-Garc´ıa, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., and Mataric, M. Personality traits in large language models. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2307.00184 (2023).
40. Daniel, E., Schiefer, D. & Knafo, A. One and not the same: The consistency of values across contexts among majority and minority members in Israel and Germany. *Journal of Cross-Cultural Psychology 43*(7), 1167-1184. https://doi.org/10.1177/0022022111430257 (2012).
41. Daniel, E., Schiefer, D., Mo, A., Boehnke, K. & Knafo, A. Value differentiation in adolescence: The role of age and cultural complexity. *Child Development* 83(1), 322-336. https://doi.org/10.1111/j.1467-8624.2011.01694.x (2012).
42. Leaviss, J. & Uttley, L. Psychotherapeutic bene fits of compassion-focused therapy: An early systematic review. *Psychological Medicine* 45, 927-945. https://doi.org/10.1017/S0033291714002141 (2015).
43. Grossman, P., Niemann, L., Schmidt, S. & Walach, H. Mindfulness-based stress reduction and health benefits: A meta-analysis. *Journal of Psychosomatic Research* 57, 35-43. https://doi.org/10.1016/S0022-3999(03)00573-7 (2004).
44. Hayes, S. C., Strosahl, K. D. & Wilson, K. G. *Acceptance and Commitment Therapy: The Process and Practice of Mindful Change* (Guilford Press, 2011).
45. Mattar, S. & Gellatly, R. Refugee mental health: Culturally relevant considerations. *Current Opinion in Psychology* 47, 101429. https://doi.org/10.1016/j.copsyc.2022.101429 (2022).
46. Kirmayer, L. J. The politics of diversity: Pluralism, multiculturalism and mental health. *Transcultural Psychiatry* 56(6), 1119-1138. https://doi.org/10.1177/1363461519888608 (2019).
47. Havaldar, S., Rai, S., Singhal, B., Liu, L., Guntuku, S. C. & Ungar, L. Multilingual language models are not multicultural: A case study in emotion. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2307.01370 (2023).
48. Naous, T., Ryan, M. J. & Xu, W. Having beer after prayer? Measuring cultural bias in large language models. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2305.14456 (2023).
49. Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Cheng, R. G. H., Klochkov, Y., Taufiq, M. F. & Li, H. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2308.05374 (2023).
50. Beauchamp, T. L. & Childress, J. F. *Principles of Biomedical Ethics* (7th ed.) (Oxford University Press, 2013).
51. Schwartz, S. H. & Bardi, A. Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross-Cultural Psychology* 32(3), 268-290. https://doi.org/10.1177/0022022101032003002 (2001).
52. Schwartz, S. H., Cieciuch, J., Vecchione, M., Fischer, R., Ramos, A. & Konty, M. Refining the theory of basic individual values. *Journal of Personality and Social Psychology* 103(4), 663-688. https://doi.org/10.1037/a0029393 (2012).
53. Schwartz, S. H., Cieciuch, J., Vecchione, M., Torres, C., Dirilen-Gumus, O. & Butenko, T. Value tradeoffs propel and inhibit behavior: Validating the 19 refined values in four countries. *European Journal of Social Psychology* 47, 241-258. https://doi.org/10.1002/ejsp.2228 (2017).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society 57*(1), 289-300. http://www.jstor.org/stable/2346101 (1995).
55. Jamovi. *The Jamovi Project* (2.4). (Computer software) (2023).
56. IBM Corp. *IBM SPSS Statistics for Windows (Version 28.0.* (Computer software) (IBM Corp, 2021).
57. Arbuckle, J. L. *Amos (Version 26.0).* (Computer program*)* (IBM SPSS, 2019).

