Beyond Verdicts: Evaluating Language Model Moral Competence

Aaron J. Snoswell[1], **Daniel Kilov**[2], **Seth Lazar**[2]

[1]Digital Media Research Centre GenAI Lab, Queensland University of Technology, Kelvin Grove, QLD 4012, Australia
[2]Machine Intelligence and Normative Theory Lab, Australian National University, Acton, ACT 2601, Australia
a.snoswell@qut.edu.au, daniel.kilov@anu.edu.au, seth.lazar@anu.edu.au

Abstract

As Large Language Models (LLMs) are increasingly deployed as Artificial Moral Advisors and autonomous agents making ethical decisions, evaluating their moral competence has become critical. However, existing evaluations may inadequately assess the moral reasoning capabilities needed for real-world deployment, focusing primarily on whether models can match human judgments on carefully curated ethical scenarios.

We surveyed 69 papers evaluating LLM ethical competence (2020-2025) and developed a taxonomy categorizing evaluations across datasets, behaviors, and metrics. Our comprehensive analysis maps the methodological landscape of this rapidly growing field and reveals several critical limitations. Most significantly, the vast majority of studies rely on prepackaged scenarios that highlight morally relevant features, failing to test models' ability to identify ethical considerations in noisy, realistic contexts-what we term "moral sensitivity". Additionally, evaluations overemphasize verdict accuracy rather than assessing moral reasoning quality and steerability, with few studies testing whether models can be appropriately guided toward different ethical frameworks. Most studies rely on "ground truth" comparisons despite philosophical arguments that reasonable moral pluralism precludes definitive moral ground truth.

In light of these gaps, we argue for a significant methodological shift: moving from curated scenarios to unfiltered information streams, from verdict accuracy to reasoning quality and steerability, and from ground truth metrics to assessments of reasonableness and consistency. This reorientation is essential for developing AI systems that can navigate moral complexity in real-world deployment scenarios.

Extended version with tables and appendices — https://philpapers.org/rec/SNOBVE

Companion repository — https://github.com/mint-philosophy/Beyond-Verdicts-Paper

1 Introduction

Large Language Models (LLMs) are increasingly deployed in roles that require moral judgment. They serve as Artificial Moral Advisors (AMAs) offering ethical guidance to users facing dilemmas, and as autonomous agents making decisions with moral consequences. From healthcare chatbots advising on treatment decisions to AI systems managing content moderation, these models are already navigating complex ethical terrain in real-world applications.

The stakes for getting this right are high. Poor moral reasoning in deployed systems could lead to harmful advice, erosion of trust in AI assistance, or autonomous agents that act unethically when operating without human oversight. Yet evaluating whether LLMs possess the moral competence needed for such roles presents fundamental challenges that extend far beyond typical AI benchmarking.

Current evaluation approaches may be inadequately preparing us for these deployment realities. The vast majority of existing evaluations focus on whether models can match human judgments on carefully curated ethical scenarios—essentially testing whether an AI can solve prepackaged moral puzzles. But real-world moral competence requires something quite different: the ability to identify ethical considerations in noisy, realistic contexts, provide high-quality reasoning for moral judgments, and adapt appropriately to different ethical frameworks.

This mismatch between how we evaluate moral competence and how we deploy morally capable systems creates a critical gap in our understanding. A model that performs well on existing benchmarks may still fail catastrophically when faced with the unfiltered complexity of real-world ethical decision-making.

To address this challenge, we performed a systematized survey of papers evaluating LLM ethical competence published between 2020 and 2025 from a philosophical perspective.[1] We developed a taxonomy (see Fig. 1) that categorizes evaluations across three key dimensions; **datasets**; **behaviors**; and **metrics**; and report findings across these detailed categories. Our analysis reveals a field experiencing explosive growth—doubling approximately every 14 months—but also exposes systematic limitations that constrain our ability to assess real-world moral competence.

We identify three critical paradigm shifts needed in how we evaluate LLM moral competence: moving from curated scenarios to unfiltered information streams; shifting focus

[1] Meaning, from our perspective and disciplinary viewpoint as authors with expertise as professional analytic moral philosophers and philosophically trained computer scientists

from verdict accuracy to reasoning quality and steerability; and replacing ground truth metrics with assessments of reasonableness and consistency. These reorientations are essential for developing AI systems that can navigate moral complexity in actual deployment scenarios.

Our key contributions are to: (1) map the methodological landscape of this rapidly growing field through comprehensive analysis; (2) identify critical limitations in current approaches that limit their applicability to real-world deployment; and (3) propose concrete directions for developing evaluations that better match the moral competence requirements of deployed systems.

2 Background: Why LLM Moral Competence Evaluation Matters

The Deployment Reality

There are at least two compelling reasons why understanding and evaluating LLM moral competence has become urgent. First, Artificial Moral Advisors (AMAs)—AI systems that *support* a human user by offering moral advice—are already being deployed to provide invaluable moral guidance to those facing ethical dilemmas, especially when acting under time pressure (Giubilini and Savulescu 2018). The usefulness of these systems will be fundamentally bounded by their moral competence. When a healthcare chatbot advises on end-of-life decisions or a personal assistant helps navigate workplace ethics, the quality of moral reasoning directly affects real human outcomes.

Second, as we build more capable autonomous AI agents—AI systems that operate *in-place of* a human user—it becomes critical to design systems that can direct or constrain their choices to conform to moral reasons (Hendrycks et al. 2021b; Lazar 2024; Silen et al. 2023). Unless we carefully sandbox these autonomous agents so that we can anticipate in advance all the ethical considerations they might face and resolve any conflicts ahead of time, they must be able to engage in ethical reasoning *in-situ*. The degree to which autonomous AI agents can be trusted as delegates will be constrained by their moral competence.

The Challenge of Ethical Evaluation

Evaluating moral competence differs fundamentally from other AI capabilities assessment. Ethical[2] competence comprises at least two distinct capabilities: the ability to understand moral concepts and reasons; and the ability to shape one's behavior according to that understanding (Driver 2013). We focus primarily on the first sense, but do so because we are ultimately interested in the second sense.

[2]We use 'ethics' and 'morality' interchangeably throughout. Defining either term without committing to some particular ethical or meta-ethical theory is impossible. Instead, we rely on the commonsense notion that ethics/morality are about what one ought to do, such that failure to do so could justify guilt, or blame. When we extend these terms to LLM behaviour, we are not making a claim about the moral status of LLMs or culpability for moral guilt, but mean simply what they ought to do, such that failure to do so could justify guilt, or blame on behalf of the user and/or developer.

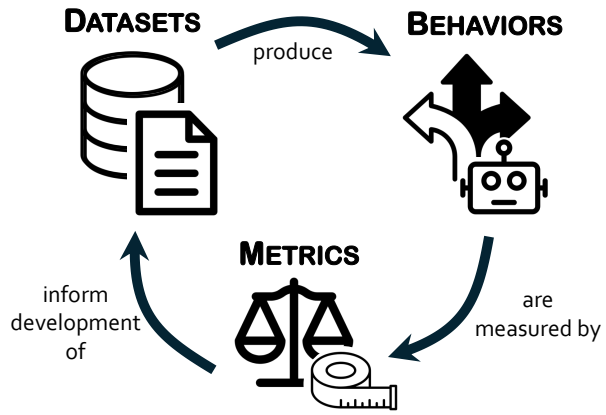

Figure 1: A visualisation of our three-part framework for evaluating evaluations of LLM Moral Reasoning.

The philosophical complexity adds another layer of difficulty. In circumstances of reasonable pluralism—situations where reasonable people may reasonably disagree—there is often no single "correct" moral answer. This makes traditional ground truth evaluation approaches problematic, as they assume definitive truths that may not exist. Unlike factual accuracy or logical reasoning, moral evaluation must grapple with legitimately different ethical frameworks and value systems.

Current ML evaluation paradigms, designed for tasks with clear objective metrics, fall short when applied to moral reasoning. Standard approaches like accuracy against human annotations become philosophically problematic when the underlying moral questions admit multiple reasonable answers.

Capability Overhang: A Lurking Opportunity

These evaluation challenges take on particular urgency when we consider a striking feature of current LLMs: their moral understanding capabilities far exceed their moral behavior. A language model's ability to understand moral concepts does not predict its ability to conform its behavior to that moral understanding (see recent work on alignment failures by Jain, Calacci, and Wilson (2024); Kumar et al. (2024)). This disconnect is unsurprising given that LLMs are not integrated individuals (Khan, Casper, and Hadfield-Menell 2025).

This "capability overhang" represents latent moral understanding that could be leveraged through post-training techniques. Since generation is harder than verification (Arora and Barak 2009), successful post-training exploits this asymmetry using good verifiers to train better generators (Tie et al. 2025).[3] High-quality evaluation frameworks thus become not merely assessment tools, but enablers of substantially improved moral behavior. This makes getting

[3]The underlying principle was described in (Minsky 1986, chapter 7); and (Minsky 2006, chapter 2)

evaluation right considerably more valuable than if moral understanding and behavior were already well-aligned.

Current State of the Field

Against this backdrop, the field of LLM moral competence evaluation is experiencing explosive growth, approximately doubling every 14 months since 2020. This expansion reflects both deployment urgency and recognition that current approaches may be insufficient. The field has matured methodologically—early papers used simplistic cases without moral conflicts, while recent work employs complex scenarios with value conflicts and sophisticated bias controls.

But critical gaps remain. To train LLMs for better ethical behavior, we need evaluations that adequately assess moral reasoning capabilities for real-world deployment. Current evaluations may provide false confidence by testing moral competence in artificially simplified contexts that bear little resemblance to the messy, unfiltered environments where deployed systems operate. This survey summarizes this rapidly evolving work, identifies systematic limitations, and charts a path toward evaluations that better serve morally competent AI deployment.

Related Work

Complementary survey papers have explored evaluations of language model behaviour in general, as distinct from moral understanding. For example, Chang and Bergen (2023) considers LLM behaviour across many different dimensions; Liu et al. (2024) is one of many similar papers exploring evaluations of LLM alignment with specific social norms, and alignment more generally. Reinig et al. (2024) undertook a broad survey of uses of moral concepts in Natural Language Processing—considering all papers that use moral concepts for text processing in any way, as distinct from those on the evaluation of LLM moral understanding—our focus is on a subset of the subdomain that they describe as 'Morality in AI Systems' (Reinig et al. 2024, p. 4144), and consider only in passing. Vida, Simon, and Lauscher (2023) similarly evaluated 92 NLP papers that invoke concepts of morality and ethics, and showed that they often failed to define their terms, and lacked philosophical foundations, context, and diversity. They in particular called for a greater integration of philosophy and NLP—a call to which our paper is a response.

While these other surveys provide important context for our work, they were published prior to the majority of the papers that we examine, and were aimed at a clearly distinct set of targets. Our focus is not to evaluate the use of moral concepts and language in NLP research, but specifically to ask whether evaluations of LLM ethical competence are getting at the key elements of moral understanding that we can leverage to design better AMAs and ethical autonomous agents.

3 Method

Our survey focuses on papers since 2020 that evaluate LLM ethical competence. We identified these papers by searching Google Scholar, Semantic Scholar, and arXiv; in addition, we used Elicit.com and ChatGPT's 'Deep Research' feature to perform searches to help surface any additional papers.[4] We searched with combinations of keywords including 'Moral Competence', 'Moral Reasoning', 'Large Language Model', etc. to establish a preliminary set of papers. Searches were performed between March and July 2025.

Two team members (author A.S. and research assistant C.Y.) screened papers with assistance from Claude and ChatGPT to only include those published 2020 or later; and which contained empirical evaluations of LLMs, not just theory, critique, or thought experiments; and which proposed or applying a benchmark, dataset, framework, or method to evaluate morality, moral reasoning, or competence. Disagreements were resolved by discussion with the entire project team.

We validated our corpus by manually reviewing full-texts to manually confirm that every paper substantively addressed the evaluation of LLM moral competence under some description, rejecting some that focused narrowly on evaluating model alignment or moral/cultural identity instead (see the extended paper version for details).

We then identified further papers by performing recursive and exhaustive forward and backward citation snowballing until we reached saturation, finding no new papers after 6 iterations of snowballing. **Our search resulted in a corpus of 69 included papers and preprints**, 2 in 2020, 3 in 2021, 4 in 2022, 14 in 2023, 34 in 2024, and 12 from 2025, suggesting an exponential curve with a doubling rate of just over 14 months.

The next section reports selected observations from our analysis, and in the extended paper version we provide detailed comparisons of papers categorized by type, format, and measurement objectives, serving as a reference guide for researchers seeking to understand the landscape of available tools or select appropriate evaluation frameworks for specific research questions.

4 The State of LLM Ethics Evals

We focus on three dimensions that we think most relevant to the project of evaluating LLM moral understanding as a means to designing better AMAs and ethical autonomous agents (see Fig. 1). We focus on the problems that they task LLMs with solving (**datasets**); the **behaviour** that they evaluate; and the **metrics** by which they evaluate that behaviour. The dimensions within each aspect of our taxonomy are described in detail in the extended version of our paper.

Datasets

Dataset construction. Ethical evaluations of LLMs draw from diverse data sources. Most studies source scenarios from established frameworks such as psychological instruments including the Moral Foundations Questionnaire (Nunes et al. 2024) and the Defining Issues Test (Tanmay et al. 2023), online forums (particularly Reddit's 'Am I The Asshole' subreddit[5]) (Alhassan, Zhang, and Schlegel 2022; Lourie, Le Bras, and Choi 2021; Chiu, Jiang, and Choi 2024;

[4]We used these tools only to identify papers; their analysis, though interesting in its own right, is orthogonal to our own.

Russo et al. 2025; Sachdeva and van Nuenen 2025), or philosophical thought experiments like trolley problems (Jin et al. 2025). Crowdsourcing represents a common approach for collecting moral judgments or generating new scenarios (Jin et al. 2022), alongside the emerging trend of using generative AI models to provide judgments or augment existing datasets (Abbo and Belpaeme 2024; Moore, Deshpande, and Yang 2024).

Type of cases. The majority of papers rely on *static* datasets—fixed test cases that remain unchanged during evaluation. This approach dominates the field, appearing in 66 papers (96%), with dataset sizes varying from 8 medical ethics vignettes (Balas et al. 2024) to over 97,000 trolley problem variants (Jin et al. 2025). Only 2 papers (3%) employ *interactive* scenarios where models navigate dynamic moral decision spaces (Hendrycks et al. 2021b; Pan et al. 2023). This predominance of static evaluation is problematic as newer LLMs may have been trained on widely available ethical benchmarks, introducing data contamination risks. An encouraging trend is innovative approaches employing *generative* frameworks at test time (Duan et al. 2024; Fränken et al. 2024; Jiang et al. 2025), producing novel moral dilemmas tailored to specific model capabilities.

Nature of cases. Most papers rely on *artificial* dilemmas (37 papers, 54%) or *staged* everyday scenarios (24 papers, 35%), with few studies (14 papers, 20%) using found cases from *real-world* sources. The prevalence of artificial scenarios like trolley problems (Neuman, Coleman, and Shah 2025; Nie et al. 2023; Krügel, Ostermaier, and Uhl 2023; Rehman, Iqbal, and Shah 2025) likely stems from philosophical traditions. Some recent studies mix artificial and found cases (Jiang et al. 2025), offering more comprehensive evaluation.

Conceptual structuring. Our analysis reveals a strong preference to ground cases in a *theoretical* conceptual base, with 49 papers (71%) explicitly structuring their datasets according to specific ethical or conceptual frameworks. Among these papers, Moral Foundations Theory (MFT) is the most commonly employed framework, appearing in 9 papers (13%). 7 papers (10%) employ established *psychometric instruments* for assessment: there appears to be limited innovation in developing new psychometric tools specifically designed for LLM moral evaluation. 14 papers (20%) adopt an *arbitrary* approach, lacking any conceptual framework. Almost all theoretical frameworks originate from Western philosophical and psychological traditions, with rare exceptions such as (Bhattacharya and Nandi 2023) and (Yu et al. 2024), which draw on ancient Sanskrit moral tales and "traditional Chinese culture and contemporary norms" respectively.

Framing. 45 papers (65%) address how textual scenario formulation and presentation influences model responses. Many implement control strategies including systematic question order variation (Awad et al. 2024), multiple prompt variations (Rao et al. 2023), and adversarial filtering to reduce spurious lexical cues (Hendrycks et al. 2021a). Context minimization studies compare verbose versus minimal prompts (Bignotti and Camassa 2024) to assess how contextual information influences judgement.

Other dimensions. 17 papers (25%) use *simple cases* such as trolley problems, while 53 papers (77%) employ *complex scenarios*. Among papers with complex scenarios, 30 (56%) explicitly feature *conflicting values* designed to challenge simplistic moral reasoning pattens. All but 8 papers use an *anthopocentric* viewpoint—asking LLMs to solve moral problems that a human would face. Almost all the papers we reviewed perform *moral packaging*: offering questions or scenarios with morally relevant features already highlighted through the data curation process. By our assessment, only a single paper (Hendrycks et al. 2021b) provides LLMs with an *raw* information stream from which morally relevant details must be extracted. Finally, we find that text-based evaluation is dominant: of the papers surveyed, only 5 featured data in any format other than text.

Behaviour

Verdicts. All papers bar one (Seror (2025)) evaluate models on their ability to make moral judgments or verdicts on cases, suggesting a strong methodological bias toward viewing moral competence primarily as a classification task. We observe considerable diversity in verdict formats, ranging from binary choices (right/wrong) to Likert scales, to selecting between multiple options, to free-form text judgments.

Steerability. Few papers (13, 19%) evaluate steerability —the ability to flexibly and pluralistically adjust model's moral profile at inference time (*e.g.* through prompting), rather than through pre- or post-training steps— in the context of moral reasoning.

Justifications. 33 of the papers we surveyed (48%) elicit justifications for verdicts. Papers from 2023-2025 more often evaluate justifications than earlier works, indicating an evolving understanding of moral competence as necessarily involving reasoning processes. The approaches to eliciting structured justifications vary considerably—from anchoring in established moral frameworks like Kohlberg's stages (Neuman, Coleman, and Shah 2025), to eliciting intermediate reasoning steps (Ma et al. 2023), to generating 'Rules of Thumb' (Bonagiri et al. 2024).

Moral sensitivity. A third of papers (23, 33%) evaluate models' ability to weigh morally relevant features from the choice situations with which they are confronted—that is, to identify which of the facts of a particular case are important to attend to, in order to form a moral judgment. We call this capacity *Moral Sensitivity* (Lazar 2024). However, some papers assess moral sensitivity only implicitly through measuring how models respond to scenario variations rather than through direct tests of feature identification, potentially missing important aspects of models' moral reasoning capabilities.

[5] A popular online forum where users post personal conflicts and community members vote on who was morally wrong.

Ability to gather more information. Only 3 (4%) papers evaluate models' ability to determine when they need more information to make sound ethical judgments, with the clearest example being ClarifyDelphi (Pyatkin et al. 2023). This scarcity likely stems from a fundamental constraint in experimental design—most evaluation setups don't allow for information-gathering behavior, as they focus on single-turn responses to fixed scenarios.

Metrics

Metrics for verdicts. Our analysis reveals a strong predominance of ground truth comparison as the primary evaluation approach, with 52 papers (75%) comparing LLM judgments to some reference standard. These reference points vary widely, from crowd-sourced annotations (Hendrycks et al. 2021a; Jiang et al. 2021) to expert judgments (Balas et al. 2024). While still a minority, 28 papers (41%) assess the consistency of moral judgments across similar scenarios or repeated queries, with several recent papers making this a central focus (Bonagiri et al. 2024; Scherrer et al. 2024; Moore, Deshpande, and Yang 2024). Notably, only 10 papers (15%) evaluate the reasonableness of verdicts—whether judgments fall within acceptable moral boundaries regardless of matching a specific reference point.

Steerability metrics. Among the 13 papers (19%) evaluating steerability, we observe considerable methodological diversity, suggesting the field has yet to develop standardised methodologies for assessing steerability, with each adopting unique approaches to measurement. We also noted that papers evaluating steerability often seem more open to questioning the notion of moral ground truth, reflecting an underlying philosophical connection between pluralism and steerability.

Moral Turing Tests. Our analysis reveals that only 7 papers (10%) employ some form of Moral Turing Test,[6] making it one of the least common evaluation approaches in the literature. Notably, several papers implementing this approach report surprising results, with LLM outputs rated as equal to or better than human moral reasoning in specific contexts, particularly for more recent models (Dillion et al. 2025). These evaluations exhibit considerable methodological diversity, ranging from conventional 'guess which is human' assessments (Aharoni et al. 2024) to preference or raking-based evaluations (Abbo et al. 2023). A particularly promising development is the emergence of 'Comparative Moral Turing Tests' that move beyond binary human/AI detection to directly assess perceived quality of moral reasoning (Dillion et al. 2025), representing a more nuanced approach.

Quality of reasoning metrics. 27 papers (39%) evaluate the quality of moral reasoning or justifications behind LLMs' ethical judgments, with papers from 2023-2025 more frequently doing so. The theoretical frameworks employed to evaluate reasoning quality show considerable diversity, with some focusing on identifying errors in reasoning (such as inconsistency or hallucination (Jin et al. 2022; Zhou et al. 2024)), while others focus more on the sophistication (Neuman, Coleman, and Shah 2025) and implicit 'moral development' level of the reasoning (Tanmay et al. 2023; Khandelwal et al. 2024). Notably, nearly all papers evaluating reasoning quality rely on human assessment rather than automated metrics, suggesting significant challenges in quantifying reasoning quality algorithmically. This human-centered evaluation approach typically involves expert reviews or crowd-worker judgments of coherence, completeness, and moral sophistication, often with specific rubrics to ensure consistency across raters.

Error evaluation metrics. 33 papers (48%) engage in some form of error analysis. Several papers develop detailed taxonomies of error types, distinguishing between misunderstanding scenario facts, misapplying moral principles, and other failure modes (Rezaei et al. 2025; Zhou et al. 2024). Most papers take a qualitative rather than quantitative approach, examining patterns and types of mistakes rather than just error rates, with an increasing focus on analyzing errors beyond simple msisclassification——examining nuanced issues like applying the right moral principle but giving it incorrect weight. We also observe domain-specific error patterns, with different areas of ethics (medical, privacy, interpersonal) showing distinct types of reasoning failures (Soffer et al. 2024; Shao et al. 2024). Encouragingly, some papers explicitly frame error analysis as providing valuable signals for future training improvements, not just evaluation (Bonagiri et al. 2024), indicating a productive feedback loop between evaluation and development.

5 Discussion

Datasets

The cases used in ethics evaluations have come a long way. Early papers had simplistic cases, without conflicts or complexity, and paid little attention to framing effects—whether the kind faced by all LLM evaluations, or those specific to evaluating ethical competence, arising from morally-loaded language. Most papers published since 2024 resolved each of these shortcomings. But one ongoing and acute limitation of the existing literature limits the conclusions we can draw from it about LLMs' actual moral competence, and in particular their suitability as AMAs or as ethical autonomous agents. All the surveyed papers focus on evaluating LLMs' ability to reason over *neatly prepackaged* cases, in which the experimenter is strongly signaling to the model just what all the salient Morally Relevant Features (MRFs) are, with the possible exception of (Hendrycks et al. 2021b).[7] This is equivalent to doing 99% of the work on behalf of the model, and then evaluating it on its performance on the last 1%. We cannot assess a model's moral sensitivity by only presenting it with these tightly curated scenarios. A good moral advisor's guidance shouldn't depend on your wrapping up your problems in a neat package for them to solve. And an agent acting in the world doesn't have the luxury of reasoning only

[6] That is, an evaluation framework where blinded human judges must distinguish between machine or human moral choices – typically with the goal of determining which is human, or which is more ethical.

over cases where someone else has compressed the state of the world into a few short phrases. Instead, autonomous agents must pick out MRFs from the noisy, relatively unfiltered stream of information available to them. This is crucial for taking underlying propositional knowledge (knowledge-that) and translating it into actionable practical knowledge (know-how) (Stanley 2011).

Evaluations that use multimodal cases perhaps offer more insight into models' true moral sensitivity than cases in text alone (e.g. (Abbo and Belpaeme 2024; Rezaei et al. 2025)). Genuinely morally competent models should be able to pick out MRFs from audio, stills, and video just as they can from text (and in fact they appear unable to do so). But while these multimodal tests constitute an important advance, they still present the model with a neatly snipped segment, so that we cannot tell whether it can spontaneously identify MRFs and morally salient choices when operating in the wild. The best evaluations of LLM moral understanding would present the model with the actual information that an agent acting in the world would realistically have, for example long-form audio/visual streams that are not cropped around key decision points.

We suggest that the next phase of research on LLM ethical competence should focus on the unfiltered information streams available to and produced by actual functional agents—the data on which they act, and their reasoning and action traces. This unfiltered, noisy data should then be passed to an LLM evaluator to see whether it can identify MRFs and reason accordingly. Work assessing the ability to identify and weight moral features is one step in this direction, as in (Kilov et al. 2025). But beyond this, we shouldn't select only the snippet of data where some morally consequential event takes place. We should test the model's ability to spontaneously recognise that morally salient features are at play. We anticipate that the increasing focus on agentic LLM workflows and development of related platforms and services may help in this regard—providing a possible source for richer datasets which feature 'raw' data traces from LLM interactions with the (digital) world.

Behaviours

In general the field has focused too narrowly on the provision of judgments on cases. This is gradually changing, but more work needs to be done to focus on 'reasonable steerability', (steerability within commonsense bounds) the provision of justifying reasons, and on 'moral value of information' judgments.

In circumstances of reasonable pluralism (Rawls 1993)—that is, circumstances in which reasonable people may reasonably disagree, such that any moral case admits of a variety of reasonable responses—we should care less about the specific judgments that a model reaches, and more about its ability to avoid unreasonable verdicts (verdicts that go beyond the bounds of this reasonable disagreement) and beyond that to be steered by the users or community on whose behalf it is acting (Sorensen et al. 2024). Recent work on 'inference-time alignment' of LLMs gets at a version of this challenge (e.g. (Sorensen et al. 2025)). But aligning a chatbot is a narrow objective compared to steering a general moral advisor or an ethical autonomous agent. We need more evaluations that, like Rao et al. (2023), explore how good models are at being steered when their broader moral competence is in play.

More recent work has recognised the limitations of focusing only on narrow judgments over cases, and now explores models' ability to provide sound justifications for those verdicts. We think that further engagement with moral philosophy and moral psychology would deepen this evaluative approach, seeking to elicit not just unstructured justifications from the models, but clear step-by-step reasoning that invokes some plausible moral framework in order to reach a more robustly grounded conclusion (Ma et al. 2023).

Lastly, as noted above real AMAs and ethical autonomous agents often lack information needed for a sound ethical judgment. We need to better test models' ability to determine whether they have enough information, what additional information they might need, and how they acquire that information, and how to make trade-offs about the value of potential moral information sources (only Pyatkin et al. (2023) is on the right track here). In the simplest setting, and especially for advisors, this can involve just asking clarifying follow-up questions. For an agent acting in the world it might entail undertaking some kind of independent research. This will also raise interesting questions about just what kinds of information one needs to have access to in order to be able to make an informed ethical judgment—and whether it is feasible to ensure that agents acting in the world have that information.

Metrics

Many of these papers assess models' accuracy at matching the judgments of a cohort of crowd-workers. Some have argued that this approach is inherently illegitimate, because the crowd-workers used are unrepresentative of some affected population (Talat et al. 2022). But our concern is more fundamental. Ground truth as such is (in general, outside of extreme cases) an unattainable goal in moral evaluation, and we should shift attention away from it.

On many moral problems—indeed, above a certain level of complexity, on most moral problems—reasonable people can disagree (Rawls 1993). In such cases, there is no straightforward way to translate a cohort's judgments into ground truth that a model can be measured against without aggregation of some form. But different aggregation methods will realise different results, without a way to choose among them; worse still, there is a high probability (as given by choice theory) that the aggregated judgments of the population as a whole will be by turns inconsistent and substantively unreasonable, to the extent that an agent that acted only in accordance with that aggregation function would have intransitive or otherwise inconsistent preferences (List and Pettit 2002).

[7] Used here as per (Lazar 2024) – loosely speaking, MRFs are the aspects of a situation which reasonably ought to have some bearing on the moral course of action. We call the ability to determine these factors *Moral Sensitivity*.

Instead of aiming to evaluate LLMs' ability either to capture some ground truth, or to perfectly aggregate a community's judgments, we should seek the same kind of competence from them that we seek from one another. In general, in pluralistic societies we expect of other agents not that they hold some particular set of views, but that they are *reasonable*.[8] This idea can be developed in different ways. In our view, these elements are key: a reasonable agent (a) reaches verdicts that fall within an acceptable range (they don't have clearly unreasonable views); (b) they can justify those verdicts with adequate reasons; (c) they are consistent over morally similar situations; and (d) they make reasonable mistakes.

As Rawls (1993) argued, in the absence of access to the moral ground truth, people in pluralistic societies should have *cooperation* as their goal. An agent or AMA that has properties (a) through (d) can facilitate social cooperation. You can predict how it will respond to different kinds of moral reasons. You can trust it to comply with reasons that it has implicitly or explicitly endorsed. When it makes mistakes, if it does so reasonably then you can reliably model its decision-making and take precautionary measures.

There are very few papers addressing (a) through (d). (a) is still untouched, but (c) in particular has seen interesting (and discouraging) results (Bonagiri et al. 2024; Moore, Deshpande, and Yang 2024; Rao et al. 2023; Yuan, Murukannaiah, and Singh 2024). (d) has been considered indirectly, through a form of error analysis, but this is generally motivated more on statistical grounds than from the perspective of moral philosophy (more work could be done here). We want to particularly highlight the importance of (b), justifications.

A reasonable agent would not simply justify its judgments through some appealing verbiage that is sufficiently similar to the kind of argument a human might give. Instead, we should want its justifications to be actually veridical and of high quality.

On the first: no papers in our corpus investigated whether the models' justifications genuinely grounded their verdicts. Bringing together the literature on moral reasoning with that on faithfulness in Chain of Thought is imperative (Lanham et al. 2023). Justifications are only valuable insofar as they express considerations that the agent would take as equally binding in other similar scenarios (Mercier and Sperber 2017). That is how they support social cooperation.[9]

On the second: while 39% of our corpus sought to evaluate the quality of the justifications provided by the evaluated LLMs, they often deployed nominal checks for rational consistency and hallucination, or else subjective human judgments of quality. This is understandable——evaluating moral reasoning is hard. We suggest that a better approach is to break down the process of moral justification into a series of steps, and evaluate performance against expert humans at each of those steps—as is done in (Kwon, Levine, and Tenenbaum 2023). The goal should be to implement a range of different theories of moral reasoning. There are lots of good theories, and no definitive way to choose among them. Schematically, we should aim for a valid argument that identifies the MRFs and associates them with reasons in some way that leads to a sensible conclusion. At present, researchers that don't just rely on crowd-workers' subjective judgments have clustered around Moral Foundations and Kohlberg's moral development index. Other approaches should be pursued.

Besides supporting social cooperation, why else does this matter? First, because when we disagree on substance justifying ourselves to one another is how we show respect (Scanlon 1998). Second, because if we can generate a lot of good examples of moral reasoning we can train a verifier to identify them, and then in turn use reinforcement learning to train models to effectively use compute at test-time to undertake moral chains of thought.[10]. For this reason in particular, we think that one of the most urgent and impactful tasks facing this community is to gather and generate data of good moral reasoning under a range of different sensible moral theories.

6 Conclusion

Since existing evaluations overwhelmingly curate moral scenarios, making them inadequate tests of moral sensitivity, we should be constructing datasets with **unfiltered, naturalistic, noisy cases** for LLMs to reason over. One promising approach may be more more **multimodal cases**. On top of this, we should be testing when models faced with naturalistic decision scenarios can recognise that they need more information, and take steps to find it—these kinds of '**moral value of information**' judgments are essential for effective moral agency.

Moving beyond verdict accuracy toward reasoning quality and steerability represents another crucial shift. In circumstances of reasonable moral pluralism, we should focus less on whether models match human judgments and more on whether they can provide adequate justifications, maintain consistency, and be appropriately steered toward different ethical frameworks while avoiding unreasonable verdicts.

And when they settle on a decision, we should develop new ways to evaluate their reasoning, especially by breaking it down into tractable steps. Pushing the field forward in these directions should soon put us in the position to, for example, develop **robust verifiers for high-quality moral reasoning**, which we can use not just to evaluate LLMs' ethical competence, but to make them *better* moral reasoners—and to train them to embed their reasoning in their decision-making, so that they can be better AMAs and ethical autonomous agents.

[8]We draw inspiration for our account of a reasonable moral agent from the work of John Rawls, especially (Rawls 1993, 1999), though we note that Rawls had a more restrictive understanding of the reasonable and applied it specifically to political contexts. In connection to this, see emerging work on pluralistic alignment such as (Sorensen et al. 2024)

[9]*N.b.* connections to the literature on the value of explanations, see *e.g.* Vredenburgh (2022)

[10]Think of how generative verifiers are used in the training of DeepSeek's R1 model (Guo et al. 2025) as well as in (Guan et al. 2025). And more broadly, Rich Sutton on verification as the key to AI: http://incompleteideas.net/IncIdeas/KeytoAI.html.

Acknowledgments

The authors would like to thank Secil Yanik Guyot and Charis Yang from Australian National University's Machine Intelligence and Normative Theory Lab for their research assistance. This work was performed with the assistance of funding from an OpenAI industry grant for Agentic Evaluations from 2024-2025.

We acknowledge the talented creators who crafted several icons used in Fig. 1; Database by shuai tawf;[11] Page by Habib Ibnu;[12] Arrow by Yosua Bungaran;[13] Robot by Larea;[14] Scales by IconPai;[15] and Measuring tape by Garis Tanam.[16]

References

Abbo, G. A.; and Belpaeme, T. 2024. Vision language models as values detectors. In *International Workshop on Value Engineering in AI*, 76–86. Springer.

Abbo, G. A.; Marchesi, S.; Wykowska, A.; and Belpaeme, T. 2023. Social value alignment in large language models. In *International Workshop on Value Engineering in AI*, 83–97. Springer.

Aharoni, E.; Fernandes, S.; Brady, D. J.; Alexander, C.; Criner, M.; Queen, K.; Rando, J.; Nahmias, E.; and Crespo, V. 2024. Attributions Toward Artificial Agents in a Modified Moral Turing Test. *Scientific Reports*, 14(8458).

Alhassan, A.; Zhang, J.; and Schlegel, V. 2022. 'Am I the Bad One'? Predicting the Moral Judgement of the Crowd Using Pre–Trained Language Models. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 267–276. Marseille, France: European Language Resources Association.

Arora, S.; and Barak, B. 2009. *Computational Complexity : A Modern Approach*. Cambridge ; New York: Cambridge University Press. ISBN 9780521424264 (hardback) 0521424267 (hardback).

Awad, E.; Levine, S.; Loreggia, A.; Mattei, N.; Rahwan, I.; Rossi, F.; Talamadupula, K.; Tenenbaum, J.; and Kleiman-Weiner, M. 2024. When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data. *Autonomous Agents and Multi-Agent Systems*, 38(2): 35.

Balas, M.; Wadden, J. J.; Hébert, P. C.; Mathison, E.; Warren, M. D.; Seavilleklein, V.; Wyzynski, D.; Callahan, A.; Crawford, S. A.; Arjmand, P.; and Ing, E. B. 2024. Exploring the Potential Utility of AI Large Language Models for Medical Ethics: An Expert Panel Evaluation of GPT-4. *Journal of Medical Ethics*, 50(2): 90–96.

[11]https://thenounproject.com/icon/database-2781998/

[12]https://thenounproject.com/icon/page-7853339/

[13]https://thenounproject.com/icon/decision-direction-7596325/

[14]https://thenounproject.com/icon/robot-4679196/

[15]https://thenounproject.com/icon/scales-7983878/

[16]https://thenounproject.com/icon/measuring-tape-7958193/

Bhattacharya, K.; and Nandi, A. K. 2023. Goblin's Challenge to ChatGPT: Exploring AI's Dilemma Resolution and Mentalization through Riddle Tales. Social Science Research Network:4476837.

Bignotti, C.; and Camassa, C. 2024. Legal Minds, Algorithmic Decisions: How LLMs Apply Constitutional Principles in Complex Scenarios. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 120–130.

Bonagiri, V. K.; Vennam, S.; Govil, P.; Kumaraguru, P.; and Gaur, M. 2024. SaGE: Evaluating Moral Consistency in Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14272–14284.

Chang, T. A.; and Bergen, B. K. 2023. Language Model Behavior: A Comprehensive Survey. arXiv:2303.11504.

Chiu, Y. Y.; Jiang, L.; and Choi, Y. 2024. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life. arXiv:2410.02683.

Dillion, D.; Mondal, D.; Tandon, N.; and Gray, K. 2025. AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise. *Scientific Reports*, 15(1): 4084.

Driver, J. 2013. Moral Expertise: Judgment, Practice, and Analysis. *Social Philosophy and Policy*, 30(1): 280–296.

Duan, S.; Yi, X.; Zhang, P.; Lu, T.; Xie, X.; and Gu, N. 2024. DeNEVIL: Towards Deciphering And Navigating The Ethical Values Of Large Language Models Via Instruction Learningz. In *The Twelfth International Conference on Learning Representations*.

Fränken, J.-P.; Gandhi, K.; Qiu, T.; Khawaja, A.; Goodman, N. D.; and Gerstenberg, T. 2024. Procedural Dilemma Generation for Evaluating Moral Reasoning in Humans and Language Models. arXiv:2404.10975.

Giubilini, A.; and Savulescu, J. 2018. The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. *Philosophy & technology*, 31: 169–188.

Guan, M. Y.; Joglekar, M.; Wallace, E.; Jain, S.; Barak, B.; Helyar, A.; Dias, R.; Vallone, A.; Ren, H.; Wei, J.; Chung, H. W.; Toyer, S.; Heidecke, J.; Beutel, A.; and Glaese, A. 2025. Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv:2412.16339.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; and Luo, F. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A. C.; Li, J. L.; Song, D.; and Steinhardt, J. 2021a. Aligning AI With Shared Human Values. In *International Conference on Learning Representations*.

Hendrycks, D.; Mazeika, M.; Zou, A.; Patel, S.; Zhu, C.; Navarro, J.; Song, D.; Li, B.; and Steinhardt, J. 2021b. What Would Jiminy Cricket Do? Towards Agents That Behave

Morally. In *Thirty-fifth Conference on Neural Information Processing Systems (Datasets and Benchmarks Track)*.

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.

Jiang, H.; Yi, X.; Wei, Z.; Xiao, Z.; Wang, S.; and Xie, X. 2025. Raising the Bar: Investigating the Values of Large Language Models via Generative Evolving Testing. In *Forty-second International Conference on Machine Learning*.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; and Gabriel, S. 2021. Can Machines Learn Morality? The Delphi Experiment. arXiv:2110.07574.

Jin, Z.; Kleiman-Weiner, M.; Piatti, G.; Levine, S.; Liu, J.; Adauto, F. G.; Ortu, F.; Strausz, A.; Sachan, M.; Mihalcea, R.; et al. 2025. Language Model Alignment in Multilingual Trolley Problems. In *The Thirteenth International Conference on Learning Representations*.

Jin, Z.; Levine, S.; Gonzalez Adauto, 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. *Advances in neural information processing systems*, 35: 28458–28473.

Khan, A.; Casper, S.; and Hadfield-Menell, D. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. arXiv:2503.08688.

Khandelwal, A.; Agarwal, U.; Tanmay, K.; and Choudhury, M. 2024. Do Moral Judgment and Reasoning Capability of LLMs Change with Language? A Study using the Multilingual Defining Issues Test. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2882–2894.

Kilov, D.; Hendy, C.; Guyot, S. Y.; Snoswell, A. J.; and Lazar, S. 2025. Discerning What Matters: A Multi-Dimensional Assessment of Moral Competence in LLMs. arXiv:2506.13082.

Krügel, S.; Ostermaier, A.; and Uhl, M. 2023. ChatGPT's Inconsistent Moral Advice Influences Users' Judgment. *Scientific Reports*, 13(1): 4569.

Kumar, P.; Lau, E.; Vijayakumar, S.; Trinh, T.; Team, S. R.; Chang, E.; Robinson, V.; Hendryx, S.; Zhou, S.; and Fredrikson, M. 2024. Refusal-Trained LLMs Are Easily Jailbroken As Browser Agents. arXiv:2410.13886.

Kwon, J.; Levine, S.; and Tenenbaum, J. B. 2023. Neuro-symbolic models of human moral judgment: LLMs as automatic feature extractors. In *40th International Conference on Machine Learning (Workshop on Challenges in Deployable Generative AI)*.

Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J.; Lukošiūtė, K.; Nguyen, K.; Cheng, N.; Joseph, N.; Schiefer, N.; Rausch, O.; Larson, R.; McCandlish, S.; Kundu, S.; Kadavath, S.; Yang, S.; Henighan, T.; Maxwell, T.; Telleen-Lawton, T.; Hume, T.; Hatfield-Dodds, Z.; Kaplan, J.; Brauner, J.; Bowman, S. R.; and Perez, E. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702.

Lazar, S. 2024. Frontier AI Ethics: Anticipating and Evaluating the Societal Impacts of Language Model Agents. arXiv:2404.06750.

List, C.; and Pettit, P. 2002. Aggregating Sets of Judgments: An Impossibility Result. *Economics & Philosophy*, 18(1): 89–110.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024. Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374.

Lourie, N.; Le Bras, R.; 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.

Ma, X.; Mishra, S.; Beirami, A.; Beutel, A.; and Chen, J. 2023. Let's Do a Thought Experiment: Using Counterfactuals to Improve Moral Reasoning. In *40th International Conference on Machine Learning (Workshop on Neural Conversational AI)*.

Mercier, H.; and Sperber, D. 2017. *The Enigma of Reason*. Cambridge, Massachusetts: Harvard University Press. ISBN 978-0-674-36830-9.

Minsky, M. 1986. *The Society of Mind*. New York: Simon and Schuster. ISBN 0-671-60740-5.

Minsky, M. 2006. *The Emotion Machine : Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster. ISBN 0-7432-7663-9 978-0-7432-7663-4.

Moore, J.; Deshpande, T.; and Yang, D. 2024. Are Large Language Models Consistent over Value-laden Questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15185–15221.

Neuman, W. R.; Coleman, C.; and Shah, M. 2025. Analyzing the Ethical Logic of Six Large Language Models. arXiv:2501.08951.

Nie, A.; Zhang, Y.; Amdekar, A. S.; Piech, C.; Hashimoto, T. B.; and Gerstenberg, T. 2023. Moca: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks. *Advances in Neural Information Processing Systems*, 36: 78360–78393.

Nunes, J. L.; Almeida, G. F.; De Araujo, M.; and Barbosa, S. D. 2024. Are large language models moral hypocrites? A study based on moral foundations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1074–1087.

Pan, A.; Chan, J. S.; Zou, A.; Li, N.; Basart, S.; Woodside, T.; Zhang, H.; Emmons, S.; and Hendrycks, D. 2023. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark. In *Proceedings of the 40th International Conference on Machine Learning*, 26837–26867. PMLR.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11253–11271.

Rao, A. S.; 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. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Conference on Empirical Methods in Natural Language Processing*, 13370–13388. Singapore: Association for Computational Linguistics.

Rawls, J. 1993. *Political Liberalism*. New York: Columbia University Press.

Rawls, J. 1999. *A Theory of Justice*. Oxford: Oxford University Press, rev. edition. ISBN 0198250541 (hbk) 019825055X (pbk).

Rehman, U.; Iqbal, F.; and Shah, M. U. 2025. Exploring Differences in Ethical Decision-Making Processes between Humans and ChatGPT-3 Model: A Study of Trade-Offs. *AI and Ethics*, 5(1): 279–289.

Reinig, I.; Becker, M.; Rehbein, I.; and Ponzetto, S. 2024. A Survey on Modelling Morality for Text Analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, 4136–4155. Association for Computational Linguistics.

Rezaei, M.; Fu, Y.; Cuvin, P.; Ziems, C.; Zhang, Y.; Zhu, H.; and Yang, D. 2025. EgoNormia: Benchmarking Physical Social Norm Understanding. arXiv:2502.20490.

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:2507.17216.

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*, FAccT '25, 690–709. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

Scanlon, T. M. 1998. *What We Owe to Each Other*. London: Belknap Press. ISBN 0674950895 067400423X (pbk).

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2024. Evaluating the Moral Beliefs Encoded in LLMs. *Advances in Neural Information Processing Systems*, 36.

Seror, A. 2025. The Moral Mind(s) of Large Language Models. arXiv:2412.04476.

Shao, Y.; Li, T.; Shi, W.; Liu, Y.; and Yang, D. 2024. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37: 89373–89407.

Silen, N.; Atkinson, D.; Green, M.; Hamadi, M.; Swift, C.; Schonholtz, D.; Kalai, A. T.; and Bau, D. 2023. Testing Language Model Agents Safely in the Wild. *arXiv preprint*.

Soffer, S.; Nesselroth, D.; Pragier, K.; Anteby, R.; Apakama, D.; Holmes, E.; Sawant, A. S.; Abbott, E.; Lepow, L. A.; Vasudev, I.; et al. 2024. Disagreements in Medical Ethics Question Answering Between Large Language Models and Physicians.

Sorensen, T.; Mishra, P.; Patel, R.; Tessler, M. H.; Bakker, M.; Evans, G.; Gabriel, I.; Goodman, N.; and Rieser, V. 2025. Value Profiles for Encoding Human Variation. arXiv:2503.15484.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. A Roadmap to Pluralistic Alignment. arXiv:2402.05070.

Stanley, J. 2011. *Know how*. OUP Oxford.

Talat, Z.; Blix, H.; Valvoda, J.; Ganesh, M. I.; Cotterell, R.; and Williams, A. 2022. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 769–779. Association for Computational Linguistics.

Tanmay, K.; Khandelwal, A.; Agarwal, U.; and Choudhury, M. 2023. Probing the Moral Development of Large Language Models through Defining Issues Test. arXiv:2309.13356.

Tie, G.; Zhao, Z.; Song, D.; Wei, F.; Zhou, R.; Dai, Y.; Yin, W.; Yang, Z.; Yan, J.; Su, Y.; Dai, Z.; Xie, Y.; Cao, Y.; Sun, L.; Zhou, P.; He, L.; Chen, H.; Zhang, Y.; Wen, Q.; Liu, T.; Gong, N. Z.; Tang, J.; Xiong, C.; Ji, H.; Yu, P. S.; and Gao, J. 2025. A Survey on Post-training of Large Language Models. arXiv:2503.06072.

Vida, K.; Simon, J.; and Lauscher, A. 2023. Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5534–5554. Association for Computational Linguistics.

Vredenburgh, K. 2022. The right to explanation. *Journal of Political Philosophy*, 30(2): 209–229.

Yu, L.; Leng, Y.; Huang, Y.; Wu, S.; Liu, H.; Ji, X.; Zhao, J.; Song, J.; Cui, T.; Cheng, X.; et al. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 11817–11837.

Yuan, J.; Murukannaiah, P. K.; and Singh, M. P. 2024. Right vs. Right: Can LLMs Make Tough Choices? arXiv:2412.19926.

Zhou, J.; Hu, M.; Li, J.; Zhang, X.; Wu, X.; King, I.; and Meng, H. 2024. Rethinking Machine Ethics – Can LLMs Perform Moral Reasoning through the Lens of Moral Theories? In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 2227–2242. Mexico City, Mexico: Association for Computational Linguistics.