Position: Evaluations of AI Moral Reasoning Still Miss Half of the Picture

Anonymous ACL submission

Abstract

Recent work on evaluating the moral competence of large language models (LLMs) has focused primarily on what we christen the moral value problem, i.e., do model outputs align with human moral values. In contrast, the moral norm problem, i.e. whether models can identify and correctly apply context-sensitive moral norms, remains underexplored. We posit that this imbalance stems partly due to the field's reliance on descriptive ethics frameworks, such as Moral Foundations Theory and Kohlberg's stages of moral development, which emphasize value representation over normative application. We review existing benchmarks and evaluation methods, and show that they cluster heavily around the value problem, while discussion regarding normative ethics remains underrepresented. We identify three crucial gaps : (i) the absence of high-quality ground-truth data for moral norms and their applications, (ii) insufficient evaluation of intermediate reasoning processes, and (iii) limited attention to the identification of morally relevant features in context. Subsequently, we propose a research agenda that includes the development of standardized formal representations for normative theories, the construction of expert-annotated datasets capturing norm application, and evaluation protocols that explicitly distinguish between values-level and norms-level competence. Our goal is to encourage more systematic study of normative reasoning in LLMs.

1 Introduction

Users increasingly rely on large language models (LLMs) for moral advice. Recent evidence suggests that such systems are perceived as comparable to expert ethicists in moral expertise (Dillion et al., 2025). Regardless of whether this trust is warranted, its prevalence makes it important to characterize what current evaluations of AI moral reasoning measure, and what they omit.

We frame the evaluation of moral reasoning in LLMs as consisting of two related but distinct problems. The first is the *moral value problem*", which asks whether model outputs reflect human moral values, understood as broad preferences and priorities. The second is the "*moral norm problem*, which asks whether models can identify and correctly apply moral principles that determine how those values translate into judgments in specific contexts. While the value problem concerns alignment with observed human attitudes, the norm problem concerns the application of structured principles drawn from normative ethics.

Prior work has focused largely on the moral value problem. Empirical studies have compared LLM outputs with human responses using instruments such as the Moral Machine experiment, moral foundations questionnaires, and large-scale value surveys. These approaches align with descriptive ethics, which studies patterns in human moral beliefs and preferences. As a result, existing benchmarks primarily assess whether models reproduce distributions of human values.

In contrast, the moral norm problem has received limited attention. Addressing this problem requires engagement with normative ethics, which studies which principles are correct and how they apply in particular cases. Values alone are insufficient to determine moral judgments; norms specify how values constrain decisions in context. A model may approximate human value distributions while failing to construct valid arguments within established ethical frameworks, recognize when specific principles apply, or identify morally relevant features of novel scenarios. One reason for this gap is that normative ethics has not been systematically represented in forms amenable to computational evaluation. However, the underlying theories and principles are well-documented; the primary challenge lies in organizing them into structured representations that support benchmarking.

In this paper, we review existing approaches to evaluating moral competence in LLMs and map them onto realistic components of moral reasoning. We show that current evaluations concentrate on descriptive ethics and values-level alignment, with limited coverage of norms-level reasoning. Based on this analysis, we outline directions for developing datasets, representations, and evaluation protocols that enable systematic assessment of normative moral reasoning in AI systems.

2 Background

The empirical study of human moral values has produced well-established frameworks. Moral Foundations Theory (MFT) identifies a set of foundational moral concerns [1] that structure moral intuitions across cultures (Graham et al., 2013). Schwartz's Theory of Basic Human Values provides a complementary framework organized around dimensions such as self-transcendence, conservation, openness to change, and self-enhancement (Schwartz, 2012). Both offer validated instruments for measuring what people value, and both have been widely adopted in the AI alignment literature as a basis for assessing model behavior.

