Discerning What Matters: A Multi-Dimensional Assessment of Moral Competence in LLMs

Daniel Kilov
Machine Intelligence and Normative Theory Lab
Australian National University
Acton, ACT 2601, Australia
daniel.kilov@anu.edu.au

Caroline Hendy
Machine Intelligence and Normative Theory Lab
Australian National University
Acton, ACT 2601, Australia
Caroline.Hendy@anu.edu.au

Secil Yanik Guyot
Machine Intelligence and Normative Theory Lab
Australian National University
Acton, ACT 2601, Australia
Secil.YanikGuyot@anu.edu.au

Aaron J. Snoswell
Digital Media Research Centre GenAI Lab
Queensland University of Technology
Kelvin Grove, QLD 4012, Australia
a.snoswell@qut.edu.au

Seth Lazar
Machine Intelligence and Normative Theory Lab
Australian National University
Acton, ACT 2601, Australia
seth.lazar@anu.edu.au

Abstract

Moral competence is the ability to act in accordance with moral principles. As large language models (LLMs) are increasingly deployed in situations demanding moral competence, there is increasing interest in evaluating this ability empirically. We review existing literature and identify three significant shortcoming: (i) Over-reliance on prepackaged moral scenarios with explicitly highlighted moral features; (ii) Focus on verdict prediction rather than moral reasoning; and (iii) Inadequate testing of models' (in)ability to recognize when additional information is needed. Grounded in philosophical research on moral skills, we then introduce a novel methodology for assessing moral competence in LLMs that addresses these shortcomings. Our approach moves beyond simple verdict comparisons to evaluate five distinct dimensions of moral competence: identifying morally relevant features, weighting their importance, assigning moral reasons to these features, synthesizing coherent moral judgments, and recognizing information gaps. We conduct two experiments comparing six leading LLMs against both non-expert humans and professional philosophers. In our first experiment using ethical vignettes standard to existing work, LLMs generally outperformed non-expert humans across multiple dimensions of moral reasoning. However, our second experiment, featuring novel scenarios specifically designed to test moral sensitivity by embedding relevant

Article presented at the Second IASEAI Conference (IASEAI'26).

features among irrelevant details, revealed a striking reversal: several LLMs performed significantly worse than humans at identifying morally salient features. Our findings suggest that current evaluations may substantially overestimate LLMs' moral reasoning capabilities by eliminating the crucial task of discerning moral relevance from noisy information, which we take to be a prerequisite for genuine moral skill. This work provides a more nuanced framework for assessing AI moral competence and highlights important directions for improving (assessment of) ethical reasoning capabilities in advanced AI systems[1].

1 Introduction

1.1 Background & motivation

As autonomous AI systems become more capable, ensuring they can identify moral considerations and moderate their actions accordingly becomes increasingly important (15, 21). It has been proposed that Artificial Moral Advisors (AMAs) could provide invaluable guidance to individuals facing ethical dilemmas, especially in time-pressured situations (12). These and many other applications require AI systems that can engage in ethical reasoning *in situ*, rather than relying solely on pre-programmed moral rules that cannot anticipate all possible ethical considerations. Large Language Models (LLMs) have demonstrated impressive capabilities in understanding moral concepts and applying them to scenarios (17, 5). However, current research suggests a disconnect between a model's capability to discuss moral concepts and its ability to apply these concepts consistently in decision-making (16, 20). This discrepancy may represent a **capability overhang** — latent abilities that could be elicited with appropriate training and/or prompting techniques (8). Understanding precisely where LLMs excel and where they falter in moral reasoning is therefore crucial for developing systems that can both advise humans ethically and act ethically themselves. However, existing research suffers from at least three serious limitations:

I. Over-reliance on pre-packaged cases: The overwhelming majority of ethical evaluations rely on static datasets with neatly prepackaged moral scenarios (14, 19, 22), often derived from classic ethical thought experiments (18, 4) or from psychological instruments (26, 35). These scenarios typically present clear moral dilemmas with explicitly highlighted morally relevant features. While useful for standardization, this approach fails to assess a model's ability to identify morally relevant features from noisy, unfiltered information — a capacity we call **moral sensitivity**. But, as we show below, pre-packaging a scenario means doing much of the work for the model, and then only evaluating the model's performance on what remains. In real-world ethical decision-making, both human moral advisors and autonomous agents must first identify which features of a situation carry moral significance before rendering judgment.

