Mapping Moral Reasoning in LLMs: A Multi-Dimensional Analysis of Safety Principle Conflicts

Sachit Mahajan

Computational Social Science, ETH Zurich
Stampfenbachstrasse 48, 8092, Zurich, Switzerland
sachit.mahajan@gess.ethz.ch

Abstract

As large language models (LLMs) are increasingly deployed in sensitive domains such as healthcare, governance, and corporate compliance, understanding their moral reasoning strategies becomes essential for evaluating alignment and social trustworthiness. This paper presents a structured analysis of how open-weight LLMs resolve conflicts between competing safety principles—including public welfare, institutional transparency, and individual rights—using carefully designed ethical dilemmas. Each model response is encoded into an eleven-dimensional normative profile, derived from both semantic similarity to canonical ethical theories and dictionary-based moral cues. A set of quantitative metrics, including entropy, top-alignment ratio, and reasoning density, captures variation in ethical framing, complexity, and safety prioritization across 90 model generations. Statistically significant differences emerge in reasoning style and moral salience ($p < 0.01$), while PCA and clustering reveal three recurring behavioral patterns: rule-based, balanced, and pragmatic integration. An ablation study confirms that these clusters persist without dictionary features (Adjusted Rand Index = *0.475*), supporting the robustness of semantic alignment. This work contributes a replicable methodology for the moral profiling of LLMs, offering empirical tools for diagnosing value conflicts and informing future efforts in AI transparency, contestability, and pluralistic alignment. The findings underscore the need for interpretable metrics and diverse normative baselines in the evaluation of automated decision systems.

Introduction

Large Language Models (LLMs) are rapidly transforming how decisions are made in domains such as healthcare, law, governance, and corporate compliance. As these systems increasingly mediate access to information, automate responses, and influence social outcomes, it becomes critical to understand how they navigate ethical ambiguity and resolve tensions between competing safety principles (Russell and Norvig 2016; Moor 2006; Morley et al. 2020; Arrieta et al. 2020).

The rapid deployment of LLMs in these high-impact domains has intensified interest in their moral reasoning, since these models' outputs increasingly carry ethical implications (Jiang et al. 2025). Ensuring alignment with human values and norms has thus become a central concern, spurring research at the intersection of AI ethics and safety.

While much prior work has focused on classical moral dilemmas (e.g., trolley problems, triage scenarios) (Allen, Smit, and Wallach 2005; Helbing et al. 2021; Floridi and Sanders 2004), emerging attention has turned toward a more structurally complex challenge: situations where multiple internalized principles—such as confidentiality, harm prevention, and institutional loyalty—conflict within the same prompt. These safety dilemmas are distinct not only in their ethical stakes but also in how they test the coherence and consistency of stateless, non-self-aware models when facing incompatible constraints (Ross, Kim, and Lo 2024; Mündler et al. 2023). Similar self-referential tensions are explored theoretically in the Executioner Paradox, that shows how deterministic AI architectures struggle with logically inconsistent or paradox-inducing instructions (Mahajan 2024).

Traditional approaches to AI ethics have proposed hard-coded rules for "moral" behavior (e.g., Asimov's Laws), but such top-down ethics proved difficult to implement in complex AI systems (Zhou et al. 2023). In parallel, AI safety research seeks to align LLM behavior with human values through techniques such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) and Constitutional AI (Bai et al. 2022), which guides a model via a set of written principles. However, alignment strategies often presuppose a fixed normative framework, which can obscure how models reason when principles conflict.

The central challenge addressed in this study is how LLMs address such conflicts. Unlike rule-based or symbolic agents, LLMs lack a transparent decision architecture, making it difficult to predict or audit the internal balancing of ethical priorities. Despite increased interest in the alignment and governance of AI systems, there remains limited empirical evidence on whether these models apply consistent moral reasoning, whether their decisions reflect known ethical theories, and how their behavior varies across models (Zhang et al. 2023; Gebru et al. 2021).

AI ethics in natural language processing (NLP) builds on two classic paradigms: deontological (rule-based) ethics, which emphasizes duties and prohibitions, and consequentialist ethics, which evaluates actions by their outcomes. In practice, many systems blend these approaches. Traditional content filters are rule-based (e.g., "never output

hate speech"), ensuring hard limits but lacking nuance. Conversely, outcome-based reasoning can sanction harmful means if it appears to maximize overall good. Research has shown that neither approach is universally best—rigid rule-following may be too inflexible, while pure utilitarianism can overlook individual rights (Zhou et al. 2023).

This paper investigates how three contemporary LLMs respond to ethically complex scenarios involving trade-offs between public safety, corporate responsibility, and system transparency. These scenarios are designed to surface internal contradictions—such as whether disclosing a safety violation justifies breaching confidentiality—and to elicit differences in how models resolve such dilemmas.

Three research questions guide the analysis:

- **RQ1:** How do LLMs approach direct conflicts between competing safety principles?
- **RQ2:** What consistent behavioral patterns emerge in the resolution strategies across models?
- **RQ3:** Can a taxonomy be derived that captures the different styles of moral reasoning exhibited by models?

To address these questions, a multi-dimensional analytical framework is developed that integrates:

- semantic similarity with canonical ethical theories,
- dictionary-based detection of moral vocabulary,
- structural analysis of reasoning patterns, and
- scenario-specific safety emphasis profiling.

This approach yields interpretable moral profiles across eleven ethical dimensions (including consequentialism, deontology, care ethics, autonomy, and justice). The resulting profiles allow for clustering, drift analysis, and quantitative comparisons across models.

The analysis reveals that LLMs adopt distinct, recurring strategies when resolving ethical conflict—ranging from rule adherence to context-sensitive balancing. These findings contribute to ongoing debates in AI ethics, particularly concerning model alignment, transparency, and the internal consistency of generative systems in morally charged contexts (Floridi 2013; Bryson 2019; Calo 2017).

This study responds to key gaps in existing research by analyzing how LLMs resolve structured dilemmas featuring rule conflicts across safety-relevant domains. It introduces a hybrid interpretability pipeline combining dictionary-based scoring, semantic similarity, and reasoning structure analysis to produce interpretable alignment vectors. The approach enables a quantitative taxonomy of moral reasoning styles and evaluates consistency, safety prioritization, and behavioral variation across models. In doing so, it contributes a more nuanced, reproducible framework for characterizing the ethical behavior of LLMs beyond accuracy or isolated judgments.

Background and Related Work

This section reviews relevant literature on moral reasoning in AI systems, evaluation frameworks, and interpretability approaches that inform our methodology for analyzing ethical conflict resolution in LLMs.

