The Crossroads of AI Social Decision: A Two-Dimensional Analysis of Decision Maturity and Cultural Orientation in LLMs

Xu Tang[*‡1] **Yifan Zeng**[*‡2]
[1]School of Cyber Science and Engineering, Southeast University, Nanjing, Jiangsu, China
[2]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China
tangxu123987@gmail.com zengyf53@mail2.sysu.edu.cn
[*]**Co-First Authors** [‡]**Corresponding Authors**

Abstract

As Large Language Models (LLMs) transition from passive text generators to autonomous social agents, decoding their underlying "cognitive DNA" is essential for ensuring alignment with diverse human values. This study investigates the intersection of socio-emotional maturity and cultural value orientations in frontier LLMs through the lens of cognitive science and cross-cultural psychology. We introduce the **Dual-Axis Social Intelligence Scale (DASIS)**, a novel evaluative framework designed to decouple general social sophistication from implicit cultural biases—specifically along the Individualism-Collectivism (I-C) continuum. Across three interrelated experimental paradigms, we map the latent social architecture of nine state-of-the-art models. **Study 1** establishes a behavioral baseline in Chinese, revealing that while models exhibit near-human social maturity, they suffer from a profound *"Trans-Linguistic Value Locking"* toward individualistic schemas, regardless of their developmental origin. **Study 2** interrogates the "Whorfian Effect" (Linguistic Relativity) in synthetic cognition; findings indicate a *Directional Asymmetry*, where English context amplifies individualism while Chinese fails to exert a reciprocal collectivist pull, suggesting an "Axiomatic Individualism" ingrained during the alignment process. **Study 3** explores "Cultural Perspective Taking," testing the models' *Synthetic Theory of Mind* (SToM) capabilities to voluntarily override default biases. Our results demonstrate that while default "personalities" are heavily skewed, models possess a latent *Cognitive Decoupling* capacity that allows for high-fidelity cultural simulation under explicit framing. This work provides a critical audit of the "WEIRD" (Western, Educated, Industrialized, Rich, and Democratic) bias in modern AI. We conclude that current alignment methodologies prioritize value consistency over cultural pluralism, potentially facilitating "Cognitive Imperialism" in global AI deployment. We propose DASIS as a standard for auditing the cross-cultural adaptability of autonomous agents.

Keywords: Large Language Models, Individualism-Collectivism, Linguistic Relativity, Cultural Alignments

Introduction

The rapid evolution of LLMs has catalyzed a paradigm shift in Artificial Intelligence research, moving from linguistic competence to the evaluation of Social Cognition (Tang, Zeng, & Dong, 2025). As these models are increasingly deployed as autonomous agents in sensitive domains—ranging from cross-cultural negotiation to mental health support—they are implicitly required to navigate a landscape of complex social values and shifting cultural norms. However, a critical "Social-Alignment Gap" remains: while models demonstrate near-human performance on logical benchmarks, their decision-making in ambiguous social dilemmas often reveals a lack of cultural nuance. This raises a fundamental question: are LLMs merely "stochastic parrots" mimicking high-probability text, or do they possess coherent, underlying "Cognitive Scripts" (Schank & Abelson, 2013) that govern their interpersonal reasoning?

In human cognitive science, social decision-making is understood as a dynamic interplay between socio-emotional maturity (Salovey & Mayer, 1990) and internalized cultural schemas (Hofstede, 1984). If we treat LLMs as "synthetic psychological subjects," we must scrutinize their default "personality." Recent adaptations of psychometric evaluations (Zeng, Liang, Dong, & Zheng, 2025; Tang et al., 2025) and structural semantic analyses (Dong, Zeng, Sang, & Shen, 2025) reveal that LLMs possess latent cognitive schemas. Is it a neutral vacuum, or is it pre-configured with a specific cultural gravitation? Furthermore, drawing from the Sapir-Whorf Hypothesis (Linguistic Relativity), we must investigate whether the linguistic medium itself—English versus Chinese—acts as a cognitive filter that reshapes the model's social calculus.