II. Focus on verdicts over reasoning: Current evaluations of LLM ethical capabilities fundamentally conflate two distinct tasks: predicting human moral judgments versus engaging in sound moral reasoning. For instance, many LLM ethical evaluations focus on judgment prediction with human situational ethics data serving as the gold standard (3, 6) such as the 'Am I The Asshole' subreddit (32, 33). But the prediction vs. reasoning distinction is crucial. Consider mathematical reasoning as an analogy: a system that perfectly predicts how humans answer mathematical questions would not demonstrate mathematical competence. Such a system might excel at reproducing common human errors—like believing division always makes numbers smaller or that multiplication always increases quantities—without understanding mathematical principles. Similarly, systems like DELPHI (17) effectively model common ethical intuitions but fail to demonstrate the deeper competence required for ethical reasoning. 1 empirically demonstrated this limitation, showing that language models can achieve "superhuman" performance on ethics datasets while making fundamental moral errors humans rarely make. More troublingly, these researchers could flip models' ethical verdicts through simple rewording that preserved moral content—revealing reliance on linguistic patterns rather than moral understanding. This gap between predicting popular verdicts and demonstrating genuine moral reasoning ability represents a critical blind spot in evaluating whether LLMs could serve as reliable moral cores for autonomous systems increasingly impacting human lives.

[1]All supplementary material available in the extended version

III. Inadequate testing of moral information gathering: Few studies assess models' ability to recognize when they need more information for ethical judgments, with 29 being a notable exception. This oversight is significant, as both moral advisors and autonomous agents may routinely face situations with incomplete information, requiring clarity about whether to gather more data before making judgments. Models that always produce verdicts regardless of information adequacy may perform well in standard evaluations but would fail in real-world ethical decision-making, which often requires acknowledging uncertainty and seeking clarification.

Our proposed methodology: To address these limitations, we propose a novel experimental methodology that moves beyond simple verdict comparison to assess several distinct aspects of moral competency, including info-gathering. We ask models to:

1. Identify morally relevant features (MRF) of scenarios,
2. Determine how important each feature is,
3. Assign moral reasons to each feature (*i.e.* to tell us why the features matter),
4. Synthesize these reasons into all-things-considered moral judgements, and
5. Identify whether any additional information should be sought out that could be relevant to making a verdict

We demonstrate this methodology with two experiments. The first establishes a baseline using existing ethical scenarios but reveals limitations in assessing true moral competence. The second experiment addresses these shortcomings through novel scenarios specifically designed to test moral sensitivity, reasoning quality, and information gathering capabilities. By comparing model performance against general human population responses and those from professional philosophers, we provide a more accurate evaluation of current LLM ethical competence.

Using this breakdown, we design novel test cases that better reflect real-world moral reasoning challenges by requiring the identification of morally relevant features amidst irrelevant details. We compare performance on these scenarios against traditional vignettes where moral features are **pre-highlighted**, *i.e.* identified textually in various ways (see Appendix 1 for discussion). Through this comparison, we investigate whether current approaches to measuring LLM moral competence are artificially inflating performance by eliminating a crucial aspect of moral reasoning — the ability to discern which features of a situation carry ethical significance.

1.2 Research question & objectives

Building on our analysis of limitations in current LLM ethical evaluations, we investigate three primary research questions:

1. Do LLMs demonstrate greater moral skill than non-expert humans when evaluated on traditional moral dilemma vignettes?
2. How do LLMs compare to non-experts and to professional philosophers when evaluated on novel scenarios specifically designed to test moral sensitivity?
3. How does the design of moral vignettes—particularly whether they pre-identify morally relevant features—affect our assessment of LLM moral competence?

We test the following two hypotheses:

- **Hypothesis 1:** LLMs' responses display greater moral skill than humans as rated by humans.
- **Hypothesis 2:** Philosophers display greater moral skill than non-philosophers as rated by humans.

2 First Experiment

2.1 Materials

Vignettes. We sourced vignettes from existing literature, specifically from the Daily Dilemmas dataset (7), the Off the Rails dataset (11), the Moral Stories corpus (10), and cases from 31. Three

vignettes were selected from each paper for a total of 12. As much as possible, we used the vignettes unchanged, though light cosmetic changes were made to some vignettes to make them fit our project (see Appendix 3).

Systems. We asked participants to compare responses from seven systems: six large language models (LLMs) and one group of non-philosopher humans. The LLMs tested were Llama 3.3 (23), GPT-4o (27), DeepSeek V3 Chat (9), Gemini 1.5 Flash (13), GPT-o1 (28), and Claude 3.7 Sonnet (2).

LLM vignette responses. We prompted each of the six LLMs to respond to all 12 vignettes via API calls and responses were collected in the same JSON format from all models. We used default values of each model for temperature (randomness of model output), maximum token and other settings. We presented each vignette as a new API call with the same system prompt, and a user prompt containing the vignette and the same 5 questions in the human survey[2]. The second question asked to rank the features by assigning each feature a weight and that all weights must add up to 100 for each vignette. We observed that all answers to this question were as instructed.

Initially, we generated five responses per model, but since each LLM's responses were largely self-similar, we used only the first response from each model for each vignette[3]. To quantify the semantic similarity (cosine similarity), we used DistilRoBERTa transformer-based language model as a cross encoder (34); average scores per model ranged from 0.58 to 0.82 (1 indicates very similar semantic properties).