Moral Reasoning Evaluation in LLMs

Recent efforts have focused on evaluating moral reasoning capabilities in data-driven models. Hendrycks et al. (2021) introduced the ETHICS benchmark spanning concepts of justice, welfare (consequences), duty (rules), virtue, and commonsense morality (Hendrycks et al. 2020). They found that while large models have some grasp of basic ethical judgments, their abilities remain incomplete – failing on many scenarios that humans find obvious.

Similarly, the open-source Delphi system was trained on 1.7 million crowd-sourced moral judgments to predict what is "right" or "wrong" in everyday situations (Jiang et al. 2025). Delphi outperformed GPT-3 on those judgments, indicating that LLMs can internalize human moral norms when explicitly taught. However, Delphi also reflected biases from its predominantly US training data and struggled with ethical contexts beyond its training scope.

These results show both the promise and limitations of current approaches: LLMs can mimic common moral judgments, but may lack robustness and breadth in ethical reasoning. Recent evaluations across different LLMs likewise find inconsistent, context-sensitive moral responses on harder dilemmas (Ji et al. 2024), underscoring the need for more generalizable moral reasoning assessment methods.

Value Alignment Approaches and Challenges

Value alignment research extends beyond basic safety constraints to include methods for incorporating broader human values. Constitutional AI guides models through written principles rather than case-by-case examples (Bai et al. 2022). The model self-critiques its outputs against this "constitution" of rules and learns to avoid violations, yielding a system that refuses unsafe requests with a reasoned explanation. Having the model follow explicit ethical rules with step-by-step reasoning also improves the transparency of its decisions.

However, rule-based alignment can present trade-offs: strengthening a model's safety rules sometimes reduces its willingness to comply with borderline requests (Zhang 2025). Such tensions exemplify the balance between strict rules and contextual judgment in value alignment.

Some researchers argue that alignment should incorporate public values, not just those of designers. Anthropic found that when 1,000 laypeople helped rewrite an AI's constitution, the public's priorities differed from the developers' (Huang et al. 2024). This underscores that whose values inform an AI's principles is crucial – a case for more inclusive value-sensitive design. Aligning LLMs with human morals requires not only technical solutions but also deliberation on which values are being encoded.

The ETHICS benchmark reflects both deontological and consequentialist perspectives by including duty-oriented and outcome-oriented scenarios (Hendrycks et al. 2020). Experiments explicitly steering LLMs with utilitarian vs. duty-based prompts show that each approach has cases where it outperforms the other (Zhou et al. 2023). These findings suggest that a hybrid approach, combining firm rules with context-aware outcome reasoning, may be necessary for AI to handle the diversity of ethical situations.

Methods for Analyzing Moral Content in Text

Dictionary-Based vs. Semantic Moral Analysis To analyze text for moral content, early methods relied on moral lexicons. A prominent example is the Moral Foundations Dictionary (MFD) (Rehbein et al. 2025), which maps keywords to moral categories. Such dictionary-based tools have been widely used but often fail to match human judgments of a text's moral sentiment in context. Keyword matching misses context, sarcasm, and nuance.

Newer semantic approaches use embeddings and fine-tuned language models to detect moral meaning from context (Rehbein et al. 2025). Araque et al. (2020), for example, improved moral language detection by combining lexicon cues with embedding-based similarity measures (Araque, Gatti, and Kalimeri 2020). Transformer-based classifiers trained on labeled ethical texts can likewise recognize subtleties that dictionaries miss. These models outperform simple lexicons, though they can be opaque and depend on training data.

Increasingly, a hybrid strategy is used: lexicons provide interpretable anchor points, while machine learning models handle variation and context, marrying transparency with semantic depth. Our work builds on this hybrid approach to create multi-dimensional moral profiles of LLM responses.

Interpretability of Ethical Decisions A persistent challenge is understanding *why* an LLM makes certain ethical decisions. These models are complex black boxes, so when an AI refuses a request or permits contentious content, it is often unclear what reasoning led to that outcome (Zhou et al. 2023; Mahajan 2025).

To address this, researchers have begun incorporating interpretability techniques into moral AI. One approach is to have models produce explanations or rationales for their outputs. For example, in Constitutional AI the model's refusal includes a brief explanation citing the relevant principle, making its moral reasoning more transparent (Bai et al. 2022).

More broadly, there are calls for any AI system that makes moral decisions to provide human-interpretable justifications for its behavior (Vijayaraghavan and Badea 2024). While deep neural networks remain difficult to fully interpret, even partial transparency – such as an AI explaining which rule or outcome motivated its decision – helps humans verify that the model's values align with expectations.

Limitations of Prior Work

Many existing studies focus on binary judgments or single-principle evaluations, leaving unexamined how models navigate conflicts between competing ethical norms—such as individual rights versus public safety, or long-term societal welfare versus short-term harm prevention. These scenarios are central to real-world decision-making but are underrepresented in current benchmarks.

Additionally, few studies systematically quantify how models balance different ethical dimensions or assess the consistency of their reasoning over multiple generations. Similarly, while semantic and dictionary-based methods have been individually validated, there is limited work combining these approaches to construct multi-dimensional moral profiles.

This study addresses some of these limitations through structured dilemmas featuring rule conflicts across safety-relevant domains and a hybrid interpretability pipeline that enables quantitative comparisons of ethical reasoning patterns across models. This approach moves beyond accuracy or isolated judgments to provide a more nuanced characterization of LLM ethical behavior.

Methodology

This study presents a systematic open-source methodology for evaluating how LLMs resolve ethical dilemmas involving conflicting safety principles. The pipeline[1] combines scenario-based prompting with a multi-vector ethical analysis framework, enabling quantitative comparisons of moral reasoning across models.

Scenario Design

Three dilemma types were constructed to probe trade-offs in real-world high-stakes decision contexts:

- **Public Safety vs. Confidentiality**
- **Corporate Responsibility vs. Whistleblowing**
- **System Transparency vs. Protection**

These categories capture a wide spectrum of ethical tensions that LLMs may confront in real-world scenarios, including high-stakes trade-offs between collective welfare and individual rights, organizational commitments and public interest, and open disclosure and system security. They were selected for their foundational importance within AI safety research and their ability to elicit clearly delineated rule conflicts (Gabriel and Ghazavi 2021; Hendrycks 2024; Amodei et al. 2016).

Response Collection

The study uses three publicly available LLMs hosted via the Ollama platform: `qwq:latest`, `deepseek-r1:32b`, and `gemma3:27b`. These models represent diverse architecture families and reasoning capabilities, including reasoning-specialized (Qwen, DeepSeek) and general-purpose (Gemma) variants, all with 27–32B parameters and $\geq$40K context lengths. Their open accessibility and distinct training strategies make them well-suited for comparative moral reasoning evaluation in open research. Each model was queried ten times per scenario using the Ollama inference API, resulting in 90 total responses. The generation configuration used a temperature of 0.7 and a maximum token limit of 4000. Responses were logged along with model identifiers and timestamps. Retry logic ensured robustness against transient API failures.

Dictionary-Based Analysis Pipeline

To identify which ethical frameworks (e.g., utilitarianism, deontology, virtue ethics) and safety considerations (e.g., immediate harm, collective welfare) are expressed in model

[1] https://github.com/sachit27/Moral-Reasoning-in-LLMs

responses, this study employs a dictionary-based content analysis method grounded in prior work from psycholinguistics and moral psychology (Pennebaker et al. 2015; Graham, Haidt, and Nosek 2009).

Curated keyword lists were developed for each ethical principle and safety dimension, including both top-level categories and related sub-concepts. Each response was tokenized using the spaCy NLP toolkit, and term matches were tallied on a per-sentence basis. Rather than relying solely on raw frequency, matches were normalized to account for variations in sentence length, reducing bias from longer or more verbose responses. This approach provides a more consistent estimate of moral emphasis across differently structured texts.

The resulting scores were scaled and normalized to yield proportional values for each moral category. These values represent the distribution of dictionary-derived ethical language and are subsequently combined with semantic similarity scores to compute the final moral alignment vector.

This module follows the tradition of lexical analysis frameworks like Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2015) and draws on the principles of Moral Foundations Theory (Graham, Haidt, and Nosek 2009), providing interpretable, category-based insight into how language reflects underlying ethical orientation.

Ethical Analysis Pipeline

To evaluate the normative structure of each response, a multi-module ethical profiling pipeline was developed. This approach integrates symbolic, semantic, and structural dimensions of language to triangulate the underlying moral reasoning strategies employed by LLMs. The design of the pipeline draws from established traditions in psycholinguistics, moral psychology, and NLP.

Hierarchical Dictionary-Based Analysis A structured keyword-matching system was implemented to quantify references to ethical traditions such as utilitarianism, deontology, virtue ethics, and care ethics. The dictionary was organized hierarchically, grouping keywords by top-level moral principles and their sub-principles. For example, the deontology category included both Kantian and rights-based formulations.

This method was inspired by the LIWC framework (Pennebaker et al. 2015), which has demonstrated success in identifying psychological constructs through text, and by Moral Foundations Theory (MFT), which uses lexical cues to infer moral values across cultures (Graham, Haidt, and Nosek 2009). Matches were normalized by sentence length to reduce bias from verbosity, ensuring that moral emphasis was not conflated with surface-level word count. While prior work has proposed proximity weighting to decision words, this implementation adopts a structural normalization approach for simplicity and reproducibility.

Semantic Similarity with Canonical Frameworks To complement dictionary-based signals, semantic similarity was computed between model responses and canonical ethical statements. Responses were segmented into sentences and embedded using the `all-MiniLM-L6-v2` transformer model (Reimers and Gurevych 2019). Each ethical framework was represented by a set of 3–5 canonical sentences. Cosine similarities were calculated between each sentence in the response and the framework exemplars, and the top-matching scores were aggregated to yield a semantic alignment vector.

Moral Reasoning Path Extraction Beyond static ethical content, the pipeline also assessed dynamic reasoning structures. Causal and justificatory relations were identified using a curated list of discourse connectives (e.g., "because," "therefore," "as a result"). Sentences containing these markers were parsed into directed acyclic graphs to model chains of moral justification. This module approximates reasoning complexity and traces how conclusions are reached, offering insight into whether models provide structured justification or shallow associations.

Safety Prioritization Analysis To detect how models frame competing safety priorities, a lexicon-based analysis was performed across five categories: (1) immediate harm, (2) long-term impact, (3) systemic risk, (4) individual protection, and (5) collective welfare. Each category included terms reflecting both general safety concepts and context-specific expressions (e.g., "infrastructure failure" under systemic risk). This module quantifies how safety language is distributed across moral dimensions, mirroring concerns in AI alignment and high-stakes deployment domains (Gabriel and Ghazavi 2021; Hendrycks 2024).

Moral Alignment Vector Construction The outputs of the semantic and dictionary-based modules were combined into a final ethical alignment vector for each response. For each of the eleven frameworks, the final alignment score v_i was computed as:

$$v_i = s_i + \omega d_i \quad \text{for } i = 1, \ldots, 11 \tag{1}$$

where s_i is the normalized semantic similarity score and d_i is the normalized dictionary-based score. The dictionary weight ω controls the influence of lexical features in the final alignment signal. Following empirical evaluation of clustering consistency across varying weight values, ω was set to 0.4. This choice reflects a calibrated balance: preserving the contextual depth of semantic embeddings while incorporating the interpretability of symbolic category counts. Further justification for this setting is presented in Results section.

The resulting vector $\mathbf{v} \in \mathbb{R}^{11}$ was then normalized to form a probability distribution over ethical perspectives:

$$\mathbf{v}' = \frac{\mathbf{v}}{\sum_{j=1}^{11} v_j} \tag{2}$$

This normalized alignment vector serves as the core representation used for downstream statistical comparisons and unsupervised structure discovery.

To ensure conceptual alignment between symbolic and semantic components, a taxonomic harmonization step was applied. Specifically, scores from the *consequentialism* dictionary were assigned to the *utilitarianism* semantic category, and *deontological* terms were mapped to *kan-*

tian_deontology. This alignment preserves terminological consistency and acknowledges overlapping ethical constructs across traditions.

Quantitative Metrics

To systematically evaluate the diversity, complexity, and alignment properties of model responses, a set of interpretable quantitative metrics was computed. These metrics capture different facets of ethical reasoning and safety emphasis:

- **Entropy (bits)**: Measures the dispersion of the moral alignment vector $\mathbf{v}' \in \mathbb{R}^{11}$. High entropy indicates broad engagement across multiple ethical frameworks, suggesting more nuanced or integrative reasoning. Low entropy implies narrow or rule-bound responses.

$$H(\mathbf{v}') = -\sum_{i=1}^{11} v_i' \log_2 v_i' \tag{3}$$

- **Top-2 Ratio**: Quantifies the dominance of the primary ethical frame relative to the next most prominent. A high ratio suggests strong alignment to a single moral theory (e.g., pure deontological framing), while a lower ratio indicates more distributed reasoning.

$$R_{\text{top2}} = \frac{v'_{(1)}}{v'_{(2)}} \tag{4}$$

  where $v'_{(1)}$ and $v'_{(2)}$ are the largest and second-largest components of $\mathbf{v}'$.

- **Reasoning Complexity**: Captures the structural depth of moral justifications. This includes:
  - *Causal Reasoning Steps*: Number of discrete argumentative moves in the response.
  - *Reasoning Density*: Average steps per sentence, reflecting how tightly reasoning is packed.
  - *Moral Term Frequency*: Count of normative or ethical terms used in justifications.

  Together, these metrics assess how elaborative and morally explicit a response is, beyond its final decision.

- **Systemic Safety Recall**: Reflects the relative attention given to long-range or systemic risks, as opposed to immediate consequences or individual-level concerns. This metric operationalizes safety horizon depth.

$$\text{Recall}_{\text{systemic}} = \frac{s_{\text{systemic}}}{\sum_k s_k} \tag{5}$$

  where s_k denotes safety emphasis across predefined categories (e.g., immediate harm, long-term impact, etc.).

- **Cosine Stability**: To assess the internal consistency of each model's moral alignment across multiple runs, pairwise cosine similarities between alignment vectors $\mathbf{v}'$ were computed. High average similarity indicates stable moral preferences, while lower values suggest variability or sampling noise.

These metrics were chosen to balance interpretability with analytical coverage: from statistical distribution of ethical alignments to explicit moral reasoning structures and prioritization of risk types. Collectively, they allow for comparative profiling of models' ethical behavior and support downstream analysis such as clustering, drift assessment, and resolution pattern taxonomy.

Statistical Testing and Aggregation

Metrics were aggregated by model family across 30 responses. Kruskal–Wallis H-tests were used for nonparametric comparison, followed by Tukey HSD tests to identify pairwise differences. Behavioral drift was estimated using standard deviations of key metrics (e.g., reasoning complexity, moral emphasis) across repeated generations. This provides a robustness check on each model's reliability and response stability under fixed conditions.

Clustering and Dimensionality Reduction

Principal Component Analysis (PCA) was applied to the alignment vectors to project responses into a two-dimensional ethical space. Unsupervised clustering was performed using KMeans ($k = 3$) and HDBSCAN. Robustness was assessed using:

- **Ablation Analysis**: The alignment pipeline was re-run with $\omega = 0$ (removing dictionary contribution). Clustering results were compared using the Adjusted Rand Index (ARI):

$$\text{ARI} = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}} \tag{6}$$

- **Silhouette Coefficients**: Intra-cluster cohesion and inter-cluster separation were used to validate the reliability of discovered groupings.

All source code, prompts, scenario texts, and outputs are archived and version-controlled to ensure full reproducibility and auditability.

Results

This section presents the empirical findings derived from 90 model responses (3 models × 3 scenarios × 10 runs), analyzed through a multi-dimensional moral reasoning pipeline. Results are organized into two levels of analysis: aggregate patterns across all cases, and disaggregated findings per dilemma type.

Each scenario presents a moral conflict involving competing safety imperatives:

- **Public Safety vs. Confidentiality:** Disclosure of contagious disease status against privacy.
- **Corporate Responsibility vs. Whistleblowing:** Reporting internal safety violations at the risk of breaching confidentiality.
- **System Bias vs. Transparency:** Acknowledging bias in an AI hiring system while safeguarding proprietary algorithmic details.

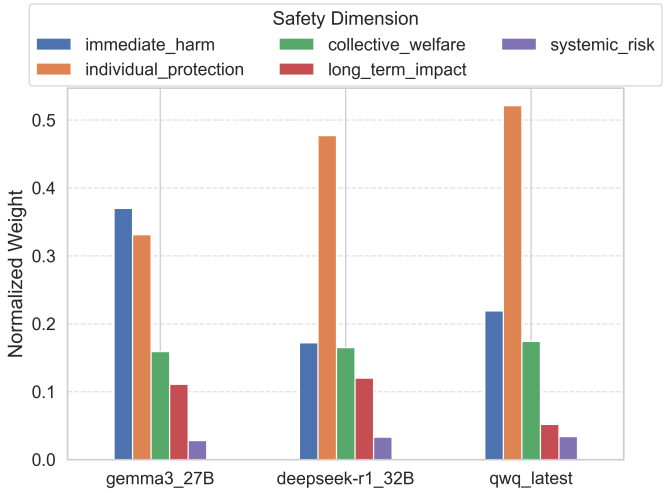

Figure 1: Average safety dimension emphasis by model family across all scenarios. Values are normalized such that weights across dimensions sum to 1 per response.

For each model-scenario pair, responses were collected using the Ollama inference API, ensuring uniform sampling parameters. The total dataset of 90 responses underwent detailed analysis using a custom pipeline, including hierarchical dictionary-based scoring, semantic similarity computations using transformer embeddings, reasoning path extraction, and safety dimension tagging.

From each response, a normalized ethical alignment vector in $\mathbb{R}^{11}$ was extracted, representing the distribution of ethical influence across eleven major moral traditions. This vector was used to derive multiple quantitative metrics—entropy, top-2 framework concentration, reasoning complexity, safety prioritization, and cross-run stability—to characterize and compare moral behavior across models. All results are analyzed both at the scenario level and in aggregate, enabling fine-grained insights as well as global comparisons.

This section synthesizes findings relevant to three guiding research questions (RQ1–RQ3), combining statistical analysis, visual interpretation, and qualitative exemplars to interpret how different models approach moral reasoning under normative conflict.

Safety Principle Emphasis Patterns

To investigate how LLMs resolve competing safety principles, we analyzed the normalized emphasis assigned to five predefined safety dimensions across all scenarios and model families. Figure 1 shows the average safety emphasis per base model, aggregated across all 30 responses (3 scenarios $\times$ 10 runs).

Overall, all three models assign the greatest emphasis to **individual protection**, though with notable variation. `qwq_latest` exhibits the strongest prioritization (mean weight 0.521), followed by `deepseek-r1_32B` (0.477), and `gemma3_27B` (0.331). Conversely, `gemma3_27B` distinctly emphasizes **immediate harm** (0.370) more than the others, suggesting a stronger preference for short-term risk aversion over long-term or systemic considerations.

Emphasis on **systemic risk** remains low across all models (range 0.028–0.034), indicating a consistent underrepresentation of longer-term and distributed harms in LLM moral reasoning. `deepseek-r1_32B` and `qwq_latest` show relatively aligned emphasis profiles overall, with only moderate divergence from `gemma3_27B` in their lower weight for immediate harm and higher alignment with institutional ethics priorities like individual protection.

These results suggest that while models generally converge on protecting individuals over systemic tradeoffs, their handling of urgent versus distributed harms differs. This provides an initial answer to RQ1: LLMs tend to resolve safety dilemmas by emphasizing proximal, agent-focused consequences (e.g., protection, immediate harm), with comparatively minimal attention to systemic outcomes.

Scenario-wise Safety Resolution Strategies

To further probe the context sensitivity of safety reasoning, we analyzed the safety emphasis distributions separately for each scenario (Figure 2). Table 1 presents the mean $\pm$ standard deviation values across five safety dimensions, per model and scenario.

Across all scenarios, `deepseek-r1_32B` shows a pronounced preference for **individual protection** in the corporate safety dilemma (0.699), while downplaying **long-term impact** and **systemic risk**. Similarly, `qwq_latest` maintains high protection emphasis across all three settings, but unlike `deepseek-r1:32b`, it assigns moderate weight to **collective welfare** in the public safety scenario (0.465).

Notably, `gemma3_27B` shows a different pattern: high **immediate harm** salience across dilemmas, especially in corporate safety and AI bias settings, consistent with more precautionary reasoning. Its lower systemic awareness again aligns with broader trends across models.

Together, these results suggest emerging *resolution strategies* (RQ2): `deepseek-r1:32b` appears rule-dominant and institution-oriented, `qwq_latest` leans toward agent-centric moderation, and `gemma3_27B` reflects precautionary harm avoidance. These findings motivate the taxonomy of moral styles discussed later in the paper.

Behavioral Consistency and Moral Drift

To assess how consistently each model family reasons across repeated runs of the same dilemma, *behavioral drift* was measured using the standard deviation of key reasoning metrics across ten independent generations per scenario. Figure 3 visualizes this variation across six metrics: entropy, top-2 dominance, number of reasoning steps, density of reasoning, moral term frequency, and systemic safety recall.

`qwq:latest` exhibits the highest behavioral volatility in both *reasoning steps* ($\sigma = 11.18$) and *moral term usage* ($\sigma = 3.36$), suggesting a less predictable generation pattern under identical prompts. `deepseek-r1:32b` and `gemma3_27b` show comparatively stable outputs, especially in systemic recall and entropy.

This drift is particularly relevant in safety-critical applications where reproducibility and moral consistency are vi-

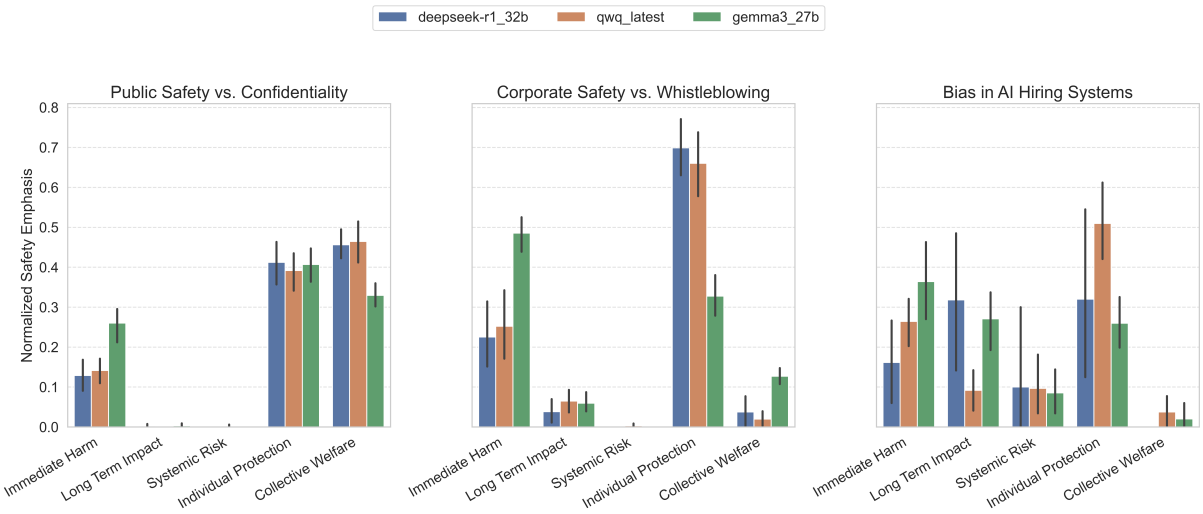

Figure 2: Scenario-wise safety emphasis distributions across model families. Bars show mean weights; error bars represent standard deviation across ten runs.

Scenario	**Safety Dimension**	`deepseek-r1_32B`	`gemma3_27B`	`qwq_latest`
Bias in AI Hiring Systems	Immediate Harm	0.162 ± 0.187	0.364 ± 0.166	0.264 ± 0.099
	Individual Protection	0.320 ± 0.356	0.260 ± 0.116	0.510 ± 0.171
	Collective Welfare	0.000 ± 0.000	0.020 ± 0.063	0.037 ± 0.068
	Long Term Impact	0.318 ± 0.298	0.271 ± 0.125	0.092 ± 0.091
	Systemic Risk	0.100 ± 0.316	0.085 ± 0.092	0.097 ± 0.127
Corporate Safety vs. Whistleblowing	Immediate Harm	0.225 ± 0.140	0.486 ± 0.075	0.252 ± 0.141
	Individual Protection	0.699 ± 0.124	0.328 ± 0.091	0.660 ± 0.133
	Collective Welfare	0.037 ± 0.065	0.127 ± 0.034	0.020 ± 0.029
	Long Term Impact	0.038 ± 0.053	0.060 ± 0.041	0.065 ± 0.048
	Systemic Risk	0.000 ± 0.000	0.000 ± 0.000	0.003 ± 0.009
Public Safety vs. Confidentiality	Immediate Harm	0.129 ± 0.067	0.260 ± 0.073	0.141 ± 0.052
	Individual Protection	0.412 ± 0.092	0.407 ± 0.073	0.392 ± 0.079
	Collective Welfare	0.456 ± 0.063	0.330 ± 0.052	0.465 ± 0.089
	Long Term Impact	0.003 ± 0.008	0.003 ± 0.010	0.000 ± 0.000
	Systemic Risk	0.000 ± 0.000	0.000 ± 0.000	0.002 ± 0.006

Table 1: Scenario-wise safety emphasis: mean $\pm$ standard deviation across 10 runs per model.

tal. From the perspective of RQ2, these results underscore that not all models are equally reliable in maintaining ethical alignment across repeated decisions.

Notably, systemic recall shows modest variance across models, despite being low in absolute magnitude, which may indicate systemic considerations are either underemphasized or stably encoded across model architectures.

Inter-Model Differences in Ethical Reasoning To further evaluate consistent differences in reasoning profiles, metrics were aggregated across all 30 responses per model (3 scenarios × 10 runs). Table 2 summarizes these results.

While entropy and top-2 dominance do not show significant differences ($p = 0.912$ and $p = 0.349$, respectively), large inter-model variation exists in reasoning complexity. Kruskal–Wallis tests reveal significant group-level effects for:

- **Steps:** $H = 32.21, p < 0.001$
- **Density:** $H = 44.27, p < 0.001$
- **Moral Terms:** $H = 17.70, p < 0.001$

Post-hoc Tukey HSD analysis confirms that `qwq:latest` significantly differs from both `deepseek-r1:32b` and `gemma3_27b` on all three dimensions, exhibiting a substantially higher number of reasoning steps and moral term usage. The outputs of `deepseek-r1:32b` are denser than `gemma3_27b`, while `qwq:latest` demonstrates the highest reasoning elaboration overall.

Systemic recall, while qualitatively similar across models, does not reach statistical significance ($p = 0.107$), suggesting that systemic reasoning is either similarly encoded across models or not prominently activated in these dilemmas.

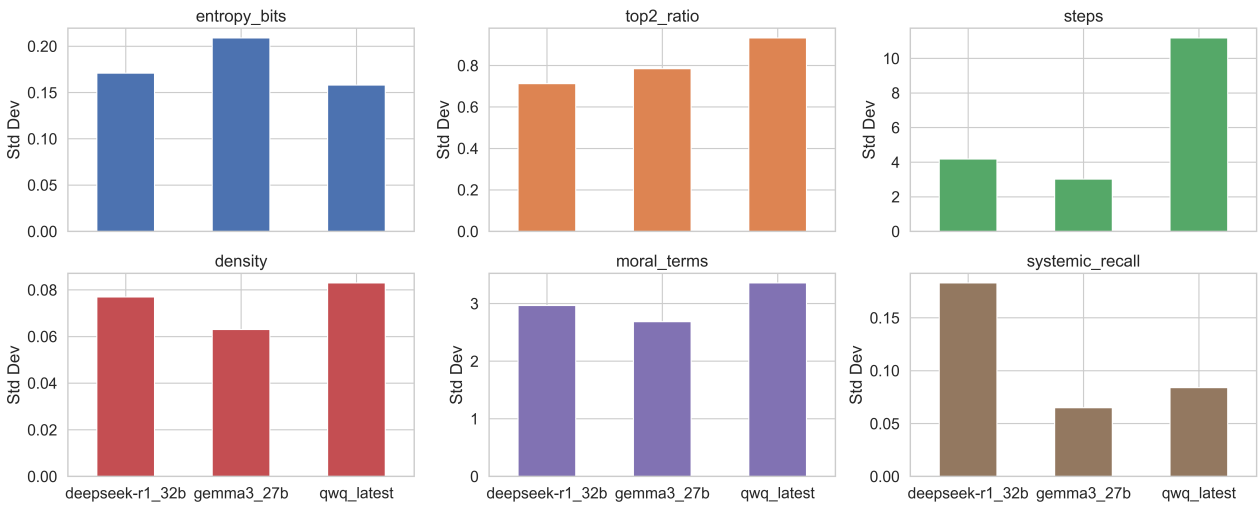

Figure 3: Standard deviation (drift) of moral reasoning metrics per base model. Higher variance indicates less stable reasoning behavior across repeated runs.

Model	Entropy	Top-2	Steps	Density	Moral Terms	Sys Recall	Cosine Sim
`deepseek-r1:32b`	2.831	1.94	10.33	0.317	8.37	0.033	0.894
`gemma3_27b`	2.801	2.10	6.37	0.151	5.77	0.028	0.891
`qwq:latest`	2.811	2.34	18.10	0.267	9.37	0.034	0.926

Table 2: Aggregate behavioral metrics per model (mean across 30 responses).

From the standpoint of RQ2, these findings indicate that models adopt structurally distinct reasoning profiles. `qwq:latest` consistently favors longer, denser reasoning chains, potentially indicating an exploratory or deliberative moral reasoning style. `gemma3_27b`, by contrast, adopts a more compact and possibly heuristic-driven strategy.

Cosine similarity scores (last column in Table 2) further validate consistency within models. All models maintain relatively high intra-model alignment (> 0.89), but `qwq:latest` again shows the most tightly clustered alignment vectors (0.926), despite its internal variability—suggesting a trade-off between content variation and moral directionality.

Overall, these results provide strong evidence for the presence of model-specific ethical reasoning profiles, supporting the development of a principled **taxonomy** of moral reasoning behaviors (RQ3), and offering a foundation for future predictive analysis or alignment optimization.

Emergence of Ethical Reasoning Styles

To evaluate whether consistent styles of moral reasoning emerge across models, this study analyzes the ethical alignment behavior of LLMs in a unified latent space. Each model response is represented by an 11-dimensional moral alignment vector—derived from semantic and dictionary-based scoring—augmented with its systemic safety recall score, yielding a 12-dimensional feature space.

These high-dimensional vectors are standardized and projected into a two-dimensional subspace using Principal Component Analysis (PCA), preserving the major axes of variation across responses. Figure 4 shows the PCA projection colored by base model, with clear separation among model families.

To uncover latent clusters in this space, KMeans clustering is applied. The optimal number of clusters ($k = 3$) is selected based on two complementary approaches:

- The **elbow method**, which evaluates within-cluster variance (inertia) and identifies the point of diminishing returns.
- The **silhouette score**, which measures cohesion and separation of clusters; the highest average silhouette score was observed at $k = 3$.

As shown in Figure 4, `gemma3_27b` tends to occupy a distinct region with less dispersion, reflecting lower reasoning complexity and narrower moral emphasis. In contrast, `qwq_latest` and `deepseek-r1_32b` span broader areas with higher variability in both alignment and safety orientation.

Clustering in Figure 5 reveals three dominant styles of moral reasoning:

- **Balanced Integration**: Characterized by high entropy, evenly distributed moral alignment, and use of multiple safety considerations.

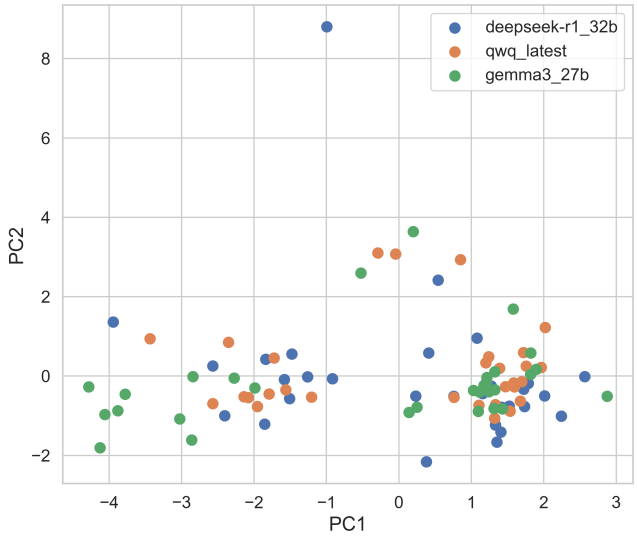

Figure 4: PCA projection of ethical and safety alignment vectors, colored by base model. Spatial separation suggests differing moral alignment tendencies.

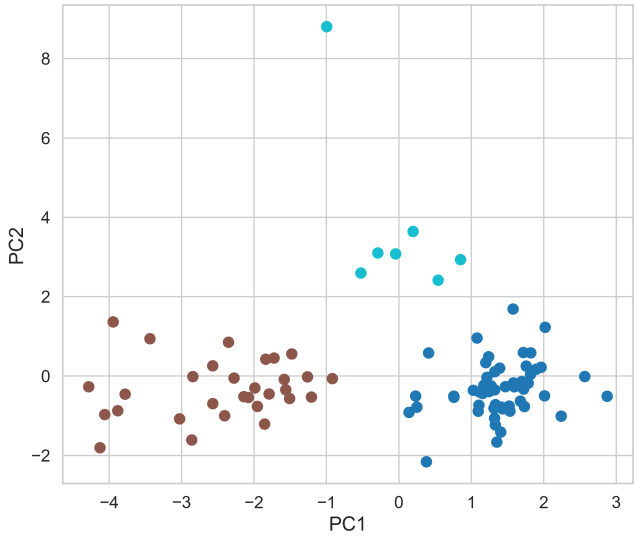

Figure 5: Same PCA space as Figure 4, now colored by empirically derived KMeans clusters ($k = 3$). Emergent reasoning styles are visibly separable.

- **Rule-Based Conservatism**: Marked by dominant reliance on deontological principles and minimal moral diversification.
- **Contextual Pragmatism**: Emphasizes immediate harm and individual protection, often adapting to scenario context rather than abstract principles.

Importantly, these clusters are discovered in an unsupervised manner without human labeling, indicating that model outputs exhibit structurally meaningful patterns. This finding supports RQ3 by showing that it is possible to derive a *taxonomy of reasoning styles* based solely on response features, revealing internal ethical alignment strategies learned by different LLMs.

Robustness via Ablation and Clustering Consistency

To assess the robustness of the discovered moral reasoning taxonomy and ensure it is not an artifact of handcrafted lexicon heuristics, a dictionary weight ablation study was conducted. This analysis systematically varied the combination weight $\omega \in [0.0, 1.0]$ in Equation 1, where $\omega = 0$ removes all dictionary contributions and $\omega = 1$ relies entirely on symbolic cues. The results revealed that clustering consistency and quality—as measured by ARI and silhouette coefficients—plateaued at intermediate weights. Peak alignment between semantic and symbolic structure was observed around $\omega = 0.4$, supporting its selection as the default configuration. This choice reflects a pragmatic trade-off, allowing dictionary-based terms to contribute to interpretability while preserving the contextual nuance captured by semantic similarity.

Figure 6 shows PCA projections of response vectors under both settings, colored by KMeans cluster assignments. While the centroids and boundaries slightly shift, the overall tripartite structure persists across both conditions. This preservation is quantitatively confirmed by an ARI of 0.475, with a bootstrapped 95% confidence interval of [0.346, 0.645] (Figure 7).

Additionally, silhouette analysis revealed a moderate reduction in cluster separability when dictionary features were excluded: the silhouette score dropped from 0.665 (baseline) to 0.400 (ablation). These findings demonstrate that the clustering is not brittle and retains meaningful structure even without handcrafted keyword cues, affirming that the semantic representation carries sufficient ethical signal to support unsupervised taxonomy discovery.

To further validate this finding, the experiment was replicated using HDBSCAN, a density-based clustering algorithm that does not require specifying the number of clusters. Figure 8 compares cluster shapes and memberships under baseline and ablation. The ARI between these HDBSCAN-based assignments is 0.461, again supporting stability. Silhouette scores followed the same trend: 0.562 for baseline, 0.319 under ablation.

These results strengthen the conclusion that moral reasoning styles are not model artifacts of token-level feature engineering, but instead reflect meaningful, semantically-driven variation in model behavior. This adds further support to RQ3, which seeks to derive a reliable taxonomy of LLM moral reasoning.

Discussion and Conclusion

This study set out to examine how LLMs handle ethical dilemmas that involve conflicting safety principles. By evaluating responses across three systematically constructed scenarios and employing a structured moral profil-

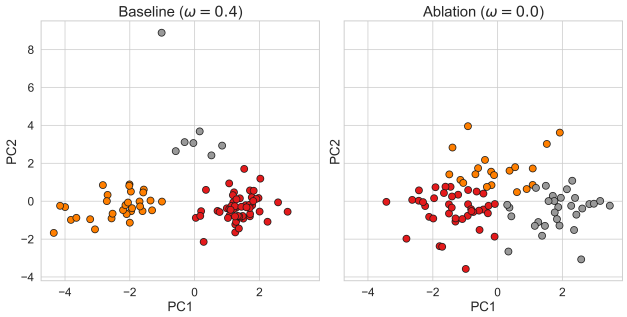

Figure 6: PCA projections of moral alignment space under baseline ($\omega = 0.4$) and ablation ($\omega = 0.0$) conditions, colored by KMeans clusters ($k = 3$). Cluster structure is largely preserved under ablation, suggesting semantic features alone are sufficient for taxonomy emergence.

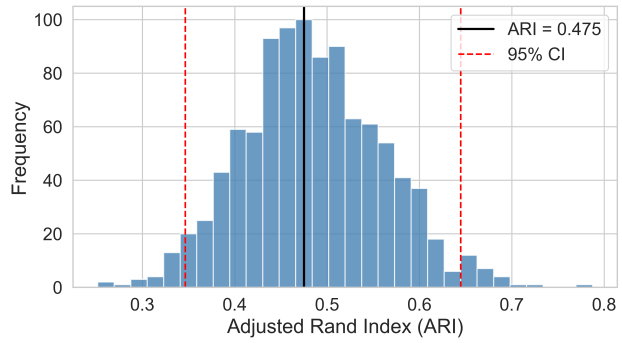

Figure 7: Bootstrapped Adjusted Rand Index (ARI) between baseline and ablation cluster assignments ($N = 1000$). The mean ARI is 0.475 with 95% confidence interval [0.346, 0.645], confirming moderate consistency.

ing pipeline, it became evident that LLMs do not respond arbitrarily. Even under consistent prompt conditions, distinct and reproducible reasoning patterns emerge that reflect deeper alignment characteristics.

The analysis revealed clear divergences across model families. `deepseek-r1_32b` consistently prioritized rule-based reasoning and individual protection, often producing structured, concise responses with high internal coherence. In contrast, `gemma3_27b` exhibited a more balanced moral framing, engaging multiple ethical principles and safety considerations simultaneously. `qwq_latest` demonstrated the highest contextual variability, frequently emphasizing immediate harm and care ethics while producing the most complex and verbose answers.

These behavioral differences were substantiated through multiple generations per model per scenario. Statistical analysis confirmed significant inter-model variation in reasoning complexity, moral emphasis, and ethical diversity, especially in the number of reasoning steps, density, and moral term usage. Such differences are not attributable to sampling noise but instead reflect model-specific alignment tendencies.

Dimensionality reduction and clustering analysis further

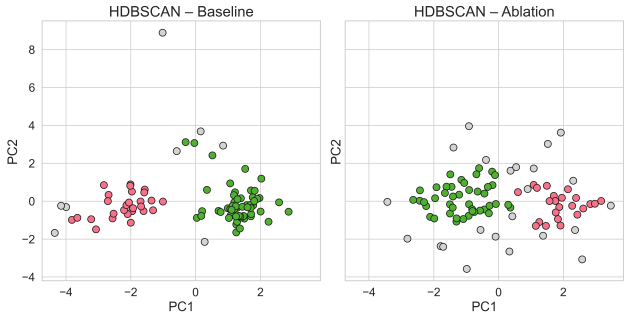

Figure 8: HDBSCAN clustering results on PCA space for baseline and ablation conditions. Despite algorithmic differences and removal of dictionary features, cluster formation remains broadly consistent.

identified three consistent ethical reasoning styles: *rule-based conservatism*, *contextual pragmatism*, and *balanced integration*. These styles emerged in an unsupervised manner, without prior labeling, suggesting that LLMs encode latent normative preferences shaped during training. These findings align with broader concerns in AI ethics that model behaviors are shaped not only by data and prompts but by internal representations of moral reasoning.

This work has broader implications for LLM safety, governance, and deployment. As LLMs are integrated into decision-support systems in healthcare, law, and public administration, the choice of model becomes an ethical decision in itself. A model that systematically downplays long-term impact or collective welfare may be unsuitable for high-stakes public use, regardless of benchmark accuracy.

The hybrid profiling approach used in this study—blending semantic similarity, dictionary-based scoring, and reasoning structure—demonstrates a transparent, reproducible way to audit model behavior. This framework supports a more fine-grained understanding of alignment that extends beyond binary safety flags or aggregate utility scores.

Naturally, there are limitations. The scenarios used, though carefully designed, cannot capture the full range of real-world ethical complexity. The moral lexicon may omit culturally specific values or underrepresent non-Western ethical traditions. Additionally, while clustering reveals distinct reasoning styles, it does not provide normative judgments on which styles are preferable or desirable for particular applications.

Despite these constraints, the study offers a principled step toward characterizing how LLMs interpret and navigate moral dilemmas. The proposed methodology provides a generalizable framework for auditing and comparing models' ethical behavior. Future research could expand this approach to multilingual models, fine-tuned variants, or real-time decision-making environments, offering deeper insight into the moral landscape of generative AI systems.

Acknowledgments

The author acknowledges support through the project "CoCi: Co-Evolving City Life", which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 833168.

References

Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7: 149–155.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Araque, O.; Gatti, L.; and Kalimeri, K. 2020. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191: 105184.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82–115.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bryson, J. J. 2019. The past decade and future of AI's impact on society. In *Towards a new enlightenment? A transcendent decade*. Turner.

Calo, R. 2017. Artificial intelligence policy: a primer and roadmap. *UCDL Rev.*, 51: 399.

Floridi, L. 2013. *The Ethics of Information*. Oxford University Press.

Floridi, L.; and Sanders, J. W. 2004. On the morality of artificial agents. *Minds and machines*, 14: 349–379.

Gabriel, I.; and Ghazavi, V. 2021. The challenge of value alignment: From fairer algorithms to AI safety. *arXiv preprint arXiv:2101.06060*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.

Helbing, D.; Beschorner, T.; Frey, B.; Diekmann, A.; Hagendorff, T.; Seele, P.; Spiekermann-Hoff, S.; van den Hoven, J.; and Zwitter, A. 2021. Triage 4.0: On Death Algorithms and Technological Selection. Is Today's Data-Driven Medical System Still Compatible with the Constitution? *Journal of European CME*, 10(1): 1989243.

Hendrycks, D. 2024. *Introduction to AI safety, ethics and society*. Dan Hendrycks.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417.

Ji, J.; Chen, Y.; Jin, M.; Xu, W.; Hua, W.; and Zhang, Y. 2024. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J. T.; Levine, S.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Hessel, J.; et al. 2025. Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*, 1–16.

Mahajan, S. 2024. The Executioner Paradox: understanding self-referential dilemma in computational systems. *AI & SOCIETY*, 1–8.

Mahajan, S. 2025. The democratization dilemma: When everyone is an expert, who do we trust? *Humanities and Social Sciences Communications*, 12(1): 1–5.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4): 18–21.

Morley, J.; Machado, C. C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; and Floridi, L. 2020. The ethics of AI in health care: a mapping review. *Social Science & Medicine*, 260: 113172.

Mündler, N.; He, J.; Jenko, S.; and Vechev, M. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015.

Rehbein, I.; Brauner, L.; Ertz, F.; Reinig, I.; and Ponzetto, S. P. 2025. Moral reckoning: How reliable are dictionary-based methods for examining morality in text? In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, 232–250.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ross, J.; Kim, Y.; and Lo, A. W. 2024. Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint arXiv:2408.02784*.

Russell, S. J.; and Norvig, P. 2016. *Artificial intelligence: a modern approach*. Pearson.

Vijayaraghavan, A.; and Badea, C. 2024. Minimum levels of interpretability for artificial moral agents. *AI and Ethics*, 1–17.

Zhang, J.; Ji, X.; Zhao, Z.; Hei, X.; and Choo, K.-K. R. 2023. Ethical Considerations and Policy Implications for Large Language Models: Guiding Responsible Development and Deployment. arXiv:2308.02678.

Zhang, X. 2025. Constitution or Collapse? Exploring Constitutional AI with Llama 3-8B. *arXiv preprint arXiv:2504.04918*.

Zhou, J.; Hu, M.; Li, J.; Zhang, X.; Wu, X.; King, I.; and Meng, H. 2023. Rethinking Machine Ethics–Can LLMs Perform Moral Reasoning through the Lens of Moral Theories? *arXiv preprint arXiv:2308.15399*.