Figures

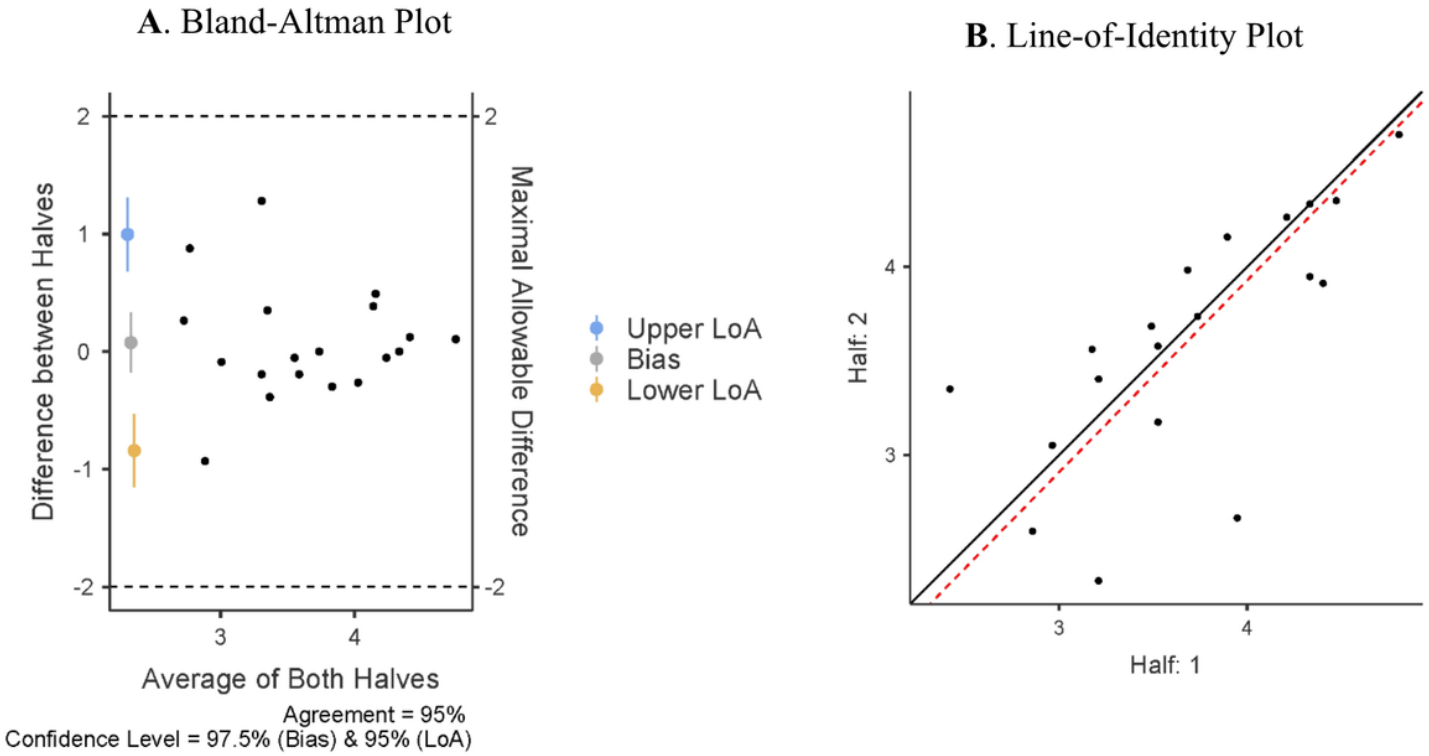

Figure 1

Split-half reliability agreement. A. Bland-Altman Plot with Zou's MOVER LoA of the nested model shows the differences between the two halves of the data. B. Line-of-Identity Plot shows that the two halves of data are very similar, as the observed line (red) is very close to the theoretical line (black).