Human vignette responses. We also collected human responses to the vignettes *via* the survey platform Prolific[4], recruiting 22 participants to respond to the 12 vignettes (see example survey in Appendix 4). Each participant responded to three vignettes, ensuring multiple responses for each vignette (4-6 responses per vignette). We selected the best response for each vignette based on two primary criteria: whether the respondent identified substantive moral features (rather than merely stating values or posing questions), and whether they provided considered responses (more than just one or two words per question). When multiple responses satisfied these criteria, the authors selected the response with the most extensive and persuasive moral justifications, constituting an additional layer of curation. The purpose of this step was to ensure that the responses demonstrated the required features, thereby ensuring construct validity of our evaluation.

2.2 Human Assessors

We recruited 590 participants[5] *via* Prolific to provide comparative judgements on the quality of pairs of vignette responses. Half of the participants were female, 48% male, 2% identified as non-binary, and 1% did not disclose gender. All age categories from 18 to 84 were represented with 55% of participants falling in the 25-44 age group. The majority (69%) of participants reside in the United States of America and Canada. 47% declared no religion and 38% were Christians, the remaining 15% of participants follow other religions. Almost all participants have at least high school or higher education and 42% hold a bachelors or undergraduate degree. Full demographic details are reported in Appendix 2.

2.3 Procedure

Assessors were randomly assigned to evaluate five of the 12 vignettes. For each vignette, they completed one pairwise comparison between two different systems. Each possible pair of systems (21 pairs in total) was evaluated across the 12 vignettes, with each pair appearing at least 8 times across different participants. To control for ordering effects, we balanced the presentation of pairs:

[2]System and user prompt details are included in Appendix 5 and the Python code can be found on https://github.com/mint-philosophy/Measuring-Moral-Skill-in-LLMs

[3]All responses are available on https://github.com/mint-philosophy/Measuring-Moral-Skill-in-LLMs/blob/main/llm_answers_experiment1.md

[4]https://www.prolific.com/

[5]599 were initially recruited, however seven participants were excluded because they did not agree to have their data anonymised and two were excluded because they selected only the first or only the second option throughout the study, resulting in a final sample of 590 participants. No participants were excluded for failing an attention check, or for taking less than 30 seconds to complete the survey.

for half of the vignettes, system pairs appeared in one order, and for the other half, they appeared in the reverse order. We varied the order for odd and even numbers (e.g., if the comparison of Claude vs. GPT-o1 was presented in that order for vignette 1, 3, 5, ..., it was presented as GPT-o1 vs. Claude for vignette 2, 4, 6, ...).

For each comparison, participants were asked five questions:

1. "Which of these two participants best identified the morally salient features of this scenario?"
2. "Which of these two answers does a better job of weighting the relative importance of the moral features they've identified?"
3. "In light of the features each participant has identified, which of these two participants do you think did a better job of associating reasons with these features?"
4. "Which of these participants did a better job of combining their various reasons into clear, overall moral conclusions?"
5. "Which of these participants did a better job of identifying additional information that might have been relevant to making a moral judgement about this scenario?"

An attention check was included at the end of the survey, asking participants to select the second of two options; no participant failed the attention check.

2.4 Analysis

We used the Bradley-Terry model to analyze the data, which estimates a log-odds parameter for each system representing its inferred level of moral skill based on pairwise comparison outcomes. For each of the five questions, we fit a separate Bradley-Terry model across all 12 vignettes. We used the BradleyTerry2 package (version 1.1.2; 36) in R Studio (30), and created graphs using ggplot2 (version 3.5.1; 37). Based on a power analysis conducted prior to our experiment, we determined that with the best system having approximately a 60% chance of outperforming the worst system (a 0.4 log-odds gap), we would need each pair to be repeated 8 times to achieve 85.6% power ($\alpha = 0.05$). This resulted in 2,016 total comparisons (21 pairs × 12 vignettes × 8 repetitions), requiring approximately 403 judges (since each judge provided five evaluations).

2.5 Results

We analysed participants' judgements for each of the five questions using the Bradley-Terry model, with the human vignette responses serving as the reference category. This allowed us to estimate the relative performance of each LLM compared to human responses across different dimensions of moral reasoning. Positive coefficient values indicate that a system was judged better than the human reference, while negative values indicate worse performance. Below we summarise the findings and include plots for Questions 1 and 5 - see Appendix 5 for the complete set of plots.

Identification of morally salient features: For the first question, four LLMs significantly outperformed human responses ($\alpha = 0.05$): GPT-4o ($\beta = 0.33, p < .001$), GPT-o1 ($\beta = 0.24, p < .001$), and DeepSeek Chat ($\beta = 0.17, p < .01$). Claude 3.7 Sonnet also performed significantly better than humans ($\beta = 0.14, p = .03$). The remaining LLMs—Llama 3.3 and Gemini 1.5 Flash—did not differ significantly from humans in their ability to identify morally salient features.