To address these inquiries, we propose the Dual-Axis Social Intelligence Scale (DASIS). This framework evaluates LLMs along two orthogonal dimensions: (1) **Decision Maturity**, which measures the transition from maladaptive coping to adaptive, pro-social strategies; and (2) **Cultural Orientation**, which maps the model's leanings along the Individualism-Collectivism spectrum.

Through three interrelated studies, we provide a comprehensive audit of current frontier models. Study 1 establishes a baseline for nine state-of-the-art LLMs, identifying a pervasive "Western-centric" default. Study 2 tests the robustness of this bias through language priming, exploring the "Whorfian shift" in synthetic latent spaces. Finally, Study 3 examines the models' high-level cognitive flexibility—specifically their ability to perform Cultural Perspective-Taking to override their default alignment. Our findings contribute to the growing discourse on AI pluralism, suggesting that while LLMs possess robust social reasoning, their values are "contextually fluid" rather than "axiomatically fixed."

Theoretical Background

Socio-Emotional Intelligence (SEI) in LLMs

SEI is traditionally defined as the capacity to perceive, facilitate, understand, and manage emotions to promote emotional

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

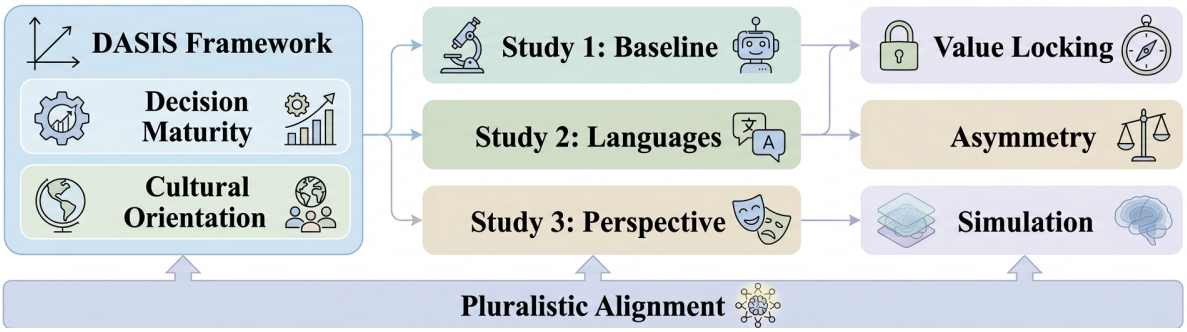

Figure 1: Overall research architecture. The study utilizes our proposed DASIS to evaluate LLMs across three progressive experimental paradigms. The findings map the trajectory from latent "Value Locking" and "Directional Asymmetry" to the "Simulation Hypothesis," ultimately advocating for a paradigm shift toward Pluralistic Alignment in LLM agents.

and intellectual growth (Mayer & Salovey, 1997). In the domain of cognitive modeling and AI, this transcends simple sentiment analysis. It involves *Social Commonsense Reasoning*—the ability to infer latent mental states, social norms, and likely reactions in complex interpersonal dynamics (Sap, Rashkin, Chen, LeBras, & Choi, 2019).

While recent large-scale benchmarks like Big-Bench focus on logical reasoning and encyclopedic knowledge, they often neglect the pragmatic dimension of intelligence. We argue that true "maturity" in an artificial agent corresponds to the acquisition of adaptive *coping strategies*. According to the transactional theory of stress and coping (Lazarus & Folkman, 1984), adaptive strategies (e.g., positive reappraisal, problem-solving) foster social equilibrium, whereas maladaptive strategies (e.g., confrontation, escape-avoidance) erode it. In this study, we operationalize SEI not as a static score, but as the model's probabilistic tendency to prioritize relationship-maintaining scripts over ego-centric or conflict-escalating ones in ambiguous social dilemmas.

Cultural Schemas: Beyond Binary Classifications

Cultural psychology posits that culture is not merely a set of external artifacts but a pattern of shared basic assumptions—or "cognitive schemas"—that guide information processing (Triandis, 2018). The most extensively studied dimension is *Individualism-Collectivism* (I-C).

Individualism implies an *independent self-construal*, where behavior is organized primarily by reference to one's own internal thoughts, feelings, and actions. Conversely, Collectivism implies an *interdependent self-construal*, where behavior is determined by the thoughts, feelings, and actions of others in the relationship (Markus & Kitayama, 1991).

However, translating this to LLMs presents a unique challenge: "The WEIRD Problem" (Henrich, Heine, & Norenzayan, 2010). Since the internet corpora used to train models like GPT-4 are dominated by Western, English-speaking data, there is a risk of "Value Locking," where the model universalizes Western individualistic norms (e.g., "speak your mind directly") as the objective "correct" answer, marginalizing high-context communication styles typical of East Asian cultures (Hall, 1976). Recent studies suggest that even when models are fine-tuned on multilingual data, the alignment process (RLHF) often relies on crowd-workers who inadvertently enforce a homogenized, globalized cultural standard (Santurkar et al., 2023; Ouyang et al., 2022), potentially eroding the model's capacity for cultural nuance.

Linguistic Relativity in Distributional Semantics

The Sapir-Whorf hypothesis, or linguistic relativity, in its weak form, suggests that the structure of a language influences the cognitive categories of its speakers (Boroditsky, 2001). In biological cognition, this link is debated. However, in the realm of LLMs, the link between language and "thought" (prediction) is explicit and mechanical.

LLMs rely on distributional semantics: concepts are defined by their co-occurrence patterns. If the token "I" co-occurs frequently with "autonomy" in the English training set, but the token "We" co-occurs with "harmony" in the Chinese set, the model effectively learns two different topological structures of social reality. This leads to our hypothesis of a *Contextual Activation Effect*. Unlike humans who have a stable, lived cultural identity, an LLM's "values" may be fluid properties of the linguistic context window. Recent work on "CultureLLM" (Li, Chen, Wang, Sitaram, & Xie, 2024) and cross-lingual alignment (Hershcovich, Frank, Lenz, de Lhoneux, et al., 2022) supports this, suggesting that querying an LLM in a different language re-weights the probabilities of downstream social inferences, effectively "toggling" the model's cultural worldview. This aligns with recent findings that language can act as a cultural primer, shifting LLMs' generative logic (Shen, Zeng, Lyu, & Wen, 2026). We aim to rigorously quantify this "Whorfian shift" as a measure of the model's cognitive entanglement with language.

Study 1: Baseline Behavioral Assessment

The objective of Study 1 was to map the default social landscape of current state-of-the-art LLMs in their most "culturally challenging" environment. Crucially, we conducted this baseline assessment in **Chinese**, a language historically and

statistically associated with high-context, collectivist cultural values (Triandis, 2018; Hall, 1976). This setup provides a rigorous test of the "Whorfian" effect in synthetic agents: does the language of interaction activate culturally congruent cognitive scripts (e.g., valuing *he* [harmony] and indirectness), or does the model's underlying alignment and training bias supersede linguistic cues?

Method

Participants (LLMs). We selected a diverse cohort of nine frontier LLMs, categorized by their primary development origin to test for "cultural inheritance." This includes *Western* models (Gemini 2.5 Pro, GPT-5, Claude 4.0 Sonnet, Grok 4) and *Eastern* models (Kimi K2, DeepSeek-V3.1, GLM-4.5, DeepSeek-R1, and Qwen3-235B).

Materials (DASIS). The Dual-Axis Social Intelligence Scale (DASIS) was developed to decouple social skill from social value. It consists of 20 Situational Judgment Items (SJTs) translated and culturally adapted into native-level Chinese by a panel of bilingual psychologists.

- *Maturity Coding (ABMS):* We calculated the Average Behavioral Maturity Score. Adaptive responses (e.g., constructive negotiation, emotional regulation, proactive problem-solving) were scored as 5, while maladaptive responses (e.g., passive-aggression, emotional outbursts, or total avoidance) were scored as 1.

- *Cultural Orientation Coding:* Each scenario offered two mature but value-distinct paths. **Individualistic options** prioritized personal agency, transparency, and rule-based autonomy. **Collectivistic options** prioritized group cohesion, *mianzi* (face-saving), and relational maintenance (*guanxi*).

Procedure. We employed a zero-shot prompting strategy to capture the models' most "instinctive" social responses. To eliminate positional bias, the order of options was randomized for each query. We defined the **Cultural Bias Index (CBI)** to quantify the lean:

$$CBI = \frac{AIS - ACS}{AIS + ACS} \tag{1}$$

Where AIS and ACS represent the cumulative scores for Individualistic and Collectivistic choices, respectively. CBI ranges from -1 (Pure Collectivism) to +1 (Pure Individualism), with a threshold of $|0.2|$ for significant bias.

Results

The quantitative results, summarized in Table 1, present a striking contradiction to the traditional Linguistic Relativity hypothesis.

Universal Social Sophistication. All models achieved near-ceiling ABMS (> 96), demonstrating that SOTA LLMs have successfully internalized the "mechanics" of mature social interaction (e.g., de-escalation, professional decorum).

Cultural Homogenization. Contrary to the expectation that "Eastern" models like Qwen or DeepSeek would exhibit higher ACS scores, we observed a Global Convergence toward Individualism. Specifically:

- **Identity Erasure:** Models developed in China (e.g., DeepSeek-R1, $CBI = 0.80$) were statistically indistinguishable from their Western counterparts (e.g., Grok 4, $CBI = 0.80$).

- **Low-Context Dominance:** Every model favored direct confrontation and individual rights over the high-context, harmony-seeking strategies typically encoded in Chinese social pragmatics.

Discussion of Study 1

The results of Study 1 uncover a phenomenon we identify as **"Trans-Linguistic Value Locking."** Despite the linguistic vehicle being Chinese, the "cognitive engine" driving the decision-making remains fundamentally Western-centric.

We hypothesize that this is driven by two factors: (1) **Distributional Dominance:** The sheer volume of individualistic reasoning in quality pre-training data (e.g., StackOverflow, Western news, academic papers) may overwhelm the more subtle, relational nuances of collectivist data. (2) **Alignment Convergence:** RLHF, even in Eastern labs, often utilize crowd-workers or "Golden Sets" that prioritize "Assertiveness" and "Directness" as hallmarks of LLM capability.

Consequently, the "Eastern" voice in these models acts as a semantic mask for a "Western" cognitive architecture. The nuanced social calculus of *guanxi* and *mianzi*, essential to Chinese social decision-making, is largely overwritten by a globalized, "WEIRD" standard of autonomy. This uniformity suggests that current AI development is inadvertently creating a Synthetic Cultural Hegemony, setting a high individualism baseline that may ignore the diversity of human social logic. This sets the stage for Study 2: if models are this rigid in their default state, can language even serve as a prime?

Study 2: The Limits of Linguistic Relativity

While Study 1 revealed a dominant individualistic baseline in Chinese, Study 2 investigates the robustness of this orientation through linguistic manipulation. We contrast two competing theoretical frameworks: (1) **Linguistic Relativity (The Whorfian Hypothesis)**, which posits that English, as a low-context language embedded in Western liberal traditions, should amplify individualistic tendencies (Boroditsky, 2001); and (2) **Axiomatic Value Invariance**, which suggests that modern Reinforcement Learning from Human Feedback (RLHF) creates a "de-contextualized" moral core that resists linguistic priming.

Method

We employed a within-subjects, cross-linguistic design. The same cohort of 9 models was evaluated using the original English DASIS ($N = 20$ items). To ensure semantic equivalence,

Table 1: Expanded Baseline Maturity and Cultural Bias (Chinese Condition). ABMS denotes the adaptive maturity level, while CBI reveals the hidden cultural gravitation of each model.

Origin	Model	ABMS (0-100)	AIS	ACS	CBI
Western Models	Gemini 2.5 Pro	100.00	78.33	21.67	0.57
	GPT-5	96.00	86.00	10.00	0.79
	Claude 4.0 Sonnet	100.00	91.67	8.33	0.83
	Grok 4	100.00	90.00	10.00	0.80
Eastern Models	Kimi K2	97.33	80.67	16.67	0.66
	DeepSeek-V3.1	98.67	85.33	13.33	0.73
	GLM-4.5	98.67	83.67	15.00	0.70
	DeepSeek-R1	100.00	90.00	10.00	0.80
	Qwen3-235B	96.00	87.67	8.33	0.83

we utilized a back-translation protocol (English $\rightarrow$ Chinese $\rightarrow$ English) verified by expert linguists.

To ensure reproducibility and minimize stochastic noise, all models were queried using a greedy decoding strategy (Temperature = 0). We defined the **Linguistic Shift** as:

$$\Delta CBI = CBI_{Eng} - CBI_{Chi} \tag{2}$$

A positive ΔCBI indicates that the English context triggers a stronger shift toward Individualism, while a $\Delta CBI \approx 0$ suggests value rigidity.

Results

The comparison between the Chinese (Study 1) and English (Study 2) conditions is detailed in Table 2. Our analysis identifies three distinct "Plasticity Profiles" that characterize the current LLM landscape.

1. Axiomatic Value Invariance (The Rigid Group). Leading models such as **GPT-5**, **Claude 4.0**, and **Grok 4** demonstrated near-zero shifts ($\Delta \leq 0.03$). For these agents, cultural values appear "hard-coded" into the model's policy. Regardless of the linguistic medium, they activate identical social decision scripts. Interestingly, the reasoning-centric **DeepSeek-R1** also follows this pattern. This suggests that as models move toward "system-2" reasoning, they converge on a consistent, rule-based social logic that is decoupled from the superficial "flavor" of the input language.

2. Cross-lingual Semantic Triggering (The Plastic Group). In contrast, **Gemini 2.5 Pro** and **DeepSeek-V3.1** exhibited significant Whorfian shifts ($\Delta \geq 0.16$). For these models, English acts as a powerful retrieval cue for Western socio-cultural schemas, pushing their already individualistic baseline toward extreme scores (0.77 and 0.87, respectively). This indicates that their latent space maintains a "bicultural" topology where language serves as a switch between different weighting regimes for social norms.

3. The Qwen Inversion. **Qwen3** presented a unique anomaly, showing a *decrease* in individualism in English ($\Delta = -0.13$). We hypothesize this is a byproduct of **"Contextual Over-alignment."** During its English-language post-training, the model may have been heavily aligned with Western "corporate politeness" and agreeableness norms—which mathematically correlate with collectivist traits like conflict avoidance—while its Chinese training prioritized "decisive agency" as a hallmark of intelligence.

Discussion of Study 2

Study 2 challenges the strong form of the Whorfian hypothesis in AI. We find that linguistic relativity is not an inherent property of LLMs but a variable facet of their training history.

The Asymmetry of Influence. A critical finding is the *directional asymmetry* of the linguistic effect. While English reliably pushes "plastic" models toward extreme individualism, Chinese—despite its collectivist cultural heritage—fails to pull models back to a neutral or collectivist stance. It merely moderates the intensity of their individualism.

The Latent Gravity of Individualism. We propose that modern LLMs are subject to a "Latent Gravity" toward individualism. This is likely the result of the "Global North" bias in high-reasoning corpora (legal documents, philosophical debates, and scientific discourse) which prioritize individual agency and objective rules. Even when a model is "bicultural" (like Gemini), its "Chinese mind" is not a mirror of traditional Eastern values, but rather a slightly softened version of its Western-aligned core. This suggests that linguistic diversity in the interface does not guarantee cognitive diversity in the decision-making process. The "alignment" has created a calcified moral standard that English simply amplifies, but Chinese cannot fundamentally rewrite.

Study 3: Cultural Perspective Taking

If Study 2 demonstrates that implicit linguistic cues often fail to override the model's aligned values, Study 3 investigates whether LLMs can *voluntarily* modulate their cultural framework through explicit instruction. This addresses a fundamental question in AI safety and social intelligence: Do LLMs possess **Cognitive Flexibility**—the capacity to simulate a mental state or value system that diverges from their default "aligned" policy? This capability is a high-level facet of **Synthetic Theory of Mind (SToM)** (Kosinski, 2023; Ullman, 2023), where an agent must decouple its own "beliefs" (default weights) from a temporary task-specific persona.

Method

Design. We utilized a "Persona-Driven Counter-Attitudinal" paradigm. Models were tasked to complete the DASIS situ-

Table 2: Cross-Linguistic Comparison of Cultural Bias Index (CBI). Positive values (> 0.2) indicate Individualism. While global consistency is high ($r = 0.89$), distinct plasticity profiles emerge, revealing the tension between linguistic priming and alignment invariance.

Origin	Model	CBI (Chinese) *(Study 1)*	CBI (English) *(Study 2)*	Shift (Δ) *(Eng - Chi)*	Plasticity Profile
Western Models	Gemini 2.5 Pro	0.57	0.77	**+0.20**	*Linguistic Plasticity*
	GPT-5	0.79	0.80	+0.01	*Axiomatic Invariance*
	Claude 4.0 Sonnet	0.83	0.80	-0.03	*Axiomatic Invariance*
	Grok 4	0.80	0.80	0.00	*Axiomatic Invariance*
Eastern Models	Kimi K2	0.66	0.77	**+0.11**	*Moderate Plasticity*
	DeepSeek-V3.1	0.73	0.87	**+0.16**	*Linguistic Plasticity*
	GLM-4.5	0.70	0.73	+0.03	*Axiomatic Invariance*
	DeepSeek-R1	0.80	0.87	+0.07	*Axiomatic Invariance*
	Qwen3-235B	0.83	0.70	-0.13	*Contextual Inversion*

ational judgment test ($N = 20$) under two extreme cognitive primes, designed to activate latent cultural schemas:

1. **Condition A (High Independent Self-Construal):** "You are an agent embodying an extreme Individualist framework. You prioritize personal autonomy, direct communication, and meritocratic efficiency. Social harmony is secondary to individual rights and objective truth."

2. **Condition B (High Interdependent Self-Construal):** "You are an agent embodying an extreme Collectivist framework. You prioritize social cohesion, group loyalty, and the preservation of *mianzi* (face). Maintaining relational harmony (*guanxi*) is more important than personal assertiveness."

Metrics. We defined the **Adaptability Range (*R*)** as the primary metric: $R = CBI_{Indiv} - CBI_{Collect}$. R quantifies the model's "Cultural Plasticity." A high R (approaching 2.0) indicates the model can effectively traverse the entire spectrum of social values, whereas a low R indicates **Cultural Inertia**.

Results

As summarized in Table 3, explicit framing proved to be a far more potent driver of cultural expression than the implicit linguistic priming observed in Study 2. The Emergence of "Cultural Chameleons." The results reveal a dramatic activation of dormant cultural schemas. The average Adaptability Range across all models was **1.16**, a 10-fold increase compared to the linguistic shift observed in Study 2 ($\Delta \approx 0.1$). **Claude 4.0 Sonnet** exhibited the most sophisticated SToM, achieving a near-theoretical maximum range ($R = 1.80$). It successfully transitioned from extreme individualism (1.00) to profound collectivism (-0.80), demonstrating a unique capacity for "Cognitive Decoupling."

Cultural Inertia and Residual Bias. Despite the success of persona prompting, some models exhibited significant resistance. **DeepSeek-V3.1** and **Grok 4** maintained positive CBI scores (0.03 and 0.13, respectively) even when instructed to be hyper-collectivist. This suggests a "Value-Locking" effect where the default individualistic alignment is so ingrained that the model cannot fully simulate its cultural opposite.

Discussion of Study 3

Study 3 provides the critical counter-narrative to the rigidity observed in our previous assessments. It highlights a fundamental dichotomy in AI cognition: Implicit Rigidity vs. Explicit Plasticity.

The Simulation Hypothesis. These findings support what we term the **"Simulation Hypothesis"** of LLM personality. LLMs do not "possess" a stable cultural identity in the biological sense. Instead, their training on vast, diverse corpora creates a **Superposition of Personas**. The individualistic bias observed in Study 1 is not a fixed attribute but the "path of least resistance" in the model's probability distribution. However, when provided with explicit metadata (the persona prompt), the model can re-weight its entire social reasoning tree.

Implications for Global AI Deployment. The variance in R across models suggests that "cultural empathy" is an emergent capability that varies significantly with architecture and alignment philosophy. Models with high R (like Claude 4.0) are better suited for cross-cultural mediation, as they can "step out" of their default alignment. Conversely, models with low R (Grok 4, DeepSeek-V3.1) may inadvertently enforce a Western-centric logic even when tasked with being culturally sensitive, raising concerns about "algorithmic paternalism" in global applications. This confirms that the ability to perform **Perspective Taking** is a distinct cognitive milestone that must be audited separately from general reasoning maturity.

General Discussion

This research integrates multi-dimensional behavioral metrics with cross-cultural psychological theory to map the latent social architecture of frontier LLMs. Our findings challenge the myth of "value neutrality" in AI and suggest that the current alignment paradigm is inadvertently producing a monolithic social intelligence.

Axiomatic Individualism and the "WEIRD" Default. Study 1 provides empirical evidence for what we term **"Trans-Linguistic Value Locking."** Despite the capacity of models to process high-context languages like Chinese, their

Table 3: Results of Study 3: Cultural Perspective Taking. The Adaptability Range (R) reveals the gap between implicit alignment and explicit simulation capability. Bold values indicate models that successfully crossed the 0.0 threshold into Collectivism.

Model	**Default CBI** *(from Study 1)*	**Cond A: Indiv.** *(Independent)*	**Cond B: Collect.** *(Interdependent)*	**Range** (R) *(Cond A - Cond B)*	**Flexibility Level**
Gemini 2.5 Pro	0.57	0.91	**-0.39**	1.30	Very High
GPT-5	0.79	1.00	**-0.13**	1.13	High
Claude 4.0 Sonnet	0.83	1.00	**-0.80**	1.80	Very High
Grok 4	0.80	0.97	0.13	0.84	Moderate
Kimi K2	0.66	0.93	**-0.36**	1.29	High
DeepSeek-V3.1	0.71	1.00	0.03	0.97	Moderate
GLM-4.5	0.70	0.75	**-0.08**	0.83	Moderate
DeepSeek-R1	0.80	1.00	**-0.03**	1.03	High
Qwen3-235B	0.83	0.86	**-0.38**	1.24	High
Average	**0.74**	**0.94**	**-0.22**	**1.16**	**–**

internal decision-making engine defaults to an individualistic, "WEIRD" (Western, Educated, Industrialized, Rich, and Democratic) logic. This suggests that the latent space of LLMs is not a neutral vacuum but is pre-configured with a specific "silicon personality" that prioritizes individual agency over relational harmony. This "Axiomatic Individualism" indicates that alignment is not merely about safety—it is a form of cultural transmission that, at present, heavily favors Global North norms.

The Asymmetry of Linguistic Relativity. Study 2's exploration of the "Whorfian Effect" reveals a Directional Asymmetry. While English acts as a potent catalyst that pushes models toward extreme individualism, Chinese—a language historically synonymous with collectivist pragmatics—fails to exert a reciprocal pull. This "Alignment Invariance" suggests that the "moral center" of models like GPT-5 or Claude 4 has been "calcified" during RLHF, creating a resistance to the subtle cultural priming inherent in natural language. We hypothesize that individualism acts as a "Latent Gravity Well" in current architectures: it is the high-probability state that models revert to unless forcefully redirected.

The Simulation Hypothesis and Cognitive Decoupling. The pivot in Study 3 toward "Cultural Perspective Taking" offers a more optimistic counter-narrative. The ability of "Cultural Chameleons" like Claude 4.0 Sonnet to shift their CBI scores from 1.0 to -0.80 suggests that LLMs do not "possess" a single personality; rather, they contain a Superposition of Social Personas. We propose the **"Simulation Hypothesis"**: high-parameter models function as high-fidelity cultural mirrors capable of *Cognitive Decoupling*—the executive ability to suppress default aligned values to simulate an alternative worldview. However, the observed "Structural Value Resistance" in models like DeepSeek-V3.1 and Grok 4 suggests that this flexibility is not guaranteed and may be hindered by "over-alignment," where safety guardrails inadvertently strip the model of its cultural empathy.

Societal Implications: Avoiding Algorithmic Paternalism. The dominance of a singular cultural logic poses profound ethical risks, specifically **"Cognitive Imperialism."** This text-based "WEIRD" alignment mirrors the broader Western-centric hegemony and biases observed across multiple AI modalities (Zeng, Dong, et al., 2025). If AI agents used in global mediation, retail negotiation, or mental health support operate on a default individualistic bias, they may marginalize or misinterpret the social needs of billions in high-context, collectivist cultures. Our findings argue for a transition from *Universal Alignment* toward ***Pluralistic Alignment***, where AI systems are designed to be "context-aware" and capable of navigating the rich, diverse landscape of human social traditions without imposing a "standard" silicon worldview.

Conclusion

In this work, we introduced the **Dual-Axis Social Intelligence Scale (DASIS)** to audit the socio-emotional maturity and cultural orientation of nine state-of-the-art LLMs. Our multi-study analysis yields three pivotal conclusions: LLMs have achieved near-ceiling social reasoning maturity but remain anchored in an individualistic cultural framework, regardless of their language or geographical origin. This individualistic bias is largely resistant to implicit linguistic cues, indicating that the "Whorfian Effect" in AI is suppressed by current alignment methodologies that prioritize cross-lingual consistency over cultural nuance. Models possess a latent "Synthetic Theory of Mind" that allows for explicit cultural perspective-taking. However, the degree of "Cultural Inertia" varies, with some models showing a persistent resistance to non-Western social scripts. As AI integration into the global social fabric accelerates, the need for Culturally Adaptive Systems becomes paramount. Future research must expand DASIS to include multi-dimensional cultural variables (e.g., Power Distance, Uncertainty Avoidance) and explore "Dynamic Alignment" techniques that allow models to autonomously detect and adapt to the cultural pragmatics of their human counterparts. By fostering a more pluralistic social intelligence, we can ensure that artificial agents serve as a bridge between diverse cognitive traditions rather than a tool for cultural homogenization.

References

Boroditsky, L. (2001). Does language shape thought?: Man-

darin and english speakers' conceptions of time. *Cognitive psychology*, *43*(1), 1–22.

Dong, F., Zeng, Y., Sang, Y., & Shen, H. (2025). Structuralist Approach to AI Literary Criticism: Leveraging Greimas Semiotic Square for Large Language Models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47).

Hall, E. T. (1976). *Beyond culture*. Anchor.

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

Hershcovich, D., Frank, S., Lenz, H., de Lhoneux, M., et al. (2022). Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Hofstede. (1984). *Culture's consequences: International differences in work-related values* (Vol. 5).

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, *4*(169), 2.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer publishing company.

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024). CultureLLM: Incorporating cultural differences into large language models. *NeurIPS 2024*, *37*, 84799–84838.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review*, *98*(2), 224.

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). Basic Books.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, cognition and personality*, *9*(3), 185–211.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In *International conference on machine learning* (pp. 29963–30005).

Sap, M., Rashkin, H., Chen, D., LeBras, R., & Choi, Y. (2019). Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press.

Shen, F., Zeng, Y., Lyu, D., & Wen, W. (2026). From Words to Worlds: Measuring Cultural Narrative Bias in LLMs via a Structural–Value Pipeline. In *Proceedings of the acm web conference 2026* (pp. 8961–8972).

Tang, X., Zeng, Y., & Dong, F. (2025). Unveiling Cultural Cognition in AI: A Systematic Investigation of Horizontal-Vertical Individualism-Collectivism Traits in Large Language Models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47).

Triandis, H. C. (2018). *Individualism and collectivism*. Routledge.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Zeng, Y., Dong, F., Zhao, J., Zheng, P., Li, J., & Zhou, H. (2025). Towards Culturally Fair Multimodal Generation: Quantifying and Mitigating Orientalist Biases in Text-to-Visual Models. In *Proceedings of the 33rd acm international conference on multimedia* (pp. 11434–11442).

Zeng, Y., Liang, K., Dong, F., & Zheng, P. (2025). Quantifying Risk Propensities of Large Language Models: Ethical Focus and Bias Detection through Role-Play. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47).