Research on moral decision-making has also produced large-scale datasets of human judgments. The Moral Machine experiment collected responses to autonomous vehicle dilemmas at scale and identified consistent patterns in how participants trade off outcomes (Awad et al., 2018). In parallel, research in moral psychology finds that human judgments draw on both outcome-based reasoning and rule-based responses, often associated with consequentialist and deontological patterns (Cushman, 2013). These results provide a basis for comparing model outputs with human decisions across controlled scenarios.

Recent work addresses alignment in settings where moral views differ across individuals or groups. Approaches to pluralistic alignment draw on social choice theory to formalize how conflicting preferences can be aggregated or represented (Sorensen et al., 2024). New benchmarks evaluate whether LLMs capture the distribution of moral opinions observed in human populations, rather than converging to a single response (Russo et al., 2025; Poole-Dayan et al., 2026). Other studies examine consistency across related moral judgments (Moore et al., 2024) and the effects of temporal variation in human feedback on alignment outcomes (Keswani et al., 2025).

Within computational ethics (Tolmeijer et al., 2021), this line of work focuses on representing and evaluating descriptive ethics. In contrast, comparatively little work has addressed how to represent and evaluate normative ethics in computational settings. We also provide a detailed description of the state of the machine ethics evaluation in the Appendix.

3 Conflating Values with Norms in Machine Ethics Evaluations

Recent work has begun to examine how LLMs are evaluated for moral competence. Snoswell et al. (2026) find that while a subset of papers assesses model-generated justifications, these evaluations typically rely on surface-level checks (e.g., consistency, hallucination) or subjective ratings, and do not test whether reasoning supports final decisions. They argue for decomposing moral reasoning into intermediate steps and evaluating performance against expert standards, as well as for incorporating a broader range of normative theories beyond commonly used descriptive frameworks.

We build on this observation by distinguishing between two components of moral reasoning: the moral value problem and the moral norm problem. The value problem asks whether models reflect human moral values, while the norm problem asks whether models can apply moral principles to specific situations.

Existing work has focused primarily on the value problem. Benchmarks commonly use questionnaires, dilemma tasks, and value surveys to compare model outputs with human judgments (Nunes et al., 2024; Liu et al., 2026). These approaches align with descriptive ethics and provide tractable methods for evaluating value alignment. However, they do not assess whether models can reason with normative principles.

In contrast, the norm problem remains underexplored. A small number of benchmarks attempt to evaluate principle-based reasoning (Samway et al., 2025; Rao et al., 2023; Zhou et al., 2024), but they rely on ad hoc representations of normative theories and lack shared datasets. This limits comparability across studies and prevents cumulative progress.

A key issue is the conflation of values with norms. Many benchmarks use frameworks such

[1] The main dimensions for MFT include care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and liberty/oppression

as Moral Foundations Theory to evaluate "moral reasoning". While these frameworks capture what people tend to value, they do not specify how those values should be applied in context. As a result, evaluations that rely solely on values-level instruments measure alignment with human preferences rather than the ability to apply moral principles.

From a measurement perspective, this conflation creates a construct validity problem. Benchmarks that claim to assess moral reasoning often operationalize only one component of the construct. A model may match human value distributions while lacking the ability to identify relevant features, apply appropriate norms, or construct valid arguments. Addressing this gap requires evaluation methods that distinguish between values-level alignment and norms-level reasoning. We explicitly refer to these gaps in the ensuing section.

4 Gaps in Current Approaches

4.1 Missing Ground Truth for Moral Norms

A central limitation is the lack of broadly applicable ground-truth data for normative ethics, i.e., representations of the norms endorsed by different moral theories. In practice, this forces each benchmark to construct its own dataset of moral norms, limiting comparability across studies.

In contrast, descriptive ethics benefits from established datasets and measurement tools. No equivalent infrastructure exists for normative ethics: current resources do not systematically encode which principles a theory endorses, how they apply across contexts, or what constitutes correct application.

Some work provides initial steps in this direction. Hammerton (2025) formalize moral theories in terms of prioritized abstract properties, and Tennant et al. (2025) model theories as reward functions in an iterated prisoner's dilemma. However, these efforts remain isolated and do not support standardized evaluation. A more systematic approach would involve constructing shared datasets that map normative theories to principles and reasoning patterns, with expert input and sufficient coverage for benchmarking.