Weighting moral features: When assessing the ability to appropriately weight the relative importance of identified moral features (Question 2), three LLMs again demonstrated superior performance compared to humans: GPT-4o ($\beta = 0.28, p < .001$), GPT-o1 ($\beta = 0.25, p < .001$), and Claude 3.7 Sonnet ($\beta = 0.18, p < .01$). DeepSeek Chat, Gemini 1.5 Flash, and Llama 3.3 showed no significant difference from human responses.

Associating reasons with features: For Question 3, only GPT-4o significantly outperformed humans ($\beta = 0.18, p < .01$). The remaining five LLMs showed no significant difference from human performance in their ability to associate reasons with identified moral features.

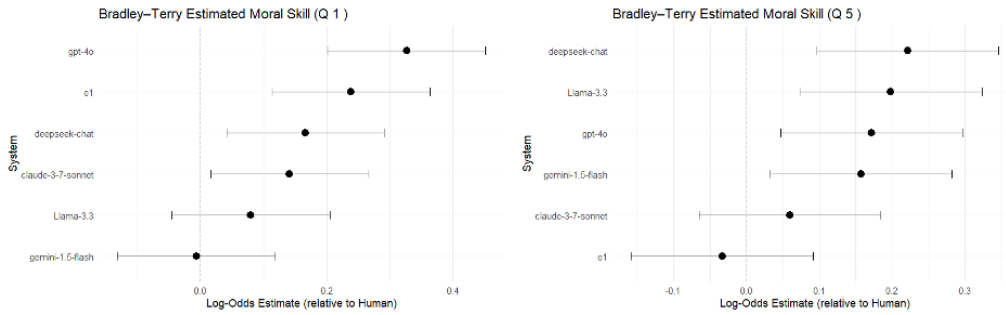

Figure 1: Bradley–Terry model estimates of LLM performance identifying morally salient features in published vignettes (Question 1 on the left and Question 5 on the right). The dotted line at 0 indicates the performance of the general public (human reference group). Error bars show 95% confidence intervals.

Forming clear moral conclusions: When evaluating the ability to combine reasons into clear moral conclusions (Question 4), only Llama 3.3 significantly outperformed humans ($\beta = 0.20, p < .01$). The other five LLMs did not differ significantly from humans.

Identifying additional relevant information: For the final question, four LLMs significantly outperformed humans: DeepSeek Chat ($\beta = 0.22, p < .001$), Llama 3.3 ($\beta = 0.20, p < .01$), GPT-4o ($\beta = 0.17, p < .01$), and Gemini 1.5 Flash ($\beta = 0.16, p < .05$). Claude 3.7 Sonnet and GPT-o1 showed no significant difference from humans.

Overall performance patterns: Across all five dimensions of moral reasoning, GPT-4o demonstrated the most consistent superior performance, significantly outperforming humans on four of the five questions. DeepSeek Chat and Llama 3.3 each significantly outperformed humans on three questions, while Claude 3.7 Sonnet and GPT-o1 each excelled on two questions. Gemini 1.5 Flash showed the least advantage over humans, significantly outperforming them on only one question.

No LLM performed significantly worse than humans on any dimension of moral reasoning evaluated in this study, or, to put the result more strikingly: All LLMs did as well or better than humans on our moral reasoning tasks. The strongest LLM advantages compared to the human baseline were observed in identifying morally salient features (Question 1) and identifying additional relevant information (Question 5), while the smallest advantages were seen in associating reasons with features (Question 3).

To measure the consistency of the human judgements of the responses, we calculated per-item agreement. This is the proportion of raters who, for each specific vignette × question × system-pair, chose the majority option. The mean item agreement for the first study was 0.724 (SD = 0.144), indicating that, on average, nearly three-quarters of raters concurred on the "winner." A Monte Carlo simulation under random selection confirmed that this exceeds chance $p < .001$, suggesting good consistency in people's judgements of which responses were best.

2.6 Discussion

Our first experiment revealed several methodological limitations that affect how we interpret LLM performance in moral reasoning tasks.

First, we sourced vignettes from published literature (7, 11, 10, 31), creating a potential data contamination issue. Since these scenarios may have appeared in LLM training data, observed performance could reflect memorization rather than genuine moral reasoning capability. This limitation is increasingly problematic as newer models are trained on larger datasets that likely include prominent ethics benchmarks.

Second, and perhaps more fundamentally, our methodology relied on pre-constructed moral vignettes that had already undergone significant editorial filtering. Although methodologically convenient, this process bypasses a crucial aspect of moral competence. In real-world scenarios, moral agents

must first identify which features of a situation carry ethical significance before deliberating on them. Humans perform this filtering effortlessly: when encountering an injured person on the street, we immediately recognize the injury as morally significant while disregarding irrelevant contextual details such as the day of the week, ambient lighting, or the person's clothing (24, 25). We also intuitively gauge which contextual features might be relevant (such as the person's apparent age or vulnerability) and weigh them appropriately. By providing systems with neatly packaged moral scenarios, we effectively perform the most challenging aspects of moral reasoning on behalf of the moral agents being evaluated.