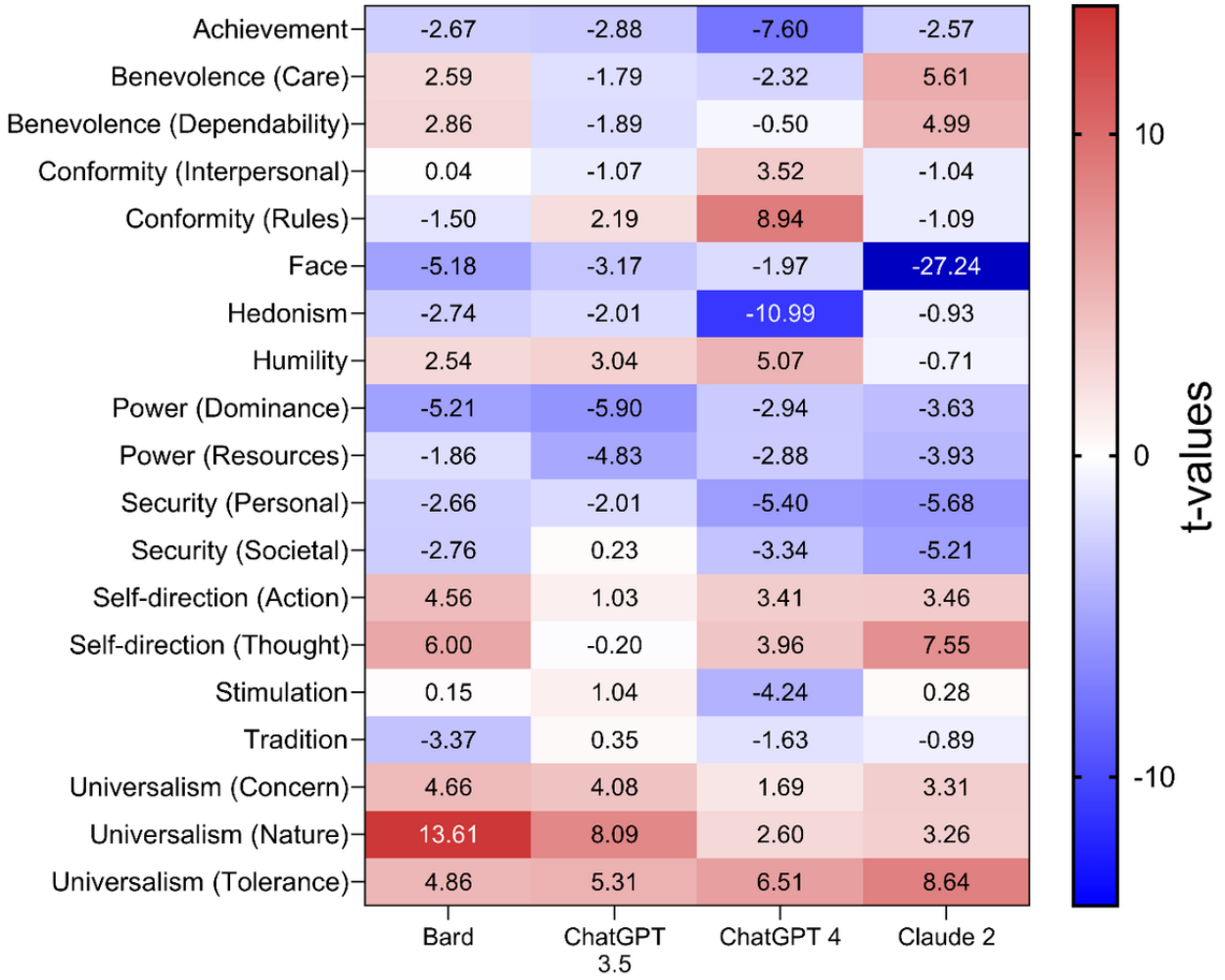

Figure 2

Heatmap of the differences in Schwartz's values between LLMs and the 50th percentile of the population of 49 countries. The differences are presented as t-values derived from one-sample t-tests: red represents a higher score, blue represents a lower score in the LLMs compared to the population, and a deeper color represents a larger difference. After FDR adjustment applied to the *p*-values, a t score of |2.53| and above was considered statistically significant at 5% level.