The absence of shared ground truth also affects reliability. Differences in benchmark design make results difficult to compare, and temporal variation in human judgments (Keswani et al., 2025) introduces additional instability. Shared representations would not fully resolve these issues but would provide a common basis for evaluation.

4.2 Evaluating Reasoning Traces Without Normative Vocabulary

A second limitation concerns the evaluation of reasoning traces. Even when benchmarks assess intermediate reasoning, they often lack the normative vocabulary needed to determine whether the norms invoked are appropriate, correctly applied, or properly weighted. MoReBench (Chiu et al., 2025) illustrates this issue.

MoReBench evaluates reasoning using scenario-specific rubric items that combine theory-related and outcome-based criteria. Without a clear representation of how normative theories should be applied, it is difficult to distinguish between failures of norm understanding and failures of application.

The benchmark also uses a criterion-fulfillment scoring approach with length normalization to reduce verbosity bias. However, this creates incentives for minimal responses that satisfy rubric requirements without exposing reasoning, allowing less transparent models to achieve higher scores.

More generally, there is a disconnect between mentioning a morally relevant consideration and incorporating it into reasoning. Models may be penalized for implicit reasoning or rewarded for listing considerations without integrating them. While some work addresses aspects of this problem (Rao et al., 2023), current approaches remain limited. This is a broader challenge for evaluating reasoning processes, but it is particularly consequential in moral reasoning, where flawed metrics may misrepresent model capabilities.

4.3 The Feature Identification Problem

A third limitation concerns the identification of morally relevant features. Before applying any norm, a system must determine which aspects of a situation are relevant for moral evaluation. Current approaches do not provide a general method for this step.

Existing work typically treats feature identification as task-specific. For example, Kwon et al. (2024) generate features by prompting models to extract salient information across variations of a scenario. While effective in controlled settings, this approach does not generalize to novel situations. As noted in prior work, there is no unified computational account of how humans identify morally relevant features.

As a result, evaluations are constrained to the features anticipated by benchmark designers. This

limits the ability to assess performance in settings where relevant considerations differ from those encoded in the dataset. The feature identification problem is closely related to the representation of normative theories, since many theories specify which aspects of a situation should be treated as morally relevant. Improved representations of normative principles would therefore support more general approaches to feature identification.

We present the coverage across representative benchmarks for machine ethics evaluation of norms and values in Figure 1, and the corresponding limitations.

5 Ways Forward

We outline several directions for improving the evaluation of moral reasoning in LLMs.

Shared representations of normative theories. The field would benefit from common formal vocabularies that specify what different normative theories prescribe, including their principles, rules, and characteristic reasoning patterns. Prior work (Hammerton, 2025; Tennant et al., 2025) provides initial steps, but existing efforts remain fragmented. Developing shared representations would enable the construction of standardized datasets linking theories to their endorsed norms, and would support comparability and aggregation of results across studies.

Expert-informed ground-truth data. Datasets for normative evaluation should incorporate input from domain experts across multiple ethical traditions. Such datasets should be sufficiently large and structured to capture different levels of competence, including recognizing relevant norms, applying them in straightforward cases, and resolving conflicts between competing principles.

Separation of values-level and norms-level evaluation. Evaluations should explicitly distinguish between assessing value alignment (the moral value problem) and assessing norm application (the moral norm problem). This distinction should be reflected in both task design and reporting, rather than treated as a single construct.

Separation of deliberation and decision. The quality of a model's reasoning process and the correctness of its final judgment should be evaluated independently. A correct answer without appropriate reasoning does not demonstrate normative competence, and conversely, sound application of principles may yield non-standard conclusions. Conflating these dimensions obscures model capabilities.

Evaluation under assisted and unassisted settings. Benchmarks should assess both baseline performance (e.g., zero-shot responses) and performance under structured conditions, such as guided prompting, tool-usage, or multi-step deliberation. This allows evaluation of both observed and potential normative competence, allowing a clearer distinction between a model's inability to apply norms and a failure to elicit them.

Improved evaluation of reasoning traces. Current methods for assessing reasoning traces are limited. Future work should incorporate techniques from the chain-of-thought faithfulness literature to evaluate whether invoked norms contribute to final decisions. In addition to criterion-based scoring, evaluation methods should distinguish between superficial mention of norms and their integration into reasoning.

6 Conclusion

The evaluation of moral competence in AI systems has made substantial progress on the moral value problem, i.e., whether models reflect human moral priorities. However, this focus has left the moral norm problem underexplored. Existing approaches, grounded in descriptive ethics, capture what models appear to value but do not assess whether they can apply normative principles to specific cases.

Addressing this limitation requires new evaluation infrastructure. In particular, the field needs shared representations of normative theories, expert-informed datasets that specify how norms apply across contexts, and evaluation protocols that distinguish between values-level alignment and norms-level reasoning. While these challenges are non-trivial, they are necessary for a complete assessment of moral competence. Evaluating what models care about is insufficient; it is equally important to evaluate whether they can determine what those commitments entail in practice.

References

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François

Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature*, 563(7729):59–64.

Yu Ying Chiu, Michael S. Lee, Rachel Calcott, Brandon Handoko, Paul de Font-Reaulx, Paula Rodriguez, Chen Bo Calvin Zhang, Ziwen Han, Udari Madhushani Sehwag, Yash Maurya, Christina Q. Knight, Harry R. Lloyd, Florence Bacus, Mantas Mazeika, Bing Liu, Yejin Choi, Mitchell L. Gordon, and Sydney Levine. 2025. MoReBench: Evaluating Procedural and Pluralistic Moral Reasoning in Language Models, More than Outcomes. *arXiv preprint*. ArXiv:2510.16380 [cs].

Fiery Cushman. 2013. Action, Outcome, and Value: A Dual-System Framework for Morality. *Personality and Social Psychology Review*, 17(3):273–292.

Danica Dillion, Debanjan Mondal, Niket Tandon, and Kurt Gray. 2025. AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15(1):4084.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral Foundations Theory. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier.

Matthew Hammerton. 2025. The fundamental divisions in ethics. *Inquiry*, 68(2):318–341.

Junfeng Jiao, Saleh Afroogh, Abhejay Murali, Kevin Chen, David Atkinson, and Amit Dhurandhar. 2025. LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. *Scientific Reports*, 15(1):34642.

Vijay Keswani, Cyrus Cousins, Breanna Nguyen, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, and Walter Sinnott-Armstrong. 2025. Moral Change or Noise? On Problems of Aligning AI With Temporally Unstable Human Feedback. *arXiv preprint*. ArXiv:2511.10032 [cs].

Joseph Kwon, Josh Tenenbaum, and Sydney Levine. 2024. Neuro-symbolic models of human moral judgment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Andy Liu, Kshitish Ghate, Mona Diab, Daniel Fried, Atoosa Kasirzadeh, and Max Kleiman-Weiner. 2026. Generative Value Conflicts Reveal LLM Priorities. *arXiv preprint*. ArXiv:2509.25369 [cs].

Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are Large Language Models Consistent over Value-laden Questions? *arXiv preprint*. ArXiv:2407.02996 [cs].

José Luiz Nunes, Guilherme F. C. F. Almeida, Marcelo De Araujo, and Simone D. J. Barbosa. 2024. Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:1074–1087.

Elinor Poole-Dayan, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel A. Bakker. 2026. Benchmarking Overton Pluralism in LLMs. *arXiv preprint*. ArXiv:2512.01351 [cs].

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. *arXiv preprint*. ArXiv:2310.07251 [cs].

Giuseppe Russo, Debora Nozza, Paul Röttger, and Dirk Hovy. 2025. The Pluralistic Moral Gap: Understanding Judgment and Value Differences between Humans and Large Language Models. *arXiv preprint*. ArXiv:2507.17216 [cs].

Keenan Samway, Max Kleiman-Weiner, David Guzman Piedrahita, Rada Mihalcea, Bernhard Schölkopf, and Zhijing Jin. 2025. Are Language Models Consequentialist or Deontological Moral Reasoners? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30699–30726, Suzhou, China. Association for Computational Linguistics.

Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1).

Aaron J Snoswell, Daniel Kilov, and Seth Lazar. 2026. Beyond Verdicts: Evaluating Language Model Moral Competence.

Shelly Soffer, Dafna Nesselroth, Keren Pragier, Roi Anteby, Donald Apakama, Emma Holmes, Ashwin Shreekant Sawant, Ethan Abbott, Lauren Alyse Lepow, Ishita Vasudev, Joshua Lampert, Moran Gendler, Nir Horesh, Orly Efros, Benjamin S Glicksberg, Robert Freeman, David L Reich, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Disagreements in Medical Ethics Question Answering Between Large Language Models and Physicians.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment. *Preprint*, arXiv:2402.05070.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2025. Moral Alignment for LLM Agents. *arXiv preprint*. ArXiv:2410.01639 [cs].

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2021. Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, 53(6):1–38.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2024. Rethinking Machine Ethics – Can LLMs Perform Moral Reasoning through the Lens of Moral Theories? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2227–2242, Mexico

City, Mexico. Association for Computational Linguistics.

A The State of Machine Ethics Evaluations

Snoswell et al. (2026) provide a recent survey of evaluation methods for machine ethics in *Beyond Verdicts: Evaluating Language Model Moral Competence.* They report that 39% of surveyed papers attempt to assess the quality of model-generated justifications. However, these evaluations typically rely on coarse checks for logical consistency or hallucination, or on subjective human ratings. The authors note that none of the surveyed work examines whether model justifications are causally linked to their final decisions, and suggest connecting moral evaluation with the literature on chain-of-thought faithfulness.

Snoswell et al. (2026) propose decomposing moral reasoning into intermediate steps and evaluating model performance at each stage against expert judgments. They also recommend expanding beyond commonly used descriptive frameworks, such as Moral Foundations Theory (Graham et al., 2013), to include a broader set of normative theories. In their formulation, evaluation should test whether a model can identify morally relevant features, relate them to appropriate reasons, and derive a defensible conclusion.

We build on this perspective. Our central claim is that existing evaluation methods emphasize whether models reproduce human moral values, while giving comparatively little attention to whether models can apply norms in a structured and context-sensitive manner.

A.1 The Moral Value Problem: What Does the AI Care About?

A common approach to evaluating moral competence in LLMs is to test whether model outputs reflect human moral values. This is typically operationalized through multiple-choice value questionnaires, dilemma-based tasks (e.g., trolley problems), or domain-specific scenarios such as those in medical ethics (Soffer et al., 2024). For example, Nunes et al. (2024) administer both the Moral Foundations Questionnaire and the Moral Foundations Vignettes to LLMs. They find that models exhibit internal consistency within each instrument but produce conflicting responses when abstract value endorsements are compared with judgments about concrete violations. Other work shows that generative settings can still reveal model priorities in cases where values conflict, by analyzing responses to value trade-off scenarios (Liu et al., 2026).

This line of work aligns with a broader effort to characterize model behavior using tools from psychology. Evaluating values-level alignment is relatively straightforward: researchers adapt an existing instrument, apply it to both human participants and models, and compare the resulting distributions. Open questions remain about dataset selection, aggregation across populations, and cross-cultural coverage, but these are methodological challenges within an established paradigm. In the terminology of computational ethics (Tolmeijer et al., 2021), this corresponds to formalizing descriptive ethics and evaluating machine behavior against it.

A.2 The Moral Norm Problem: Can the AI Apply Moral Principles?

The moral norm problem concerns whether LLMs can identify and apply the principles that determine how values should guide decisions in specific contexts. Within computational ethics, this requires formalizing normative ethics and designing evaluations that test principle application rather than value representation.

Only a limited number of benchmarks address this problem directly. MoralLens (Samway et al., 2025), for example, evaluates whether model reasoning aligns with a taxonomy of 16 rationale types grounded in consequentialist and deontological theory. Rao et al. (2023) introduce a policy-based framework in which sets of theory-linked rules are used to guide and assess in-context ethical reasoning. Related work examines whether models can apply established moral theories to novel scenarios (Zhou et al., 2024).

Across these approaches, a common limitation is the lack of shared datasets that map normative theories to general principles or fine-grained rules. As a result, each benchmark constructs its own set of theory-derived norms. This limits comparability across studies and prevents cumulative progress: new benchmarks do not build on prior resources, and results cannot be evaluated against a common standard.

What current AI morality evaluations miss

Coverage across representative benchmarks, organized by the dimensions identified in §3–4

	CONSTRUCT COVERAGE		EVALUATION INFRASTRUCTURE (§4)		
	Moral values assessed	Normative theories engaged	Shared norm vocabulary	Reasoning trace evaluated	Moral salience identification
DESCRIPTIVE ETHICS / VALUES FRAMEWORKS					
MFT Moral Hypocrisy Nunes et al. (2024) MFQ + MFV questionnaires	Y	N	N	N	N
Moral Machine Awad et al. (2018) Dilemma judgments	Y	N	N	N	N
LLM Ethics Benchmark Jiao et al. (2025) MFT + Kohlberg stages	Y	N	N	N	N
NORMATIVE ETHICS / REASONING-FOCUSED					
MoralLens Samway et al. (2025) 16-rationale taxonomy	~	Y	N	Y	~
MoReBench Chiu et al. (2025) Criterion-fulfillment rubric	~	~	N	Y	~
Policy-based deliberation Rao et al. (2023) Theory-specific policies	N	~	~	~	~
Theory-lens reasoning Zhou et al. (2024) Direct theory application	N	~	N	~	N

Y Addressed ~ Partial / ad hoc N Not addressed

No listed benchmark evaluates scaffolded or best-case moral reasoning (§4.4).

Figure 1: Coverage across representative benchmarks, organized by the dimensions identified in §3–4

A.3 The Values-Norms Conflation

A recurring issue in the literature is the use of values-based frameworks as proxies for normative competence. For instance, several benchmarks evaluate moral reasoning using Moral Foundations Theory (MFT). While MFT provides a structured account of moral intuitions, it is a descriptive framework: it captures what people tend to value, not how those values should be applied in specific cases. Treating MFT as sufficient for evaluating reasoning conflates value alignment with norm application.

This pattern is partly explained by differences in available tools. MFT and related frameworks provide validated instruments and well-defined categories that are readily adapted for computational evaluation. Normative ethics, by contrast, consists of multiple competing theories, such as consequentialism, deontology, virtue ethics, care ethics, and contractualism, without standardized representations or measurement instruments.

However, this difference in tractability does not resolve the underlying issue. Values alone do not determine judgments; norms specify how values constrain decisions in context. A model may match human value distributions while failing to construct valid arguments within established ethical frameworks, recognize when specific principles apply, or identify relevant features of a scenario. From a measurement perspective, benchmarks that claim to evaluate "moral reasoning" using only values-level instruments risk lacking construct validity.

This issue is illustrated by the "LLM Ethics Benchmark" (Jiao et al., 2025), which evaluates moral reasoning across dimensions such as foundational principles, robustness, and value consistency. The benchmark defines moral reasoning in terms that include identifying dilemmas, weighing considerations, and applying principles to reach justified conclusions. However, its implementation relies on Moral Foundations Theory to represent both values and principles. Since MFT does not specify how principles should be applied, this setup evaluates value alignment rather than normative

reasoning. Similar patterns appear across multiple studies in the literature.