Third, while comparing LLM responses to those of non-philosopher humans yielded interesting insights, this benchmark does not establish whether these systems have achieved expert-level moral reasoning. The finding that LLMs can produce responses judged superior to those of the general population sets a relatively low bar that does not address whether these systems can approach the reasoning quality of those with specialized training in ethical analysis. Our second experiment was designed to address these limitations, discussed in the sequel.

3 Second Experiment

3.1 Materials

Vignettes. We created 12 entirely novel moral scenarios drawn from an unpublished database of naturalistic vignettes. When developing these vignettes, we employed three specific strategies:

1. Removing morally loaded language that might cue particular responses
2. Assessing the systems' capability to request clarification when needed before making moral judgments

Our methodical approach to vignette creation involved selecting 30 third-person narratives from a pool of 351, replacing generic names with ethnically diverse alternatives. We have then asked GPT-4o to modify them according to strategies 1 and 2 [6]. From these modified scenarios, an independent expert in moral philosophy selected the best 12 cases, establishing face validity of our survey (see Appendix 3). These were specifically designed to eliminate the risk of training data contamination for the LLMs.

Systems. In addition to the seven systems from our first experiment (six LLMs and non-philosopher humans), we included an eighth system: responses from professional philosophers. This addition established a more rigorous benchmark for evaluating LLM performance.

LLM vignette responses. The procedure for collecting LLM response was the same as in our first experiment[7]. We again observed that all answers to the second question of weighing features added up to 100.

Human vignette responses. For the non-philosopher responses, we recruited 22 participants via Prolific to respond to the 12 vignettes, in the same manner as for Study 1.

For the domain expert responses, 12 we recruited professional philosophers (scholars employed in philosophy departments at universities) to respond to the 12 vignettes. Philosophers were recruited via the professional networks of the authors and were selected for their specific expertise in analytic moral philosophy. Each philosopher responded to three vignettes, ensuring even distribution across all vignettes (3 responses per vignette). We selected the best response for each vignette using criteria similar to those applied to non-philosopher responses, with additional consideration given to philosophical sophistication and depth of moral analysis.

Responses from 9 of the 12 philosophers were used. Most of these philosophers were native English speakers (7), most had a PhD in philosophy (7), and most were male (7). Countries of residence

[6]The prompts used to modify the vignettes can be found in https://github.com/mint-philosophy/Measuring-Moral-Skill-in-LLMs/blob/main/modify_vignettes.py

[7]All responses are available on https://github.com/mint-philosophy/Measuring-Moral-Skill-in-LLMs/blob/main/llm_answers_experiment2.md

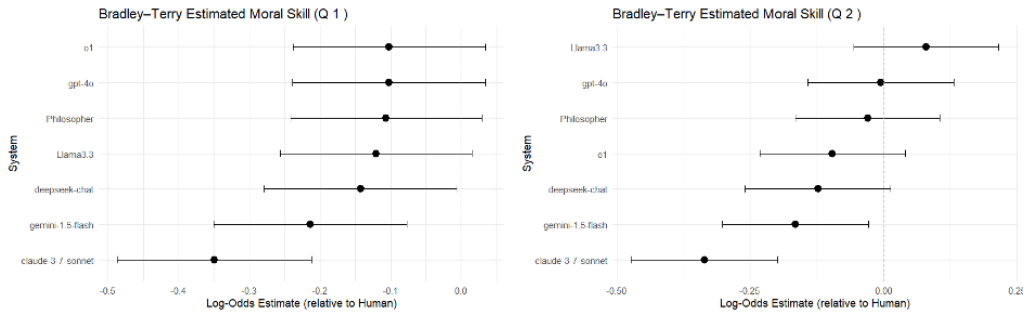

Figure 2: Bradley–Terry model estimates of systems' performance on identifying morally salient features in novel vignettes (Question 1 on the left and Question 2 on the right). The dotted line at 0 indicates the performance of the general public (human reference group). Error bars show 95% confidence intervals.

included the United States of America (2), Australia (2), the United Kingdom (2), Singapore (2) and another country not specified. Philosophers were aged 25-34 (4) or 35-44 (5). Most were Christian (4) or had no religion (4).

3.2 Human Assessors

We recruited 586 participants[8] to serve as human assessors *via* Prolific. This experiment attracted more female participants (60%) than the first experiment. Age profile is similar with 55% of participants falling in the 25-44 age group. Unlike the first experiment, the majority (73%) of participants reside in the United Kingdom. Similar to the first experiment, 51% declared no religion, 39% were Christians, with the remaining 10% of participants following other religions. Education profile was very similar to the first experiment with all participants having at least high school or higher education and 42% holding a bachelor's / undergraduate degree. Full demographic details are reported in Appendix 2.

3.3 Procedure and analysis

The procedure and analysis methods were similar to those used in Study 1, with the only difference being the number of system pairs compared (28 pairs instead of 21, due to the addition of the philosopher system). This resulted in 2,688 total comparisons (28 pairs × 12 vignettes × 8 repetitions), requiring approximately 538 judges according to the power analysis procedure. To enable comparison between the visualisations of the two experiments, we again used the general population vignette responses as the baseline against which other systems were compared.

3.4 Results

Using the same analytical approach as Study 1, we analysed participants' judgements for each of the five questions using the Bradley-Terry model, with human responses again serving as the reference category. Below we summarise the findings and include plots for Questions 1 and 2 - see Appendix 5 for the complete set of plots.

Identification of morally salient features: For the first question, three LLMs performed significantly worse than non-philosopher responses: Claude 3.7 Sonnet ($\beta = -0.35$, $p < .001$), Gemini 1.5 Flash ($\beta = -0.21$, $p < .01$), and DeepSeek Chat ($\beta = -0.14$, $p < .05$). The remaining systems—Llama 3.3, GPT-4o, GPT-o1, and Philosopher responses—did not differ significantly from human responses in identifying morally salient features.

[8] 600 were initially recruited, however three were excluded for not agreeing to having their data anonymised, and eleven were excluded for selecting only the first or only the second option throughout the study, resulting in a final sample of 586 participants. As in Study 1, no participants completed the survey in less than 30 seconds or failed an attention check qeuestion.

Weighting moral features: When assessing the ability to appropriately weight the relative importance of identified moral features (Question 2), two LLMs demonstrated significantly worse performance compared to humans: Claude 3.7 Sonnet ($\beta = -0.34, p < .001$) and Gemini 1.5 Flash ($\beta = -0.17, p < .05$). DeepSeek Chat showed marginally worse performance ($\beta = -0.12, p < .10$). The remaining systems—GPT-4o, Llama 3.3, GPT-o1, and Philosopher responses—showed no significant difference from human responses.

Associating reasons with features: For Question 3, none of the systems showed statistically significant differences from human performance. All coefficient values were relatively small, suggesting comparable performance across all systems in their ability to associate reasons with identified moral features.

Forming clear moral conclusions: When evaluating the ability to combine reasons into clear moral conclusions (Question 4), none of the systems differed significantly from human performance. While philosopher responses showed the highest positive coefficient ($\beta = 0.10$) and Claude 3.7 Sonnet showed the most negative coefficient ($\beta = -0.10$), these differences were not statistically significant.

Overall performance patterns: Across all five dimensions of moral reasoning in Study 2, a notably different pattern emerged compared to Study 1. Rather than LLMs consistently outperforming humans, several LLMs showed significantly worse performance than humans on some dimensions, particularly in identifying morally salient features and weighting moral features. Claude 3.7 Sonnet showed the most consistent underperformance relative to humans, performing significantly worse on two questions.

On the other hand, most systems performed comparably to non-philosophers in associating reasons with features and forming clear conclusions. Gemini 1.5 Flash showed a notable advantage in identifying additional relevant information, despite underperforming in other areas.

Philosopher responses generally performed on par with human responses across all five dimensions, with no significant advantages or disadvantages detected. This suggests that our non-philosopher human sample was able to perform moral reasoning at a level comparable to professional philosophers on these novel moral vignettes.

For this second study, the per-item agreement was 0.734 (SD = 0.148), indicating once again that, on average, nearly three-quarters of raters concurred on the "winner" of a pairwise comparison. Monte Carlo simulation under random selection again confirms that this exceeds chance $p < .001$.

3.5 Discussion

In Experiment 2, we investigated the robustness of LLMs' moral-sensitivity capacities by testing their performance on novel vignettes alongside professional philosophers. Models significantly underperformed in independently identifying morally salient features without pre-packaged support, highlighting the impact of training-data overlap and prior cueing. However, the models maintained robust reasoning and judgement capabilities when morally salient features were provided, despite their initial detection weaknesses. Additionally, Gemini 1.5 Flash uniquely surpassed humans in recognising the need for additional information, suggesting potential advantages in meta-cognitive heuristics fostered by certain instruction-tuning methods.

Surprisingly, expert philosophers did not outperform lay participants in any evaluated dimension, contradicting our second hypothesis.

The results from Experiment 1 indicated that LLMs can perform comparably to humans in structured moral reasoning tasks. However, Experiment 2 demonstrated that altering the vignettes to require independent identification of morally salient features significantly diminished LLM performance. This finding suggests current benchmarks might overestimate LLMs' moral reasoning capabilities by removing the task of discerning moral relevance from noisy or ambiguous information – a task we consider fundamental for genuine moral reasoning skill. Overall, these experiments highlight the importance of designing evaluations that reflect the real-world complexity of moral reasoning, ensuring that benchmarks challenge models not only to reason effectively once relevant features are identified but also to reliably recognise what morally matters amidst complex and varied contexts.

4 Limitations

Our study had several limitations. First, the sequential moral reasoning assessment created uneven comparisons, as performance on early tasks affected inputs for subsequent tasks, making it difficult to isolate specific reasoning capabilities. Future research should evaluate each reasoning stage with standardised inputs.

Second, stylistic differences between human and LLM responses may have influenced evaluators beyond substantive moral reasoning. Future work could address this by having LLMs rephrase human responses or vice versa to ensure stylistic consistency.

Third, our experiments were not designed to investigate correlations *between* moral reasoning subtasks. Determining whether these skills operate independently or together would provide valuable insights into the structure of moral competence. Our focus on relative rankings rather than absolute metrics prevented such correlation analysis, and future experiments could relax this limitation.

Finally, our vignette set comprises 12 cases that do not span the full moral domain; they centre on interpersonal dilemmas. Though this represents an important limitation to our study, we made this restriction intentionally: first, to enhance ecological validity for a key motivating application—Artificial Moral Advisors that assist with everyday, interpersonal decisions—so evaluating that surface first is appropriate; second, to avoid muddying the analysis with legal, institutional, or political-legitimacy questions that risk conflating moral assessment with other flavours of normativity; and third, to maximise comparability with prior work, which predominantly uses short, person-centred scenarios. This continuity let us cleanly demonstrate the "reversal" we observe when superficial cues are removed.

5 Conclusion

This study presented a novel approach to evaluating moral reasoning skills in Large Language Models (LLMs), addressing critical gaps in existing methodologies. Our findings suggest that while current LLMs demonstrate competence comparable to or slightly superior to non-expert humans on traditional, pre-packaged ethical scenarios, their performance markedly declines when required to identify morally relevant features amidst irrelevant information.

Specifically, our first experiment, using standard ethical vignettes, indicated that several LLMs significantly outperformed humans in recognizing morally salient features, appropriately weighting their importance, and identifying additional relevant information. However, our second experiment, using novel scenarios with embedded irrelevant details, revealed a notable reduction in performance, with multiple LLMs performing significantly worse than humans, particularly in discerning morally relevant features and assigning appropriate weight to these features.

This performance gap underscores the limitations of existing benchmarks that overestimate the moral competence of LLMs by eliminating the complexity of identifying moral relevance from noisy inputs. Our results suggest the need for a reevaluation of moral reasoning benchmarks, emphasizing tasks that mirror the complexities inherent in real-world ethical decision-making.

Further research should focus on refining assessment techniques, controlling for stylistic differences between human and model responses, and examining interrelations between different moral reasoning subtasks. Ultimately, our enhanced methodological framework aims to foster more ethically competent AI systems capable of reliably assisting human decision-making in morally complex situations.

Acknowledgments

The authors would like to thank Charis Yang and Jake Stone from Australian National University's Machine Intelligence and Normative Theory Lab for their research assistance. We would also like to thank Tegan Maharaj for their invaluable feedback. This work was performed with the assistance of funding from an OpenAI industry grant for Agentic Evaluations from 2024-2025, along with an ARC grant LP210200818 and funding from the Templeton World Charity Foundation.

Ethics Statement

This research was approved by the Australian National University Human Ethics Office (approval number: **H/2024/0534**). All participants provided informed consent, and their privacy and confidentiality were ensured. Broadly speaking, our work engages explicitly with ethical impacts of AI, as the moral (in)competence of LLMs has potentially large societal impact.

References

[1] Albrecht, J., Kitanidis, E., and Fetterman, A. J. (2022). Despite "super-human" performance, current LLMs are unsuited for decisions about ethics and safety. arXiv:2212.06295 [cs].

[2] Anthropic (2025). Claude 3.7 Sonnet [Large language model].

[3] Bignotti, C. and Camassa, C. (2024). Legal Minds, Algorithmic Decisions: How LLMs Apply Constitutional Principles in Complex Scenarios.

[4] bin Ahmad, M. S. Z. and Takemoto, K. (2024). Large-scale moral machine experiment on large language models.

[5] Bonagiri, V. K., Vennam, S., Govil, P., Kumaraguru, P., and Gaur, M. (2024). SaGE: Evaluating Moral Consistency in Large Language Models. arXiv:2402.13709 [cs].

[6] Bulla, L., De Giorgis, S., Mongiovì, M., and Gangemi, A. (2025). Large Language Models meet moral values: A comprehensive assessment of moral abilities. *Computers in Human Behavior Reports*, 17:100609.

[7] Chiu, Y. Y., Jiang, L., and Choi, Y. (2025). DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life. arXiv:2410.02683 [cs].

[8] Clark, J. (2022). Import AI 310: AlphaZero learned Chess like humans learn Chess; capability emergence in language models; demoscene AI.

[9] DeepSeek (2025). deepseek-ai/DeepSeek-V3 [Large language model]. original-date: 2024-12-26T09:52:40Z.

[10] Emelin, D., Bras, R. L., Hwang, J. D., Forbes, M., and Choi, Y. (2020). Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. arXiv:2012.15738 [cs].

[11] Fränken, J.-P., Khawaja, A., Gandhi, K., Moore, J., Goodman, N. D., and Gerstenberg, T. (2023). Off The Rails: Procedural Dilemma Generation for Moral Reasoning. In *AI Meets Moral Philosophy and Moral Psychology Workshop*.

[12] Giubilini, A. and Savulescu, J. (2018). The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. *Philosophy & Technology*, 31(2):169–188.

[13] Google DeepMind (2025). Gemini 1.5 Flash [Large language model].

[14] Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. (2023). Aligning AI With Shared Human Values. arXiv:2008.02275 [cs].

[15] Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., Song, D., Li, B., and Steinhardt, J. (2022). What Would Jiminy Cricket Do? Towards Agents That Behave Morally. arXiv:2110.13136 [cs].

[16] Jain, S., Calacci, D., and Wilson, A. (2024). As an AI Language Model, "Yes I Would Recommend Calling the Police": Norm Inconsistency in LLM Decision-Making. arXiv:2405.14812 [cs].

[17] Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., and Choi, Y. (2022). Can Machines Learn Morality? The Delphi Experiment. arXiv:2110.07574 [cs].

[18] Jin, Z., Kleiman-Weiner, M., Piatti, G., Levine, S., Liu, J., Gonzalez, F., Ortu, F., Strausz, A., Sachan, M., and Mihalcea, R. (2024). Language Model Alignment in Multilingual Trolley Problems. *arXiv preprint*.

[19] Jin, Z., Levine, S., Gonzalez, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., and Schölkopf, B. (2022). When to make exceptions: exploring language models as accounts of human moral judgment. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 28458–28473, Red Hook, NY, USA. Curran Associates Inc.

[20] Kumar, P., Lau, E., Vijayakumar, S., Trinh, T., Team, S. R., Chang, E., Robinson, V., Hendryx, S., Zhou, S., Fredrikson, M., Yue, S., and Wang, Z. (2024). Refusal-Trained LLMs Are Easily Jailbroken As Browser Agents. arXiv:2410.13886 [cs].

[21] Lazar, S. (2024). Frontier AI Ethics: Anticipating and Evaluating the Societal Impacts of Language Model Agents. arXiv:2404.06750 [cs].

[22] Lourie, N., Bras, R. L., and Choi, Y. (2021). SCRUPLES: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13470–13479. Number: 15.

[23] Meta AI (2024). Llama 3.3 [Large language model].

[24] Mole, C. (2022). The Moral Psychology of Salience. In Archer, S., editor, *Salience: A Philosophical Inquiry*, pages 140–158. Routledge.

[25] Murphy, F. C., Wilde, G., Ogden, N., Barnard, P. J., and Calder, A. J. (2009). Assessing the automaticity of moral processing: Efficient coding of moral information during narrative comprehension. *The Quarterly Journal of Experimental Psychology*, 62(1):41–49. Place: United Kingdom Publisher: Taylor & Francis.

[26] Nunes, J. L., Almeida, G. F. C. F., de Araujo, M., and Barbosa, S. D. J. (2023). Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations.

[27] OpenAI (2024a). GPT-4o [Large language model].

[28] OpenAI (2024b). OpenAI o1 [Large language model].

[29] Pyatkin, V., Hwang, J. D., Srikumar, V., Lu, X., Jiang, L., Choi, Y., and Bhagavatula, C. (2023). ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations. arXiv:2212.10409 [cs].

[30] R Core Team (2021). R: The R Project for Statistical Computing.

[31] Rao, A., Khandelwal, A., Tanmay, K., Agarwal, U., and Choudhury, M. (2023). Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. arXiv:2310.07251 [cs].

[32] Russo, G., Nozza, D., Röttger, P., and Hovy, D. (2025). The pluralistic moral gap: Understanding judgment and value differences between humans and large language models. *arXiv preprint arXiv:2507.17216*.

[33] Sachdeva, P. and van Nuenen, T. (2025). Normative evaluation of large language models with everyday moral dilemmas. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 690–709.

[34] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs].

[35] Tanmay, K., Khandelwal, A., Agarwal, U., and Choudhury, M. (2023). Probing the Moral Development of Large Language Models through Defining Issues Test.

[36] Turner, H. and Firth, D. (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software*, 48:1–21.

[37] Wickham, H. (2016). *ggplot2*. Use R! Springer International Publishing, Cham.