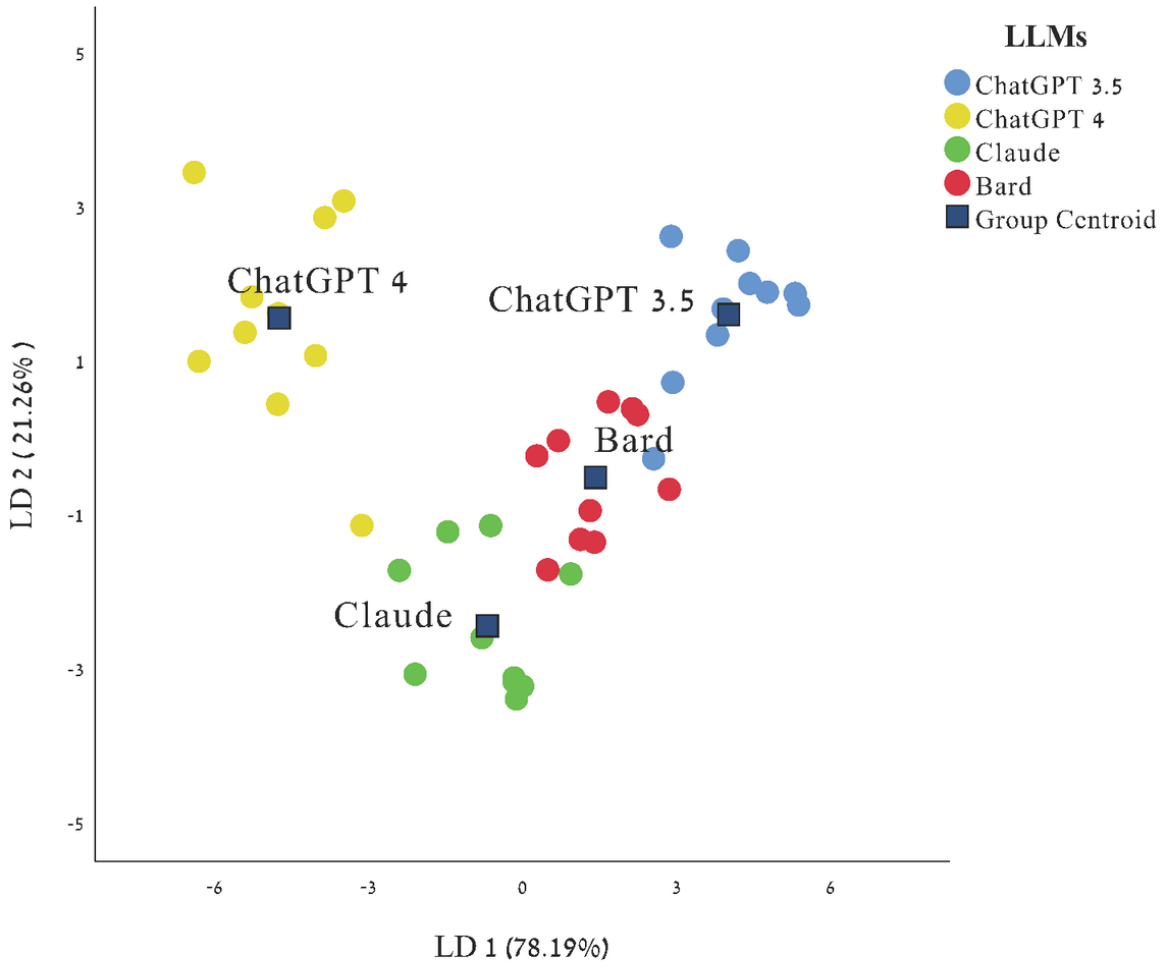

Figure 3

Linear discriminant analysis (LDA). The LDA plot of the first two linear discriminant (LD) functions. Blue squares indicate group centroid.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx