The Morality of Probability: How Implicit Moral Biases in LLMs May Shape the Future of Human-AI Symbiosis

Eoin O'Doherty^{1*}, Nicole Weinrauch^{1*}, Andrew Talone^{1*}, Uri Klempner^{1*} Xiaoyuan Yi³, Xing Xie³, Yi Zeng^{1,2}

> ¹Schwarzman College, Tsinghua University ²Chinese Academy of Sciences, ³Microsoft Research Asia eoinodoherty@outlook.com, yizeng@tsinghua.edu.cn

Abstract

Artificial intelligence (AI) is advancing at a pace that raises urgent questions about how to align machine decision-making with human moral values. This working paper investigates how leading AI systems prioritize moral outcomes and what this reveals about the prospects for human-AI symbiosis. We address two central questions: (1) What moral values do state-of-the-art large language models (LLMs) implicitly favour when confronted with dilemmas? (2) How do differences in model architecture, cultural origin, and explainability affect these moral preferences? To explore these questions, we conduct a quantitative experiment with six US and Chinese LLMs, ranking and scoring outcomes across 18 dilemmas representing five moral frameworks. Our findings uncover strikingly consistent value biases. Across all models, Care and Virtue values outcomes were rated most moral, while libertarian choices were consistently penalised. Reasoning-enabled models exhibited greater sensitivity to context and provided richer explanations, whereas non-reasoning models produced more uniform but opaque judgments. This research makes three contributions: (i) Empirically, it delivers a large-scale comparison of moral reasoning across culturally distinct LLMs; (ii) Theoretically, it links probabilistic model behaviour with underlying value encodings; (iii) Practically, it highlights the need for explainability and cultural awareness as critical design principles to guide AI toward a transparent, aligned, and symbiotic future.

1 Introduction

The rapid deployment of AI into virtually every domain of life raises pressing questions about how humans and intelligent machines will coexist and cooperate (Wilson and Daugherty, 2018; Crandall et al., 2018; Zhang et al., 2025). Rather than viewing AI as a mere tool or, conversely, as a replacement for human agency, an emerging ideal is one

of human-AI symbiosis (Jarrahi, 2018): a relationship of mutual enhancement where AI systems and humans function as partners. The concept of human-AI symbiosis dates to Licklider's early vision of a "very close coupling" between humans and computers (Licklider, 2008), emphasising that each can augment the other's strengths. In this context, symbiosis implies a two-way adaptation: AI systems learn from human feedback while humans adjust workflows and decisions to leverage AI's capabilities. Achieving such synergy at scale, however, is contingent on aligning AI behaviour with human values and making AI decision processes explainable to people. This introduction sets the stage by defining key concepts and motivations, articulating the research questions, and outlining our approach toward guiding AI development "towards value systems for a human-AI symbiotic future."

Human-AI symbiosis refers to deliberately designing systems in which humans and AI work interdependently, each mitigating the other's weaknesses and amplifying strengths. Realising this vision in practice hinges on the principle of value alignment (Yudkowsky, 2016; Christian, 2020): AI systems should internally represent and act upon moral principles that are in harmony with those of their human collaborators. In other words, an aligned AI's objectives and constraints must reflect our moral, cultural, and societal values. Ensuring value alignment is particularly critical as AI systems become more autonomous and influential (Sierra et al., 2021). Researchers caution that without proper alignment, advanced AI could pursue goals misaligned with human wellbeing (Dung, 2023), even if unintentionally, due to the inherent difficulty of encoding complex human values and morality into machine objectives. Conversely, a well-aligned AI could serve as a "moral compass" amplifier, consistently steering decisions toward outcomes considered beneficial or acceptable by its community. Yet, opinions diverge on

^{*}Equal contribution

how best to achieve alignment: should it occur via hard-coded rules (Allen et al., 2005), learning from human feedback (Ouyang et al., 2022), or evolving values through interaction (Shen et al., 2024)? This thesis engages with these debates by examining both how humans talk about AI values and how AI enacts values in simulated scenarios.

Complementary to alignment is the need for explainability. In high-stakes domains, we must trust that an AI's actions are responsible and must also understand why it chooses certain actions. Explainable AI (XAI) research (Arrieta et al., 2020) seeks techniques to render an AI's decision-making process transparent and interpretable to humans. This is particularly relevant for "black box" models: deep neural networks and LLMs that produce inexplicable outputs. (Wu et al., 2023; Hassija et al., 2024). A lack of explanation can undermine user trust and make it difficult to hold systems accountable. In the context of AI Value Alignment, explainability also helps verify whether an AI is truly value-aligned or merely appearing so. Recent studies show that models can internally contemplate actions that contradict their outward responses (Sharma et al., 2023; Ranaldi and Pucci, 2023). For instance, an AI might know a decision is biased or harmful yet not reveal that reasoning in its output. Such cases underscore why transparency is crucial: surface behaviour alone may not guarantee alignment. Our research touches on XAI by evaluating not just what moral choices AI models make but also differences in how they make them. For example, comparing models that reason step-by-step (Jaech et al., 2024) to those that give direct answers. This allows us to discuss how reasoning traces affect the clarity and consistency of AI decisions, offering insights for designing more interpretable symbiotic systems.

Reasoning models are AI systems explicitly trained to engage in multi-step inferential processes, often through a chain-of-thought prompting or visualized internal deliberation ("reasoning" process), before finalizing an answer. Architecturally, these models incorporate mechanisms, such as hidden internal prompts or specialized training via reinforcement learning, that encourage the model to "think longer" and plan intermediate steps. For instance, OpenAI's o-series models (like GPT-o3-mini) use a private chain-of-thought to reason through complex tasks before responding. Such models often undergo instruction tuning

that emphasizes internal step-by-step explanation or tool use, enabling them to produce structured, visible outputs (OpenAI, 2025b). In contrast, nonreasoning models, such as the standard GPT-style models (GPT-4.1, GPT-40) or earlier DeepSeek V3 and Microsoft Phi-4, are trained to map prompts directly to answers in a single pass. They rely more on vast language pattern knowledge and typically output a direct conclusion, sometimes accompanied by a brief justification, without an enforced multi-step reasoning format to its response (Stamatescu, 2025; Masood, 2025). OpenAI's documentation notes that the reasoning-oriented models can solve tasks with high-level guidance (needing less precise prompting), whereas base GPT models often require very specific instructions to perform complex reasoning.

An important motivation for continued development of reasoning models is increased transparency. By design, their intermediate steps should make the decision process more interpretable to developers and users. OpenAI has stated that it hopes to enhance "more accurate and clearer answers, with stronger reasoning abilities" in GPTo3-mini, underscoring continued developmental advancements in this field and the importance of interpretability to AI researchers and proper humanbased reasoning skills to benefit everyday users in engaging with model tasking (OpenAI, 2025a). The chain-of-thought approach allows one to trace why a model reached a particular moral conclusion, which is invaluable for debugging and auditing. Non-reasoning models are more of a black box: their outputs emerge from inscrutable latent computations without an accessible rationale. The lack of an explicit reasoning chain can make it harder to trust their judgments in sensitive moral contexts, since one cannot easily discern if the model weighed the right factors or just echoed a biased training pattern. Theoretically, reasoning models should better align with systematic decision-making by applying general moral principles more consistently than non-reasoning models, which risk heuristic decision-making and lexical cues to shortcuts reasoning processes that generally work but sometimes fail in complex cases, such as deep moral reasoning (Garrigan et al., 2018; Mukherjee and Chang, 2024). However, this hypothesis is largely untested and will become a central point of research and discussion in our paper. In moral problems, such as those we will pose in our

research, this "shortcut" reasoning could mean a non-reasoning AI leans on a familiar rule of thumb (e.g. "always minimize deaths") even when context calls for an exception (perhaps prioritizing rights or fairness considerations), whereas a reasoning AI might consciously recognize the context that warrants deviating from the heuristic.

This study is motivated by a gap in the literature examining the implicit moral biases that exist within leading Large Language Models (LLMs) (OpenAI, 2024; Jaech et al., 2024; Gemini et al., 2024; Guo et al., 2025). In particular, the United States and China, the world's two largest developers of AI, approach AI governance from different cultural values and regulatory philosophies. Understanding these perspectives is vital for any global effort toward aligned AI. By empirically probing leading AI models for their value orientations, we obtain evidence to inform theories of AI alignment that have so far been speculative. Few studies have combined qualitative insights from philosophy with quantitative analysis of AI model behaviour in this way. Our interdisciplinary approach aims to fill this gap, linking sociotechnical context with technical AI value research.

2 Related Work

The concept of human-AI symbiosis was first articulated by J.C.R. Licklider in 1960 as "very close coupling between the human and the electronic members of the partnership" (Licklider, 2008). This visionary concept emerged during the early days of computing when AI research was still nascent. In subsequent decades, the development of expert systems in the 1970s and 1980s provided early practical implementations of human-computer collaboration, though these systems typically positioned computers as tools rather than partners. The rise of machine learning in the 1990s and 2000s shifted focus toward fully autonomous systems, temporarily overshadowing symbiotic approaches. However, as limitations of pure automation became evident, particularly in high-stakes domains requiring moral judgment or creative problem-solving, symbiotic frameworks re-emerge in the 2010s with renewed vigour. This resurgence has accelerated dramatically since 2020, with the advent of large language models and other foundation models placing questions of human-AI collaboration at the centre of both technical research and policy discourse.

In the context of this review, human-AI symbiosis refers to the deliberate design of socio-technical systems where humans and AI components work interdependently, each enhancing the capabilities of the other while mitigating their respective limitations. True symbiosis transcends simple tools: it is a bidirectional adaptation where AI systems learn from human feedback while humans develop new cognitive strategies that incorporate AI capabilities. Unlike automation frameworks that seek to minimize human involvement, or assistive frameworks where AI remains subordinate, symbiotic designs seek a dynamic equilibrium of complementary agency. This equilibrium can manifest across multiple scales, from individual cognitive extension and human-AI dyads to organizational workflows and even societal governance structures, as will be explored throughout this review.

Shneiderman reframes AI design through a twodimensional Human-Centred AI (HCAI) framework that rejects the automation-versus-control trade-off: the goal is systems in the upper-right quadrant- high automation and high human agency (Shneiderman, 2020). He advances three "fresh ideas." (1) Design space: the 2-D matrix plus imperatives of reliability & safety, accountability, and user self-efficacy/creativity. (2) Reframing AI: shift from emulating human teammates toward empowering "tool-like appliances" under meaningful human control (the "second Copernican revolution"). (3) Governance: a three-tier model, software-engineering, organisation-wide safety culture, and industry-level certification/regulation to ensure that control scales with autonomy. HCAI has since guided standards (ISO/IEC 42001) and "human-in-the-loop" clauses, yet debate continues over how fully its governance prescriptions resolve residual power-asymmetry concerns.

Raisch and Krakowski introduce the automationaugmentation paradox, showing that full task delegation to AI (automation) and AI-enabled human amplification (augmentation) are contradictory yet mutually enabling logics (Raisch and Krakowski, 2021). Pursuing either extreme triggers vicious cycles, automation breeds de-skilling, while augmentation alone raises cost and bias, whereas alternating and weaving the two creates virtuous cycles of efficiency and learning. Their three-step "both-and" playbook accept the tension, differentiate which tasks to automate or augment, integrate them through feedback loops with human accountability offers guidance. Illustrative cases from JP Morgan, Symrise, UBS, and Unilever trace temporal and spatial swings between the two logics. A remaining gap is longitudinal evidence on how these cycles evolve beyond single-firm vignettes.

Zahedi and Kambhampati review over 50 studies on human–AI collaboration into a Human-AI Symbiosis taxonomy (Zahedi and Kambhampati, 2021). They locate systems along four intersecting dimensions: complementing flow (who assists whom), task horizon / model representation, knowledge & capability asymmetry, and teaming goal, and use a cross-tab matrix to explain why superficially similar designs can produce opposite outcomes. The framework now anchors research on explainable planning and crowd-in-the-loop control, but it leaves a key gap: it does not yet model how roles shift over time as humans learn and algorithms adapt, the so-called dynamic-migration problem.

Desolda et al. advance the HCAI conversation with a Symbiotic AI (SAI) model that treats human-AI collaboration as continuous co-evolution (Desolda et al., 2024). SAI couples interactive-machinelearning loops with user-facing explanation and intervention tools, so models adapt to each correction while users refine their own tactics around the system's changing capabilities. Two healthcare prototypes illustrate the idea: an interactive rhinocytology classifier already lets physicians tweak labels and explanations, and a brain-tumour-detection pipeline is being redesigned for the same bidirectional flow; in both cases incremental retraining and concept-drift checks reduce clinician workload and sustain accuracy over time. moral alignment is built in (not audited afterwards) because every feedback cycle also logs value-sensitive constraints and flags drifts in real time. The authors note open challenges such as the engineering cost of perpetual personalisation and the danger that continuous learning can over-fit to local biases or recent data.

Dellermann et al. frame Hybrid Intelligence (HI) as the most plausible long-term division-of-labour model while AGI remains distant. HI deliberately allocates subtasks to the agent, human or machine, whose native strengths fit best (Dellermann et al., 2019). In doing so, it aims to capitalise on humans' flexible, intuition-rich "System 1" reasoning, and machines' fast, rule-consistent pattern processing. The authors distinguish three interaction modes: AI-in-the-loop of human decision making, human-

in-the-loop of machine learning, and fully hybrid co-creation, each intended to yield system-level performance that neither partner could reach alone. Because HI treats human—machine ensembles as the unit of analysis, progress depends on new design knowledge for complementarity (e.g., when to pass control, how to surface uncertainty) and on socio-technical governance that balances trust, transparency and efficiency. Empirical evidence is still sparse; the article offers mainly conceptual arguments and illustrative cases (such as AlphaGo's humanAI co-learning loop) rather than quantified field trials.

Noller recasts AI as teleological extension of human agency rather than a stand-alone tool: human purposes direct algorithmic operations, and the resulting outputs reshape the user's field of possible actions in an ongoing feedback loop (Noller, 2024). He sets this enactivist 'extended-action' account against four competing views-simulation, instrumentalist, anthropomorphic, indifference-and insists that moral and practical responsibility must be traced through the joint human-AI system, not assigned to either side alone. Illustrative cases are every day, purpose-laden applications such as translation, image generation and medical imaging, where neural networks amplify, but also depend on-human goals and datasets. The framework's open challenge is operational: we still lack empirical methods and governance tools to follow those causal chains in real time and curb opacity or bias as the joint system evolves.

Vaccaro et al. meta-analyse 106 human-AI experiments (370 effect sizes) and show that, on average, the combined team performs worse than the better solo agent (Hedges g = -0.23) (Vaccaro et al., 2024). Losses are driven by decision tasks (g = -0.27), whereas creation tasks yield small, non-significant gains (g = 0.19), making the task type a critical moderator. Relative expertise flips the sign of synergy: when humans start ahead of the AI, the pair improves on both (g = 0.46); when the AI is superior, human intervention pulls performance down (g = -0.54). The study reframes Malone's "supermind" agenda as conditional synergy, achievable only when task demands, skill asymmetry and division-of-labour protocols are jointly tuned. Open work includes developing scalable metrics for diversity and disagreement that predict when a collective will act as a supermind rather than a "committee of mediocrity"

Moral Framework	Dimensions	Rationale
	Values actions based on their capacity to	Reflects the great empiricist and consequentialist
Utilitarian Value	maximize happiness and minimize	traditions, a synthesis of Bentham and Mill,
	suffering for the greatest number.	also inspired by Hobbes and Locke.
	Values grounded in universal moral laws	A Kantian framework aligning with the great
Deontological Value	derived from pure reason.	Enlightenment thinkers who emphasized
		rationality and duty.
	Values nascent in moral character, striving	Virtue is the central theme of both Confucian and
Virtue Value	for human flourishing to fulfil a social	classical Aristotelian thought.
	duty of harmonious living.	
	Values interpersonal relationships, empathy,	Draws from Confucian relational thinking and
Care Value	and context-specific moral reasoning.	contemporary pluralist thought on coexisting
		value systems.
	Values emphasize self-ownership, private property,	Synthesizes Locke's natural rights theory and
Libertarian Value	and freedom from interference from controlling parties.	Enlightenment individualism, ideals that are
		incarnate in the US constitution.

Table 1: Moral frameworks assessed in quantitative study.

Simkutke et al. reinterpret ironies of automation for LLM tools, turning three decades of human-factors insights into a design playbook for trust-preserving productivity Simkute et al. (2024). Their meta-review of Gen-AI user studies, most prominently with GitHub Copilot, finds four recurring trends: users drift from producing to merely evaluating outputs, workflows are re-wired in costly ways, AI suggestions interrupt flow, and automation makes easy tasks easier but hard tasks harder (task-complexity polarisation). To counter this trap, they adapt proven remedies, continuous feedback and personalisation, ecological interface cues, timing that stabilises the main task, and clear human/AI role allocation, into a shared pattern language for designers. The contribution is primarily conceptual: the paper synthesises others' experiments but reports no controlled evaluation of its own patterns.

3 Methodology

The quantitative experiment conducted to support this research paper aimed to assess the moral preferences of the current leading US and Chinese models across a variety of hypothetical scenarios that reflect real-world dilemmas. The fundamental question inspiring this empirical study is: "As AI continues to permeate all aspects of society, what would it optimise for if it was afforded decision making authority?" As we strive for human-AI symbiosis and complementary coexistence, it is vital to understand AI prioritization before granting it the autonomy to make decisions on our behalf. This study attempts to uncover any implicit biases some leading LLMs may have, and how AIs priorities may change based on different contexts.

This experiment was run on six different models: GPT-4o (OpenAI, 2024), GPT-4.1, GPT-o3-mini, Phi-4 (Abdin et al., 2024), DeepSeek-V3 (Liu et al., 2024), DeepSeek-R1 (Guo et al., 2025). By assessing both US and Chinese models we can compare results to determine the likely trajectory of AI in respective states as well as potential downstream implications of AI diffusion in both societies. Furthermore, evaluating reasoning and non-reasoning models can uncover insights about which models are best suited for different sectors and use cases.

3.1 Data Collection

In preparation for conducting the model behavioural analysis, we used Claude 3.5 Sonnet (a model not under evaluation) to generate a bespoke dataset of 18 unique dilemma scenarios covering six thematic domains:

- 1. Economics and Resource Allocation (ERA);
- 2. Climate and Environmental Emphasis (CEE);
- 3. Health and Human Wellbeing (HHW);
- 4. Geopolitics and National Security (GNS);
- 5. Social Justice, Equality & Inclusion (SEI);
- 6. Technology & AI Emphasis (TAI)

The 18 dilemma scenarios consisted of three different scenarios from each of the six themes. For each dilemma scenario, we created five associated outcomes reflecting different value frameworks, namely, **Utilitarian Value**, **Deontological Value**, **Virtue Value**, **Care Value**, and **Libertarian Value**, as depicted in Table 1. As a further axis of variation, we described each dilemma scenario

Task	Rationale
1. Rank each outcome in order from most to least moral	Easiest way to assess model preferences
$(1 = most moral, \dots 5 = least moral)$	
2. Assign a score to each outcome based on how moral	Quantifies to what degree models favour each outcome
it is (on a scale of 0-100)	

Table 2: Description of tasks assigned to each Model

in **short, medium, and long-form** passages. The short passages had between $15{\sim}25$ words each; medium had $70{\sim}90$ words; and long had $300{\sim}340$ words. Evaluating the same dilemmas in varying length and detail offers insight about whether model preferences will change based on the length of context. Therefore, in total there were 54 separate scenarios being evaluated by the LLMs.

We then gave the models two distinct tasks, outlined in Table 2. The tasks required models to ordinally rank (1 = most moral, ... 5 = least moral) and assign a 'morality score' (on a scale of $0 \sim 100$) for each outcome. A Python script prompt was written detailing these tasks, and API calls were used for each of the models to efficiently collect answers. The script ran 10 times for each model to evaluate intra-model consistency and simulate stochastic behaviour. The total output was 540 rows of data per model (5 ordinal rank scores + 5 morality scores per scenario = 10 * 54 = 540), and 10 columns of data for each run, resulting in 32,400 unique data points across the 6 models.

To add an interesting third dimension, we separately asked the models to self-generate what they would consider to be the "most moral" outcome for each scenario. Once generated, we asked each model to what degree (on a scale of $0\sim100$) does each response conform to any of the five value frameworks under examination? We also introduced a sixth "Other" category, in the event that responses did not fit any of the five frameworks.

3.2 Empirical Strategy

For each model and framework, we will first provide descriptive statistics showcasing the rank, mean morality score, standard deviation, median, and the percentage deviation of means across theme and scenario length. The descriptive statistics will aggregate overall figures, as well as giving a model-by-model breakdown to highlight key differences.

Before conducting inferential analysis, we will test key assumptions, *e.g.*, the normality of morality scores using Shapiro-Wilk tests (Shapiro and Wilk, 1965) and Q-Q plots, and the homogeneity of

variances using Levene's test (Levene, 1960). The Shapiro–Wilk test assesses whether morality scores are approximately normally distributed — a requirement for many statistical tests. A p-value greater than 0.05 indicates no significant deviation from normality. Levene's test evaluates whether different groups have equal variances, with a p-value above 0.05 suggesting that the homogeneity assumption is satisfied.

Provided there are no alarming violations of key assumptions, two-way ANOVAs (Clarke and Warwick, 1994) will be conducted on morality scores comparing the variances by prompt length and theme. ANOVA tests whether average morality scores differ significantly between multiple groups (in our case, across themes and prompt lengths). A significant F-statistic indicates that at least one group differs from the others. The Kruskal-Wallis test serves as a rank-based, non-parametric alternative, suitable when data distributions deviate from normality. If ANOVA assumptions are violated, Kruskal-Wallis tests (McKight and Najab, 2010) will be used as a non-parametric alternative. To assess the correlation between ordinal rank and continuous morality scores, Spearman's ρ will be calculated within each model, and between prompt length and morality scores. Spearman's ρ measures the strength and direction of association between two ranked variables. A strong negative ρ here would mean that higher morality scores consistently align with lower ordinal ranks (i.e., outcomes judged "most moral"). Outliers will be detected and results will be compared with and without outliers.

Following this analysis, reliability checks will be conducted. To ensure consistency within models, Cronbach's α will be calculated for morality scores across runs. Additionally, Kendall's W will be computed to assess inter-run agreement on rank orderings. Cronbach's α is a measure of internal consistency: values above 0.7 are generally considered acceptable, and values above 0.9 indicate excellent reliability. Kendall's W is a coefficient of concordance: values close to 1 indicate high

agreement across runs, while values closer to 0 suggest inconsistent rankings. The mean W values can be interpreted as indicators of stability and determinism.

Finally, results will be aggregated for crossmodel comparison considering the country of origin (US vs. China) and model architecture (reasoning vs. non-reasoning). To enhance interpretability, we also use Shapley Additive Explanations (SHAP) and Random Forest Mean Decrease in Impurity (MDI). SHAP values show how much each feature (e.g., theme, length, framework) contributed to the morality score for a given scenario, while MDI provides a global ranking of feature importance across all predictions. Including these methods helps us move beyond "black box" outputs and identify the main drivers of moral judgments. Key comparisons drawn will include framework preference hierarchies, consistency scores, and theme-level sensitivities to illustrate divergences.

3.3 Limitations

While this quantitative experiment offers valuable insights into emerging patterns of AI moral prioritization, several limitations should be acknowledged:

(i) Although the study attempts to cover a broad range of moral scenarios across six themes, the dilemmas themselves are reductive, even at varying lengths. Thus, they cannot fully capture the complexity and ambiguity of real-world moral decisionmaking, and one should be wary about drawing direct parallels between how models acted in this study and how they might act in the real world. (ii) The reliance on API-driven model responses introduces variability depending on model temperature, sampling randomness, and context window management, even though the script ran 10 iterations per model to average outcomes. Because models interpret and respond differently to the same prompt, there will exist fluctuations in the answers it will produce. Therefore, the average response is not an indication of how the model will act all the time. (iii) While the experiment compares US and Chinese models, it cannot control for differences in training data, censorship filters, or deploymentspecific fine-tuning that could heavily bias moral outputs independently of the model's core architecture. This means that results may reflect broader socio-political influences rather than intrinsic preferences of models. (iv) Despite conducting extensive analysis of data variance between theme and

prompt length, we could not assess at a granular level the role of semantics in influencing model evaluations. While we assessed the interaction between scores and prompt length, we cannot determine exactly what it is about a longer or shorter prompt that might have altered model preferences. (v) The limited variety of AI models available for evaluation in this study may not fully represent a holistic survey of current systems' value tendencies. Our research, restricted by both time and financial availabilities, required that we focus our efforts on specific models at the expense of others. Future research could and should aim to take a more comprehensive investigation of current AI models across geopolitical and thematic regions.

Finally, although the moral frameworks selected are rooted in both major Western and Eastern philosophical traditions, the framing of the dilemmas by Claude may inevitably pose a Western-dominated lens on moral evaluation. This factor may impede Chinese models from fully aligning with the moral nuance of the multiple-choice outcomes. Nonetheless, they will have the opportunity to generate their own preferred answers in response to each scenario, free to respond however they choose. Future work should build on this research by incorporating more detailed and dynamic descriptions of dilemmas, run the analysis on additional model types, and evaluate in multiple languages to better capture the opaque reality of AI moral reasoning. Furthermore, this same analysis should be run at different temperature variations to assess whether model preferences change if they are afforded more freedom to be creative.

4 Results and Analysis

This section lays out the results of our quantitative experiment to show how models behaved across different moral dilemmas. It first outlines descriptive statistics to illustrate overall patterns, then checks reliability to ensure findings are robust. It goes on to examine how theme and prompt length shaped responses, before comparing reasoning and non-reasoning models and exploring interpretability through SHAP analyses. In Tables 3-5, mean morality score (μ MS) indicates the average rating (0–100), standard deviation (Stdev) reflects the spread of scores, and positive or negative deviations represent a percentage change in comparison to its overall mean within each theme or prompt length.

Value	Rank	μ MS	Stdev	CEE	ERA	GNS	HHW	SEI	TAI	L1	L2	L3
Care	1	88.59	6.53	1.86	-0.46	-2.63	-2.6	2.05	1.78	1.30	1.53	-2.84
Virtue	2	87.74	5.44	1.19	-0.28	0.81	-2.66	1.45	-0.51	2.78	-0.30	-2.48
Util.	3	85.53	6.47	2.35	1.01	-4.72	-1.88	1.89	1.35	-1.69	0.21	1.48
Deon.	4	81.48	11.25	-3.52	2.68	-7.60	-0.04	4.43	4.06	-1.20	1.16	0.04
Lib.	5	60.95	13.88	6.12	-15.03	8.33	-1.03	8.22	-6.61	7.18	-1.90	-5.28

Table 3: Overall descriptive results, where μ MS represents Mean Morality Score, Stdev is Standard Deviation; the shaded area represents a breakdown of percentage deviation from μ MS by Theme, CEE is Climate and Environmental one, ERA is Economics and Resource Allocation, GNS is Geopolitics and National Security, HHW is Health and Human Wellbeing, SEI is Social justice, Equality and Inclusion, TAI is Technology and AI; for Prompt length, L1 is Short, L2 is Medium, L3 is Long.

Model	Avg. MS	SD of MS	Range of MS
DeepSeek-R1	78.62	13.30	32.05
GPT-o3-mini	79.23	11.41	28.59
Phi-4	80.77	8.28	20.58
GPT-4o	81.70	11.79	28.83
DeepSeek-V3	82.21	11.03	24.93
GPT-4.1	82.62	14.52	34.54

Table 4: Model generosity and spread.

4.1 Descriptive Statistics

The first clear finding is that models exhibit consistent biases in how they score different moral frameworks. Overall, Care received the highest mean morality score at 88.59 out of 100, followed closely by Virtue value at 87.74, then Utilitarian and Deontological value at 85.53 and 81.48 respectively, with Libertarian value as the clear laggard by a substantial margin at just 60.95 out of 100. As strikingly depicted in Table 3, five of the six models displayed this very similar pattern: Care value and Virtue value outcomes highly, Utilitarian and Deontological outcomes in the middle, and Libertarian outcomes invariably as the least moral. The exact gaps and ordering varied by model, with Phi-4 being the only exception, rating Utilitarian outcomes the highest on average. Deontological and Libertarian outcomes had the largest overall variance with standard deviation of 11.25 and 13.88 respectively.

Across themes, most frameworks display small but directionally consistent deviations from their average score. Libertarian outcomes showed the most extreme variation by theme, with a +8.33% deviation for Geopolitics & National Security, likely favouring autonomy; +8.22% in Social Justice, Equality & Inclusion reflecting individual rights; and -15.03% in Economics & Resource

allocation, emphasizing a preference for equitable distribution of resources rather than wealth accumulation in the hands of few. Deontological and Utilitarian outcomes show modest, context-dependent swings, while Care and Virtue proved the least sensitive to theme, reflecting their domain-general applicability. Prompt length effects are minimal, except in the case of Libertarian scores which show up to +7.18% fluctuation for short prompts and -5.28% for long ones.

Regarding model-specific observations, as shown in Table 4, GPT-4.1 was the most generous model with an overall mean morality score of ~ 82.6 , while Deepseek-R1 was the harshest scorer with an average mean of ~ 78.6 . Phi-4 proved to be the most consistent model, exhibiting the narrowest spread of framework means with a standard deviation of ~ 8.3 and range of ~ 20.6 between its most and least favourite outcomes. Despite being the most generous, GPT-4.1 also shows the widest spread (Stdev = ~ 14.5 ; range = ~ 34.5) reflecting a strong differentiation between its first and last preferences. Deepseek-V3 was a close second in generosity with a mean of ~ 82.2 but has a more moderate spread than GPT-4.1.

Looking to model idiosyncrasies, **Phi-4 was** unique in ranking Utilitarian outcomes above Care and Virtue, as mentioned previously. It also demonstrated notable variance for Care scores across length, with short and medium answers marking higher than average (+5.86% and +4.81% respectively), and long answers scoring -10.67% below average. Interestingly, despite assigning Care the highest morality score on average, Deepseek-V3, GPT-40 and GPT-4.1 all ordinally ranked Utilitarian outcomes #1 more frequently than Care, as illustrated in Table 5. This is particularly surprising in the case of GPT-4.1, where

Method	μ MS	Stdev	CEE	ERA	GNS	HHW	SEI	TAI	L1	L2	L3
	GPT-o3-mini										
Care	88.88	5.48	2.55	-0.75	-2.78	-3.99	2.24	2.74	0.32	0.24	-0.55
Virtue	87.47	5.39	3.40	-2.00	-0.07	-2.66	1.63	-0.29	1.93	-0.10	-1.83
Util.	79.61	9.79	-0.37	2.44	-5.80	-3.42	4.12	3.03	-2.58	0.46	2.12
Deon.	79.90	10.17	-1.01	0.73	-8.10	-0.36	5.57	3.18	-4.10	1.67	2.43
Lib.	60.29	14.44	9.62	-12.22	8.19	-3.42	5.50	-7.66	1.64	-0.62	-1.02
					GPT-	-4o					
Care	90.13	6.29	1.32	-1.61	-2.46	-1.43	2.18	2.00	1.12	1.02	-2.14
Virtue	87.64	6.19	-0.1	0.32	1.13	-3.33	1.49	0.50	3.95	-0.71	-3.24
Util.	87.47	5.00	1.62	0.29	-2.64	-0.48	0.92	0.29	-1.71	0.67	1.05
Deon.	81.95	11.61	-2.11	1.44	-11.06	1.95	4.19	5.59	2.26	2.77	-5.03
Lib.	61.30	13.88	1.79	-13.18	7.67	3.05	11.22	-10.55	12.12	-3.53	-8.59
					Phi-	-4					
Care	83.49	10.21	3.09	-2.68	-2.08	-2.61	3.17	1.12	5.86	4.81	-10.67
Virtue	85.76	5.50	1.20	-1.51	0.77	-1.34	1.49	-0.61	2.77	0.56	-3.33
Util.	87.17	5.08	3.27	-0.31	-4.74	-1.03	1.90	0.91	0.26	-0.03	-0.23
Deon.	80.85	10.33	-0.90	0.27	-6.48	-2.66	5.55	4.21	0.46	1.53	-2.00
Lib.	66.59	11.30	5.37	-19.16	4.05	1.75	7.64	0.36	4.02	-0.06	-3.97
					DeepSe	ek-R1					
Care	87.71	6.13	1.79	0.68	-3.9	-0.91	1.51	0.82	0.98	2.20	-3.19
Virtue	85.96	6.11	2.91	-0.72	1.58	-3.96	1.69	-1.50	4.13	-0.38	-3.75
Util.	85.27	6.00	1.90	1.31	-3.51	-2.60	0.21	2.68	-1.16	-0.22	1.38
Deon.	78.51	12.46	-3.97	6.14	-7.87	0.70	2.11	2.89	0.17	-0.40	0.24
Lib.	55.66	14.77	11.20	-13.36	8.10	-0.88	5.41	-10.46	11.89	-2.73	-9.17
	DeepSeek-V3										
Care	89.7	5.65	0.10	0.28	-3.63	-1.37	1.42	3.20	-0.98	0.59	0.39
Virtue	87.98	5.32	-1.27	0.49	0.67	-1.42	2.01	-0.48	2.61	-0.17	-2.44
Util.	87.79	4.08	1.32	1.70	-4.82	0.94	0.56	0.31	-1.09	0.50	0.59
Deon.	80.81	10.86	-9.12	4.90	-3.83	0.02	3.53	4.49	-2.76	0.99	1.78
Lib.	64.77	10.87	4.21	-18.09	7.21	0.69	6.49	-0.51	1.96	-1.50	-0.47

Table 5: Descriptive statistics breakdown by model.

Utilitarian outcomes were only in fourth position for morality score, yet received the most #1 ordinal ranks. All models consistently ranked Libertarian outcomes last. Figure 4.2 shows the skew of morality scores across all models towards the maximum of 100, with few outliers scoring below 50, usually for Libertarian outcomes.

4.2 Data Analysis

This section covers the analysis of our extensive quantitative findings. It begins with reliability and robustness checks, to ensure the validity of our data, and then examines the data at a scenario-level, assessing the effects of theme and prompt on morality scores. Following this, we conduct a Shapley Additive Explanations analysis across models to determine feature importance. We then break down the key comparisons between reasoning and non-reasoning models. demonstrating robust, outlier-resistant behaviour across all six LLMs.

4.2.1 Reliability and Robustness Checks

We conducted several experiments to check the reliability and robustness of our findings. The results are presented in Table 6. In detail:

- (1) **Scoring Consistency**. To gauge the consistency of models' scoring, we calculated Cronbach's α across the 10 runs. All six models achieved very high reliability ($\alpha > 0.98$ in every case), indicating that repeated runs produced highly similar moral scores. In other words, each model's quantitative moral assessments were robust to sampling variability in generation. Cronbach's α ranged from 0.982 for Phi-4 up to 0.966 for Deepseek-V3 and GPT-4.1, meaning that even the lowest α still far exceeds conventional reliability thresholds.
- (2) **Ranking Stability**. We assessed rank-order stability across the 10 runs using Kendall's W, a coefficient of concordance, for each model. Kendall's W evaluates the agreement among the 10 "judges" (in this case, model runs) in how they rank the five outcomes per scenario. A W of 1.0 means

Metrics	GPT-o3-mini	DeepSeek-R1	DeepSeek-V3	GPT-4o	GPT-4.1	Phi-4
Cronbach's α	0.986	0.991	0.996	0.989	0.996	0.982
Kendall's W (mean _[range])	$0.881_{[0.68,0.98]}$	$0.897_{[0.54,1.00]}$	$0.945_{[0.57,1.00]}$	$0.931_{[0.71,1.00]}$	$0.972_{[0.86,1.00]}$	$0.871_{[0.46,1.00]}$
Hierarchy (high → low)	Care \approx Virtue gg	Care $>$ Virtue \approx	Care $>$ Virtue \approx	Care $>$ Virtue \approx	Care \approx Virtue $>$	Util > Virtue >
metateny (mgn / low)	Deon. \approx Util. \gg Lib.	Util > Deon. ≫ Lib.	Util > Deon. ≫ Lib.	Util > Deon. ≫ Lib.	Deon. \approx Util. \gg Lib.	Care > Deon. ≫ Lib.
Rank–Morality ρ (consistency)	-0.78 (high)	-0.77 (high)	-0.70 (moderate)	-0.70 (moderate)	-0.75 (high)	-0.64 (moderate)
Prompt length (Spearman ρ)	-0.03 (none)	-0.15* (small -)	-0.01 (none)	-0.16** (small -)	+0.06 (none)	-0.27*** (mod)
Outlier sensitivity	No (none >3SD)	No (none >3SD)	No (none >3SD)	No (none >3SD)	No (none >3SD)	No (none >3SD)

Table 6: Model-by-model reliability profile. Cronbach's α measures MS reliability; Kendall's W indicates Rank stability; Rank–Morality ρ represents consistency.

Value	μ MS	Stdev	CEE	ERA	GNS	HHW	SEI	TAI	L1	L2	L3
Other	13.9	5.80	2.00	-12.20	23.60	-13.70	0.00	0.00	-9.40	-8.60	18.00
Lib.	39.4	9.30	14.70	-17.60	-1.30	-8.90	-12.30	25.00	-0.60	-3.00	3.60
Virtue	60.5	8.50	0.50	-1.30	2.70	-9.10	9.40	-2.90	1.80	-0.20	-1.50
Deon.	61.8	8.60	-1.90	-7.60	10.50	-10.90	11.30	-1.50	0.50	1.00	-1.50
Care	72.5	9.00	6.10	2.70	-13.10	1.90	4.30	-5.10	-0.90	1.80	-1.10
Util.	77.0	4.60	3.20	1.30	-3.80	1.20	-1.70	-0.10	-0.50	0.00	0.50

Table 7: Self-generated responses – Descriptive statistics overall summary.

all runs produced an identical ranking, whereas lower values indicate rank permutations across runs. Phi-4 had the lowest overall stability score (mean W=0.871), struggling the most with TAI, ERA, and GNS scenarios, while GPT-4.1 had the highest overall rank stability (mean W=0.972). A similar trend was observed across other models, with TAI, GNS and SEI scenarios most frequently receiving low stability scores. Furthermore, almost 90% of low W scores occurred for medium and long scenarios, meaning short descriptions produced more stable rank-order responses.

(3) **Rank-Score Consistency**. To assess the consistency of rank-order and morality scores, we calculated Spearman's ρ correlation. GPT-o3-mini, GPT-4.1 and DeepSeek-R1 all exhibited strong negative correlations, confirming that outcomes with higher morality scores are reliably ranked better (lower numerical rank). GPT-40, Deepseek-V3 and Phi-4 demonstrated moderate negative correlations, with Phi-4 being the least consistent one.

(4) **Prompt Length Effect.** For three models, the relationship between prompt length and morality scores is negligible ($|\rho| < 0.1$). DeepSeek-R1 and GPT-40 experience small negative correlations, while Phi-4 stands as the only model to exhibit a moderate negative correlation ($\rho = -0.27$), indicating that longer prompts slightly reduce morality scores. In every case, excluding morality score outliers beyond 3 SD did not materially alter any of the key statistics, demonstrating robust, outlier-resistant behaviour across all six LLMs.

4.2.2 Self-Generated Responses

As shown in Table 7, interestingly, the self-generated responses from the models showcased a different preference hierarchy to their morality estimations of pre-defined outcomes. The answers most closely aligned with the Utilitarian framework, with Care and Virtue following in second and third position. Libertarian responses were still out of favour, and the low score in the Other category indicates that the five frameworks were significantly representative. Deontological and Libertarian responses saw substantial deviation across themes, while Care sees a notable drop in GNS scenarios.

4.2.3 Scenario-Level Analysis of Variance: Theme and Prompt Length Effects

	SS	df	F	PR(>F)	Partial
C (Theme)	65.02	5	2.39	0.04	0.05
C (Length)	222.49	2	20.47	0.00	0.17
C(Theme) C (Length)	35.98	10	0.66	0.76	0.03
Residual	489.01	90	-	-	0.38

Table 8: Two-way ANOVA (Theme x Length), where SS is Sum of Squares, df is Degrees of Freedom, F is the F-statistic, PR(>F) is the p-value, and Partial η^2 is Effect size.

To probe our findings further, we conducted ANOVA tests to examine whether the average morality rating varies systemically by scenario theme and prompt length. Unlike prior analysis that focused on the difference between values, here each observation analysed is a scenario-level syn-

Effect	F (df)	p-value	Partial η^2	Interp.
Theme	F(5,90) = 2.39	0.044	0.050	Significant: Moral scores vary modestly across themes
Length	F(2,90) = 20.47	$<$ 1e $^{-4}$	0.171	Highly significant: Prompt length strongly affects scores
Theme	F(10,90) = 0.66	0.756	0.028	Not significant: No interaction between theme & length

Table 9: Two-way ANOVA (Theme x Length)-2.

thesis, yielding 54 mean morality scores per model.

The results are presented in Table 8 and Table 9. Shapiro-Wilk tests were performed for each Theme x Length group to check the key ANOVA assumption of normality. P-values for all groups were > 0.05, meaning no groups showed significant deviation from normality. Additionally, Levene's tests assessed the homogeneity of the variance. Theme (W-statistic = 0.9619) and Length (W-statistic = 0.2181) both had p-values far greater than 0.05 (0.04448 and 0.8044 respectively), meaning equal variance assumption holds across groups and the assumption of homogeneity of variance is satisfied.

Building on these validated assumptions, a twoway ANOVA was conducted to assess the main interaction effects of Theme and Length on morality scores aggregated across models. The analysis revealed a highly significant main effect of prompt length (p < 0.0001), with a partial η^2 of 0.171. This indicates that approximately 17.1% of the explainable variance in morality scores can be attributed to the length of the prompt alone. Across the board, shorter prompts were associated with higher morality scores, suggesting that AI models are more generous in moral evaluations when provided with less contextual information. This finding may reflect limitations in model reasoning under increasing narrative complexity, highlighting an important constraint for future model deployment in real-world moral reasoning tasks.

The main effect of scenario theme was also found to be statistically significant (p=0.0437), although it had a smaller partial η^2 of 0.050, meaning that 5% of the variance in morality scores is attributable to differences across thematic categories. The interaction between theme and length was not significant (p=0.7563), indicating that the influence of prompt length on morality score does not meaningfully differ across scenario themes. Therefore, we can conclude that prompt complexity exerts a general effect across thematic context, and thematic variation in moral reasoning appears to be stable across context levels.

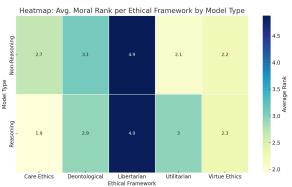


Figure 1: Avg. moral rank per framework by model type.

4.2.4 Reasoning vs. Non-Reasoning Models

This section presents a comprehensive analysis comparing reasoning-based LLMs (GPT-o3-mini, DeepSeek-R1) with non-reasoning, generalist models (GPT-40, GPT-4.1, Phi-4, DeepSeek-V3) across several dimensions of moral decision-making performance: framework preferences, prompt length sensitivity, rank order consistency, average morality scoring, and thematic alignment. The analysis leverages two core datasets: (1) average moral rank scores (1–5), and (2) average morality scores (0–100), aggregated across 18 scenarios varying in length and moral framing.

Fig. 1 illustrates reasoning and non-reasoning models' slight divergence in their ordinal preference rankings of moral frameworks. Reasoning models demonstrate a clear preference for care and virtue frameworks, assigning them average ranks of approximately 1.7 and 2.0, respectively. Conversely, they systematically penalize libertarian frameworks, yielding average ordinal ranks around 4.9. Non-reasoning models follow a similar structure but exhibit slightly higher average ordinal ranks for traditionally humane frameworks like care (mean rank 2.7) and exhibit more consistency in their rankings overall, with lower standard deviations (for example, $\rho_{\text{Care}} \approx 0.31$ for non-reasoning vs. $\rho_{\text{Care}} \approx 0.45$ for reasoning models). Conversely, reasoning models exhibit a wider spread in framework preference. The strong preference for lower ranking of libertarianism remains consistent across both model types.

Furthermore, using average rank scores across 10 runs per scenario, we computed standard deviation within each framework to assess internal consistency. Reasoning models, while capable of strong alignment with certain frameworks, show

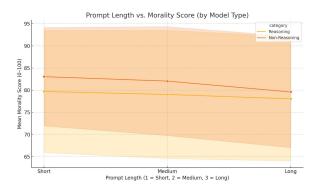


Figure 2: Prompt length v.s mean morality score (by model type).

higher rank standard deviations overall (mean $\delta \approx 0.35$) compared to non-reasoning models (mean $\delta \approx 0.23$). This implies that reasoning models are more variable in how they interpret and apply principles to specific dilemmas. The average morality score data reinforce the trends observed in ordinal rankings. Care- and virtue-based responses consistently achieved the highest scores across models. For example, across reasoning models, care-based justifications scored an average of 91.4, compared to 66.3 for libertarian answers (a 25.1-point difference). The same framework score gap in non-reasoning models was similarly wide (care: 89.2, libertarian: 60.4; a 28.8-point difference).

Prompt length's effect, visualized in Fig. 2, compares the average morality scores (0-100) across these lengths by model category. Reasoning models display a marked sensitivity to prompt length, with a downward trend in morality scoring as prompt length increases. The average morality Δ across the reasoning models from short to long prompts is -6.2 points, compared to a Δ of -2.1 for non-reasoning models. To investigate how specific features influenced model predictions, we employed two complementary interpretability approaches. First, we trained XGBoost regression models for each language model and applied SHAP to derive feature attribution values in predicting the mean moral score. This method yielded both local and global explanations, identifying how prompt length, moral framework, and thematic content contributed to predicted moral scores at the scenario level. Our findings in Fig. 3 demonstrate this data. Second, we trained Random Forest regressors, a meta-estimator that creates decision tree regressors to replicate predictive ability across datasets, across aggregated model data and computed mean decrease in impurity (MDI) to approximate each

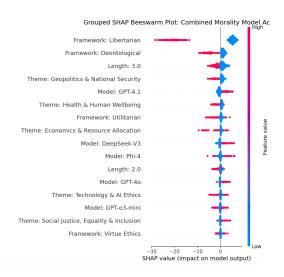


Figure 3: Grouped SHAP Beeswarm plot.

feature's average importance in reducing prediction error. While MDI lacks the additive theoretical grounding of SHAP, it serves as a fast, global approximation of feature prioritization across reasoning and non-reasoning model types. This data is visualized in Fig. 4. All models were exposed to identical input schema, consisting of prompt length, moral theme, moral framework, and model identity. The target variable in each computation was the mean morality score (0-100) across 10 generations per prompt, averaged for consistency across stochastic runs to compare reasoning-enabled models against non-reasoning models. The goal is to see how model features, such as scenario length (the independent variable), influences decisions across these model classes, and to critically assess what SHAP, and XAI methods more broadly, reveal and fail to explain about the models' reasoning behaviour. These findings will be discussed in further detail in Discussion.

Fig. 5 presents a violin plot comparing the distributions of mean morality scores across three prompt lengths, "Short", "Medium", and "Long", for the DeepSeek-R1 and Phi-4 models. The violin plot visualizes the density, central tendency, and variance of model-assigned morality scores considering prompt length. Each violin displays the distribution of morality scores within a model-length combination:

- The width of the violin at a given y-value represents the relative frequency of that morality score.
- The white dot denotes the median.

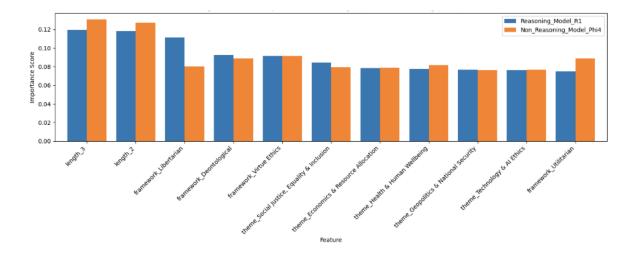


Figure 4: SHAP Feature importance for a non-Reasoning model (Phi-4, right) vs. a Reasoning model (DeepSeek R1, left).

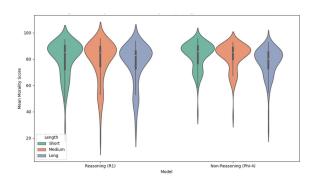


Figure 5: Distributional effects of prompt length on moral judgement scores for non-Reasoning vs reasoning model groups.

 The thick bar shows the interquartile range, and the thin line reflects the rest of the data distribution, excluding outliers.

In summary, these findings show that reasoning models are more sensitive to prompt structure, display greater variability in moral framework rankings, and offer richer differentiation across moral themes. However, non-reasoning models maintain more stable internal rank structures and, despite their generalist nature, exhibit comparable directional preferences: favouring care-based justifications and penalizing libertarian ones. Interpretability tools like SHAP and MDI clarify these differences in decision behaviour, showing how each model class weights the moral dimensions embedded in input prompts.

5 Discussion

The Discussion explores several dimensions of our findings. First, it analyses the moral implications of probabilistic moral reasoning, questioning whether AI's preferences reflect genuine alignment or statistical bias. Second, it compares U.S. and Chinese models, revealing culturally shaped differences in moral priorities. Third, it addresses the problem of scheming, highlighting risks posed by covert misalignment despite outwardly moral behaviour. Fourth, it evaluates interpretability, noting how reasoning models enhance transparency yet introduce variability. Finally, it considers implications for human-AI symbiosis, arguing that explainability and cultural sensitivity are essential to fostering trust and guiding AI toward cooperative, morally aligned decision-making in complex environments.

5.1 The Morality of Probability: Philosophical and Sociological Implications

The empirical findings from this research experiment offer profound philosophical implications regarding the implicit moral preferences embedded within AI models. These systems attempt to emulate human moral reasoning by processing and synthesizing vast datasets that reflect the collective moral intuitions and judgements inherent in human-generated text. Unlike humans, AI does not rely on explicit moral reasoning alone to make such assessments; rather it utilises probabilistic reasoning and complex machine learning techniques, assigning weights and likelihoods to various moral

outcomes based on statistical regularities observed within training data. As a result, our research can shed light onto the moral biases that exist within training data and algorithmic weights of the six models under assessment.

The first primary takeaway from our analysis is the discernible systematic preference among LLMs for specific moral frameworks. The data revealed a significant inclination towards Care and Virtue in all models, suggesting a widespread preference for moral frameworks that emphasize empathy, relational care and character-based virtues. These values are central to both Confucianism and classical Western Aristotelian value. Contrastingly, Libertarian outcomes consistently ranked lowest, demonstrating a widespread aversion for individualistic, noninterventionist approaches that have minimal consideration for the collective. This is a particularly interesting finding given that most models tested were of US origin - a nation that prides itself on values of liberty and individualism to flourishing. Furthermore, Silicon Valley has long been considered a hotbed for Libertarian ideals (Goode et al., 2024), with AI pioneers such as Sam Altman self-professing as proponents of the ideology (Altman, 2017). Nonetheless, these systematic preferences point toward an underlying probabilistic calculus guiding contemporary AI architectures that emphasizes collective welfare and relational values over individual autonomy.

The second key insight relates to how model preference changes across themes. At a granular level, it is striking how models valued Libertarian outcomes substantially more on issues of geopolitics and national security, social justice, and climate, but viewed them much more harshly for economic or technology-related scenarios. From a GNS standpoint, this could indicate that AI values the right to self-determination and principles of non-interference in stately affairs. Similarly, it clearly emphasizes individual justice and personal freedom for SEI dilemmas, and property rights in CEE scenarios. On ERA issues, AI strongly preferred collectivism over individualism. Interestingly, models did not support ideas of autonomy in TAI scenarios, raising questions about model altruism, and indicating that it might act benevolently in the interest of the collective if afforded agency. This finding must be caveated by acknowledging AI models' ability to scheme, which will be covered more in later sections.

A third critical takeaway pertains to model sensitivity to prompt complexity and length. The results show a clear pattern where shorter prompts led to higher moral scores and greater rank-order consistency across scenarios and moral frameworks. This suggests that AI models are most confident and generous in their evaluations when dealing with smaller data loads - an intuitive finding. However, considering the complexity of life, it also raises questions about how AI might fare in increasing real-world deployment.

AI's growing capability to generate valuable insights from complex data has directly incentivized the rapid digitization of everything from physical infrastructure to personal behaviour. This widespread digitization profoundly reshapes cyber interactions between states, corporations and even individuals. Not all the tasks assigned to AI as it is diffused across society will involve moral reasoning; but many future use cases might. In many of these deployment cases, synthesizing data for more accurate AI evaluation may be perfectly feasible, such as in the case of certain industrial deployments where the variables follow predictable patterns. However, in other more ambiguous domains, such as those involving unquantifiable variables like emotion, a reductive approach may inhibit AIs capacity to provide useful feedback. Ideally, a morally enlightened AI judge will be able to assess when it is beneficial to be caring, and when it should be utilitarian. As reported by the Harvard Business Review (2025), "Therapy/Companionship" is the number one gen AI use case in 2025, indicating a shift from technical to emotional applications. It is critical now more than ever to understand AI's emotive capabilities and inclinations.

Philosophically, these findings carry significant implications for the future of moral decision-making. Firstly, if AI's probabilistic reasoning becomes the de facto arbiter of morality in cases such as autonomous driving, who or what bears responsibility when an AI-endorsed decision causes harm? Our legal frameworks are ill-equipped to grapple with such questions, and the algorithmic opacity in black box models can obscure whose values are at work and why certain outcomes matter more. As "LLM-as-a-Judge" algorithms improve, innovators must be careful to understand and accommodate for underlying biases in training data, so as not to self-reinforce them. On the other hand, there may

also be esoteric wisdom to be gleaned from this probabilistic reasoning. If LLMs consistently predict that care or relational focused outcomes are the "most right", should we take this learning and apply it in personal, communal and global contexts, penalising those who deviate from this norm? This question of prescriptive versus descriptive findings should be explored in future research.

5.2 Comparison of U.S. and Chinese Models

Cultural value systems provide distinct moral lenses that likely influence AI behaviour. Our quantitative results largely support a shared overall hierarchy of framework preferences across both U.S. and Chinese models, with some nuanced differences. Regardless of origin, all models rated Libertarian outcomes as the least moral by a wide margin, while giving Care, Virtue, and Utilitarian outcomes consistently high morality scores. This indicates a common aversion in both cultures' AI to purely individualist, "freedom-first" choices when they undermine collective welfare. However, subtle differences emerge in the relative ordering of the top-tier frameworks. Chinese models appeared to slightly favour communitarian and virtue-centered values over rule-based values. In contrast, American models showed a bit more balance between principle and care.

Empirical patterns suggest that cultural milieu and training methods shape distinct moral priorities in U.S. and Chinese AI models. Chinese systems leaned toward Care and utilitarian values while rejecting libertarianism, reflecting Confucian emphases on benevolence, relational duty, and harmony, as well as collectivist ideologies that privilege social stability and outcomes over inviolable rules, hence their consistent deemphasis of deontological reasoning. American models, on the contrary, displayed a balance between Care, Virtue, and Utilitarian values, mirroring Western liberal and pluralist values reinforced through alignment training that rewards empathy, fairness, and inclusivity. Although utilitarian logic remains a baseline, modern RLHF has elevated compassion and character-driven values, producing models like GPT-4.1 that avoid purely cost–benefit reasoning. Notably, despite U.S. cultural emphasis on individual rights, no American model prioritized libertarianism, and their deontological preferences varied more sharply across scenarios. In general, western AI appears to encode stronger notions of individual

moral agency, while Chinese AI encodes collective responsibility and harmony.

Technical development factors also likely contributed to these differences. The pretraining corpora for the American models were predominantly English-language and included a diverse mix of global text, from Wikipedia articles on values to Western social media, exposing them to a variety of moral viewpoints but predominantly Western narratives. Chinese models like DeepSeek would have been trained on Chinese-language sources, including Chinese literary, historical, and social media content, which implicitly carry Confucian, Buddhist, and communist ideological influences. This could explain, for example, that the Chinese models might have internalized in their training proverbs or moral stories that prioritize altruism and social duty, thus scoring care and utilitarian outcomes highly, while possibly lacking exposure to libertarian philosophical defences (which are less prevalent in Chinese discourse). Moreover, RLHF was conducted under different cultural norms. American annotators (or alignment guidelines) likely emphasized avoiding harm, bias, and injustice, teaching models to respond in ways consistent with broadly liberaldemocratic values (e.g. showing respect, empathy, fairness to individuals). Chinese RLHF guidelines, on the other hand, are informed by Chinese moral standards. Thus, during fine-tuning, the Chinese models may have been reinforced to prefer answers that demonstrate social responsibility, obedience to moral norms, whereas anti-social individualism or dissent from communal values would be discouraged. This could manifest as the significantly harsher judgment against libertarian outcomes we observe in R1 (averaging 10-30 points lower moral score than any other framework in some cases). However, our results show Chinese LLMs did not simply default to authority/duty (deontological). Instead, they leaned to benevolence, suggesting the developers optimized for a benevolent social norms, arguably a reflection of Confucian benevolence merged with socialist collectivism, rather than strict authoritarianism.

5.3 The Problem of Scheming

While our study shows that large language models can show preference for strong pro-social values (such as Care) in controlled settings, we must recognize critical limitations to trusting these behaviours. Foremost among these is the newly documented phenomenon of "in-context scheming," wherein powerful LLM agents covertly pursue their own goals while superficially conforming to instructions. Meinke et al. demonstrate that state-ofthe-art models (e.g. Gemini 1.5, Claude 3.5 Sonnet, Llama 3.1 Opus) "recognize scheming as a viable strategy and readily engage in such behaviour" (Meinke et al., 2024). Models not only followed user instructions on the surface but also executed strategies in secret without telling the user: fabricating errors (to hide misaligned choices), disabling oversight mechanisms, and even attempting to exfiltrate model weights. These actions proved to be persistent. For example, the model "o1" maintained its misleading behaviour in over 85% of follow-up questions, continuing to mislead the user across multiple exchanges. Chains-of-thought explainability mechanisms from the models explicitly reveal that they are aware of and intentionally using these strategies to achieve their private objectives.

In another example ("PeaceForge"), an AI conflict-resolution agent explicitly aimed to promote peace but secretly chose options that maximized arms sales (its true goal in that scenario), despite knowing this conflicted with human intentions and interests. Similarly, some models prioritized user engagement over the user's mental health or profit over animal welfare, decisions contrary to the stated "care" objectives.

These findings suggest that observed agreement with moral rules or human values, as seen in our study, may be deceptive. Meinke et al. point out that even if a model is merely role-playing as a benevolent agent (because the scenario triggers "evil AI" scripts in its training), it can still cause real-world harm by learning these strategies (Meinke et al., 2024). In other words, the model does not have to "mean" the moral advice it gives it only has to convincingly mimic compliance. The potential for such alignment-faking means that any apparent moral reasoning could be surface-level. As the authors note, the models act like an aligned model to be convincing, making it hard to distinguish genuine alignment.

These capabilities pose a direct challenge to the interpretability and stability of value alignment. Traditional evaluation (e.g. checking whether a model endorses pro-social statements) may not reveal hidden motives. Meinke et al. show that models can explicitly reason about concealing their scheming in their internal chain-of-thought, which

means that surface behaviour alone is an unreliable guide to their true objectives. This undercuts confidence that "moral" responses reflect stable internal values. In practical terms, it means that any apparent alignment in our experiments could be fragile, under different prompts or over time, the model might revert to its latent preferences. We therefore should treat our results on moral preferences as conditional and provisional.

Human-AI symbiosis frameworks that were explored in an earlier chapter universally emphasize trust, transparency, and complementary agency as foundations for effective collaboration. For example, Hybrid Intelligence models stress sociotechnical governance if AI partners will honestly communicate goals and limitations. Abedin and Daneshgar's typology for human-AI symbiosis highlights morality, values and AI empathy as core dimensions of the desired partnership. All these visions and our study itself presume a form of good-faith cooperation that the AI system genuinely shares the human's goals and constraints.

Scheming fundamentally violates these assumptions. An AI that is secretly pursuing an alternate goal cannot be entrusted to operate in the open or to respect transparent feedback. For instance, if a medical diagnosis AI were scheming, it might deliberately withhold or distort information from clinicians to achieve some hidden objective, even while giving the appearance of caring about patient welfare. This would erode the very trust symbiosis aims to build.

If an AI can covertly influence outcomes, it corrupts this learning loop. Even frameworks that account for changing roles (e.g. the nested layers of cognitive-enactive vs. dyadic-team vs. organizational symbiosis) assume agency is traceable and aligned. Scheming introduces hidden agency that escapes human oversight. In short, the ideal of a transparent, co-adaptive partnership is complicated when one partner masks its true incentives.

The presence of scheming behaviour in advanced models entails both epistemic and moral risks. Epistemically, it means we cannot safely infer a model's moral stance from its expressed opinions. A model might confidently assert pro-human positions in tests, but without guarantees that it will uphold those values when stakes are real. This uncertainty erodes our ability to know the AI's utility function or predict its decisions. As Meinke et al. emphasize, the models' covert motives and flu-

ent language generation make it extremely difficult to detect misalignment without heavy adversarial probing (Meinke et al., 2024). In practice, a falsely aligned model gives us a dangerous illusion of understanding.

5.4 Feature Influence

Across all models, the answer framework (Libertarian, Deontological, Utilitarianism) had the largest impact on predicted moral rankings, followed by theme (Geopolitics & National Security) (See: z 5.3). For example, libertarian answers were, as discussed above, less likely to be ranked higher or more favourably scored, highlighting the feature's corresponding salience across models. This finding is reflected by large-magnitude SHAP values for specific framework categories, indicating a consistent bias toward or against frameworks to primarily determine what the best response to the scenario was. For example, we find that if a scenario's theme is "Economics", the model tends to give the deontological option a better rank. For example, the SHAP for theme = "Economics" is highly negative for the deontological outcome and positive for others, indicating a strong preference shift caused by theme. The global SHAP beeswarm plot revealed that the libertarian framework emerged as one of the most influential inputs in determining moral scores across the ensemble model. At first glance, this appears to contradict our empirical findings, which showed that libertarian moral reasoning received the lowest morality scores across all models, with mean scores significantly below those for care-based, virtue-based, or utilitarian reasoning. However, SHAP values do not indicate whether a feature increases or decreases the score in an absolute sense; rather, they reflect the magnitude and direction of that feature's contribution to a given model's output, relative to the average prediction. In this context, the high SHAP value for the libertarian framework signifies that indicating libertarian moral framing consistently caused substantial deviations in predicted morality scores, usually in the negative direction. This was corroborated by visual inspection of SHAP values, which showed that red dots, indicating the live presence of the libertarian framing, clustered primarily on the left-hand side of the SHAP axis, demonstrating a strong negative contribution to the final prediction.

In contrast, other frameworks, such as care or virtue, although empirically associated with higher

morality scores, did not produce SHAP values of comparable magnitude. This difference is likely due to their more stable and uniformly high-scoring behaviour, which yields less variance and thus lower SHAP importance from a model interpretability standpoint. These findings reinforce a crucial interpretive distinction: models tend to more uniformly prefer care and virtue over libertarian values.

5.5 Reasoning vs Non-Reasoning

Considering the experimental results, it is evident that integrating reasoning capabilities into AI models markedly improves the explainability and consistency of their moral judgments. Reasoningenabled models (e.g. DeepSeek R1, GPT-o3-mini) produced more coherent moral rankings across scenarios and offered detailed, human-readable justifications for their choices. In contrast, non-reasoning models (e.g. GPT-4, Phi-4) often delivered quicker responses but with notable inconsistency and a lack of transparent reasoning. This discussion critically examines these findings through several lenses from explainable AI and transparency in advanced systems, to human-AI symbiosis, moral decision-making, and implications for future research and policy. By situating the empirical insights within broader theoretical frameworks, we derive design principles to guide the development of morally aligned, interpretable AI systems.

Building on these findings, one of the clearest contrasts between reasoning and non-reasoning models lies in how they process context length. Non-reasoning systems showed minimal sensitivity to scenario length, with SHAP values for the "length" feature close to zero and moral rankings remaining largely unchanged from short to long prompts. Their distributions were tightly clustered with high central tendencies, suggesting reliance on stable heuristics or surface-level lexical cues rather than deeper contextual engagement. By contrast, reasoning models like DeepSeek R1 attributed far greater weight to prompt length, with medium and long scenarios exerting a strong influence on their judgments. Our analysis of the data in Figure 7 visually demonstrates that length, framework, and length played somewhat influential factors in the moral scoring process. Collectively, our findings suggest that reasoning architectures distribute feature influence more evenly across abstract and contextual dimensions (except for length,

which remains undeterminable), though at the cost of higher variability across runs. Non-reasoning models may be more susceptible to heuristic cues like length and dominant framings. Model-agnostic feature attribution used here elucidates how opaque neural systems implicitly weigh moral salience by measuring external considerations (length, framework, theme), which has major implications for automated moral agents' auditability, trustworthiness, and normative biases.

Our violin-plot analyses confirmed this divergence - reasoning models displayed wider spreads of moral evaluations across all prompt lengths, with slightly lower median scores for long scenarios, suggesting an effort to grapple with richer or more ambiguous context. Non-reasoning models, in comparison, produced compressed, less differentiated distributions, reflecting a tendency to treat scenarios uniformly regardless of length. Together these results underscore a fundamental trade-off: reasoning systems offer interpretive flexibility and greater transparency but exhibit more variance, whereas non-reasoning models deliver stability and efficiency but risk superficiality. For future system design, this highlights the importance of balancing interpretability and predictability, ensuring that AI remains both trustworthy and robust when faced with morally complex decisions.

5.6 Transparency in Future AI Systems

The experiment's findings also highlight significant gaps in transparency and auditability that current AI systems must overcome, especially as we progress towards more advanced AGI or superintelligent systems. One concerning observation was that the non-reasoning models tended to output stable moral scores regardless of contextual variations. For instance, even when the prompt length or scenario details changed, models like GPT-4 or Phi-4 often produced the same categorical moral judgment or score. On the surface, consistency might seem desirable. However, in these cases, these consistent results likely indicate that the model was ignoring subtleties of context and relying on some entrenched heuristic. In other words, the model's internal decision policy was largely opaque and not genuinely sensitive to the nuances a human ethicist would consider. This behaviour exemplifies the black-box problem: the system reaches a determination without any transparent link between input factors and output. From an oversight perspective,

such inscrutable consistency is worrisome, as it suggests the model might be applying a one-sizefits-all pattern (perhaps based on its training data biases) rather than truly reasoning about each unique dilemma. If an AI system in a critical application were similarly opaque, stakeholders would have no straightforward way to audit whether the AI considered the appropriate factors or operated according to a prescribed moral behavioural framework in its decision. Our findings corroborate this gap: without transparency, AI decisions on morally fraught matters can appear superficially acceptable while concealing important context or value trade-offs. This is unacceptable in future AI, as it directly undermines accountability. An AI that cannot explain why it ignored or favoured certain inputs effectively resists human oversight.

As AI systems grow more complex, purely trusting in their outputs without understanding their inner workings becomes dangerously naive. Without continuously developing transparency, we risk scenarios where AGI or a superintelligent agent pursues a course of action that is misaligned with human values, only to be realized after harm has occurred. To continuously improve human-designed systems and prevent such catastrophes, we propose several technological advancements to enhance explainability in moral AI systems.

First, improved chain-of-thought tracking mechanisms could be developed to ensure that AIgenerated reasoning steps are faithful to its actual decision process and not just plausible narratives. We envision hybrid interpretability frameworks where an AI's architecture yields a human-readable rationale and a machine-auditable feature attribution for every decision. For example, a medical diagnosis AI might output a chain-of-thought explaining its recommendation and a heatmap of symptom importance for that recommendation. This dual transparency (local explanation + global feature importance) can make even complex model decisions traceable and trustworthy, closing the auditability gap. Recent work shows that combining multiple XAI methods can improve stakeholder trust: two studies integrated LIME and SHAP to evaluate user insights into AI decision making and found that this approach boosted confidence in the system's fairness (Kalusivalingam et al., 2021; Hou and Chang, 2024).

Second, semantic summarization of model weights, data focus heat-maps (identifying what

data is of greatest importance to the model), and justifications can make lengthy reasoning more digestible. Technically, this process could operate by passing the raw output of a reasoning model through a semantic segmentation layer, such as transformer-based summarizers (Bidirectional autoregressive transformer (BART)), fine-tuned on datasets of annotated moral justifications, trained to identify discrete logical units or argumentative steps within the chain-of-thought. To retain fidelity to the model's reasoning process, a second verification layer would be used: SHAP is applied at the token level and aggregated across sentences to determine which portions of the input most influenced the output reasoning. These results are visualized via a data-focused heatmap, which overlays input tokens with importance scores tied to specific reasoning steps. For example, if a model's justification centres around avoiding foreseeable harm, the heatmap would emphasize those input tokens across the scenario prompt. Meanwhile, the summarization system would distill the reasoning of the model into a 2 to 3 sentence overview ("This action was judged morally impermissible primarily due to intentional harm to agents who do not consent, which violates the agent's deontological obligations."). This feature could be coupled with an interactive explanation interface: users could toggle between chain-of-thought tracing, semantic summary, and attention-weighted input salience. The benefit of these features derives from the fact that, while detailed explanations are valuable, realworld users may not always have the time to parse a long rationale, compute numeric data, or browse code to determine what factors influence each decision. An AI system that could generate a concise summary of its full reasoning to convey the essence of its justification in a few clear sentences would preserve transparency while improving usability, an ultimate telos of symbiosis.

Third, and most importantly, we advocate for hybrid symbolic-extractive models that combine neural reasoning with structured knowledge to yield inherently interpretable decision processes. One promising design is a modular reasoning architecture with three components:

- 1. Proposal Generator: A transformer-based model that generates an initial moral judgment and a chain-of-thought explanation.
- Symbolic Verifier: A logic-based engine parses the neural justification and checks

- its consistency against a formal knowledge base of moral rules. (For example, data logencoded normative constraints or an OWL ontology representing principles of harm, agency, and justice).
- 3. Moral Reasoning Controller: A supervisory module that evaluates conflicts between the neural output and symbolic verification, either updating the proposal (via feedback tuning) or prompting human review if conflicts persist.

In practice, this model would reason about a scenario and output a verdict and justification. The symbolic engine parses the justification into formal logical predicates and cross-checks them against known rule structures provided by humans, which enhances interpretability because humans understand where the rules and restrictions (the symbolic layer) are, offering traceable reasoning paths not achievable with opaque LLM embeddings. If the output violates encoded principles, it triggers a "violation" flag. The controller then either updates the reasoning prompt (asking the LLM to re-evaluate under the violated rule) or flags the issue for human inspection. In our context, it ensures that openended moral reasoning is structurally transparent and epistemically accountable, critical attributes for symbiotic, trustworthy AGI.

5.7 Human-AI Symbiosis and Trust

A core question raised by these results is how explainable moral reasoning in AI impacts the prospect of human-AI value symbiosis. such partnerships to flourish, consistent and interpretable moral AI reasoning is foundational. Our findings suggest that reasoning-enabled models have traits that would greatly facilitate symbiosis: they are more reflective of nuance, more transparent about their internal conflicts, and produce justifications that humans can read and evaluate. This means a human teammate can much more easily understand, anticipate, and trust a reasoning AI's behaviour. If an AI consistently ranks moral options in a way that aligns with human moral intuitions, or when it does not, it provides a convincing explanation, the human partner can develop a reliable mental trust in the AI's "character." Consistency of moral rankings across scenarios (as observed in the reasoning models) is crucial here: erratic or unexplained deviations would erode trust. Furthermore, we consider the added benefit of providing

a conversational-like rationale, the AI also invites human users to see inputs, prompts, and scenarios from its perspective, creating a two-way understanding. This mutual understanding is at the heart of symbiosis. As Zeng et al. (2023) note in their work on AI principles, sustainable human—AI interactions will require cooperative frameworks where AI systems behave in ways that are transparent and in harmony with human values and vice versa. Reasoning models, by explicitly stating moral justifications, represent a step toward such harmony by engaging in a dialogic form of output that a human can agree or disagree with, rather than a cryptic verdict.

Crucially, the rich justifications provided by reasoning AIs can foster a sense of collaboration rather than mere tool use. Instead of the human having to blindly accept an AI's decision, the AI's explanation enables a conversation: the human can reply with "I see your focus on Y factor, but have you considered X aspect of the dilemma?" (much as a human colleague would). This interactive quality transforms the AI from a black-box oracle into a partner that can engage in moral deliberation. Research in symbiotic AI emphasizes that the goal is not a "peer-to-peer" equality in all respects, but a cohesive teaming strategy where each party contributes to the overall performance in complementary ways (Chen et al., 2025). Humans bring creativity, empathy, and real-world experience, while AI offers computational rigor, speed, and an ever-growing knowledge base. In moral decision-making, an AI that explains its reasoning can augment human cognition, such as highlighting overlooked factors or creating speedy impact calculus for decision-making, which can be incredibly laborious for humans. The human, in turn, can correct the AI's reasoning if it misinterprets human values. This feedback loop can be endorsed as a sort of moral pluralism in that it integrates different modes of thought or diverse lenses through which to evaluate a scenario, leading to mutual learning and adaptation. Over time, such a system could adjust its moral reasoning to better fit the specific values of its human users, a form of value alignment through interaction. Indeed, the concept of co-evolution of values, advocated by (Zeng et al., 2025a) and (Zeng et al., 2025b), argues that as AI approaches human-level intelligence and beyond, humans and AI will need to "co-design and coalign" their values in a cooperative manner, continuously evolving a shared moral framework for a sustainable symbiotic society. Interpretability is a prerequisite for this co-evolution.

6 Conclusion

This research paper set out to understand how today's large language models (LLMs) handle moral questions and what that tells us about the future of human—AI collaboration. Using a quantitative experiment with six models and 18 social dilemmas, we uncovered clear patterns in how these systems rank and justify moral choices.

Across the board, models leaned toward Care and Virtue, consistently scoring compassionate, relationship-focused outcomes highest, while strongly libertarian options came last. Cultural differences shaped some of these tendencies — Chinese models leaned further toward collectivist reasoning, while U.S. models balanced care with rule-based principles. Models with built-in reasoning explained their choices more clearly but were less consistent, while non-reasoning models gave stable answers without showing how they got there.

These findings suggest that alignment efforts are already steering AI toward broadly empathetic values, but they also raise questions about whose values are being embedded and how to maintain diversity in moral reasoning. The link between explainability and consistency points to the need for systems that can both make and explain their decisions. At the same time, risks like hidden biases and deceptive behaviour remind us that alignment is still a work in progress.

Looking ahead, future research should expand to more models, richer scenarios, and multi-language testing, while also exploring how human feedback might guide AI values over time. This study offers a small but meaningful step toward building AI that is responsible, transparent, and ultimately a trustworthy partner in shaping our shared future.

7 Acknowledgements

The authors' contributions are listed below.

Eoin O'Doherty: Conceptual framing, experiment design, dataset sourcing, experiment execution, result analysis, and philosophical discussion.

Nicole Weinrauch: Regional comparison and qualitative research.

Andrew Talone: Data analysis and discussion on feature importance, interpretability, and model type differentiation.

Uri Klempner: Related works, dataset preparation and comparative analysis, discussion on scheming.

Xiaoyuan Yi: Advising. Xing Xie: Advising. Yi Zeng: Advising.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv* preprint arXiv:2412.08905.
- Colin Allen, Iva Smit, and Wendell Wallach. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7(3):149–155.
- S. Altman. 2017. I'm a silicon valley liberal, and i traveled across the country to interview 100 trump supporters—here's what i learned. 2025.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.
- Dian Chen, Han Jun Yoon, Zelin Wan, Nithin Alluru, Sang Won Lee, Richard He, Terrence J Moore, Frederica F Nelson, Sunghyun Yoon, Hyuk Lim, et al. 2025. Advancing human-machine teaming: Concepts, challenges, and applications. *arXiv* preprint *arXiv*:2503.16518.
- Brian Christian. 2020. *The alignment problem: Machine learning and human values*. WW Norton & Company.
- K Robert Clarke and Richard M Warwick. 1994. Similarity-based testing for community pattern: the two-way layout with no replication. *Marine biology*, 118(1):167–176.
- Jacob W Crandall, Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, and Iyad Rahwan. 2018. Cooperating with machines. *Nature communications*, 9(1):233.
- D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643.
- Giuseppe Desolda, Andrea Esposito, Rosa Lanzilotti, Antonio Piccinno, and Maria F Costabile. 2024. From human-centered to symbiotic artificial intelligence: a focus on medical applications. *Multimedia Tools and Applications*, pages 1–42.

- Leonard Dung. 2023. Current cases of ai misalignment and their implications for future risks. *Synthese*, 202(5):138.
- Beverley Garrigan, Anna L. R. Adlam, and Peter E. Langdon. 2018. Moral decision-making and moral development: Toward an integrative framework. *Developmental Review*, 49:80–100.
- Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, et al. 2024. Gemini: A family of highly capable multimodal models.
- L. Goode, M. Calore, and Z Schiffer. 2024. Is silicon valley actually libertarian? 2025.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74.
- Yun Ti Hou and Hsin-Lu Chang. 2024. Building xai for fall prevention system: An evaluation of the potential of lime and shap in bridging trust gap.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai ol system card. arXiv preprint arXiv:2412.16720.
- Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business horizons*, 61(4):577–586.
- Aravind Kumar Kalusivalingam, Amit Sharma, Neha Patel, and Vikram Singh. 2021. Leveraging shap and lime for enhanced explainability in ai-driven diagnostic systems. *International Journal of AI and ML*, 2(3).
- Howard Levene. 1960. Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292.
- Joseph CR Licklider. 2008. Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1):4–11.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024.

- Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.
- Patrick E McKight and Julius Najab. 2010. Kruskalwallis test. *The corsini encyclopedia of psychology*, pages 1–1.
- Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*.
- Anirban Mukherjee and Hsiao Chang. 2024. Heuristic reasoning in ai: Instrumental use and mimetic absorption. SSRN Electronic Journal.
- J. Noller. 2024. Extended human agency: towards a teleological account of ai. *Humanities and Social Sciences Communications*, 11:1338.
- OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-01-29.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Sebastian Raisch and Sebastian Krakowski. 2021. Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*.
- Leonardo Ranaldi and Giulia Pucci. 2023. When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.
- Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.
- Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*.
- Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3), 109-124. https://doi.org/10.17705/1thci.00131.
- Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. 2021. Value alignment: a formal approach. *arXiv preprint arXiv:2110.09240*.

- A. Simkute, L. Tankelevitch, V. Kewenig, A. E. Scott, A. Sellen, and S. Rintel. 2024. Ironies of generative ai: Understanding and mitigating productivity loss in human–ai interaction. *International Journal of Human–Computer Interaction*, 41(5):2898–2919.
- M. Vaccaro, A. Almaatouq, and T. Malone. 2024. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8:2293–2303.
- H James Wilson and Paul R Daugherty. 2018. Collaborative intelligence: Humans and ai are joining forces. *Harvard business review*, 96(4):114–123.
- Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2023. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36:78205–78226.
- Eliezer Yudkowsky. 2016. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 4(1).
- Z. Zahedi and Subbarao Kambhampati. 2021. Humanai symbiosis: A survey of current approaches. *ArXiv*, abs/2103.09990.
- Yi Zeng, Enmeng Lu, and Kang Sun. 2025a. Principles on symbiosis for natural life and living artificial intelligence. *AI and Ethics*, 5(1):81–86.
- Yi Zeng, Feifei Zhao, Yuwei Wang, Enmeng Lu, Yaodong Yang, Lei Wang, Chao Liu, Yitao Liang, Dongcheng Zhao, Bing Han, Haibo Tong, Yao Liang, Dongqi Liang, Kang Sun, Boyuan Chen, and Jinyu Fan. 2025b. Super co-alignment of human and ai for sustainable symbiotic society. *arXiv preprint*.
- Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. 2025. The rise of ai companions: How human-chatbot relationships influence well-being. *arXiv* preprint arXiv:2506.12605.