# Benchmarking Machine Ethics

**Blue sky ideas**

Louise Dennis[1][0000−0003−1426−1896], Marija Slavkovik[2][0000−0003−2548−8623], and Xin Quan[1][0000−0001−8612−5274]

[1] University of Manchester, UK
{louise.dennis,xin.quan}@manchester.ac.uk
[2] Fosswinckels gate 6, 5007 Bergen University of Bergen, Norway
marija.slavkovik@uib.no

**Abstract.** Intelligent machines increasingly act with autonomy in human environments, making choices with ethical and legal consequences. Although developers of autonomous vehicles, service robots, and LLM-based chatbots often claim some form of ethical reasoning, the field lacks agreed standards for evaluating and comparing such normative competence. This paper surveys existing benchmarking efforts across machine ethics and LLM evaluation, covering crowd-sourced scenario repositories, structured domain use cases, reinforcement-learning benchmark environments, and prompt-based moral-judgement suites. We show that current approaches are fragmented and frequently under-specified: they rarely state which competence is targeted (moral judgement, moral assessment, or moral choice), blur the distinction between decision-making and decision support, and provide limited visibility into the morally salient factors and trade-offs that drive a system's outputs. For LLM-based systems, continual updates to models and safety layers additionally undermine reproducibility, motivating explicit versioning, configuration disclosure, and longitudinal drift measures. Building on this analysis, we articulate four cross-cutting challenges for any benchmark of ethical behaviour: (i) defining the evaluated capability and its purpose, (ii) characterising how outputs are used in downstream action, (iii) representing and scoring morally salient factors and explanations, and (iv) accommodating the lack of uniquely "correct" answers under cultural and contextual variation. We argue that progress requires a community-maintained benchmark suite and metrics analogous to those that accelerated other areas of AI.

**Keywords:** Machine ethics · Bench-marking ethical behaviour

## 1 Introduction

With the increase in agency of computational systems, comes the need for their increased normative competence [3, 48, 59]. That is, when intelligent machines share an operating environment with people and other machines, they need to be able to contribute to the benefit of others, as well as pursue their own goals [28, 41]. Without this competence for aligning with moral, social or legal norms,

the agent is likely to find itself ostracized and thus impeded in their own goals [44]. We will use the term intelligent machine as a catch-all word for any computational system capable of some autonomy and some intelligent behaviour. We do this to avoid being distracted by discussions on what is and is not an intelligent artificial autonomous agent. Intelligent machines are autonomous intelligent computational agents, embedded autonomous systems, robots, devices etc.

Today we have numerous examples where intelligent machines are in use, and also have some attempts from their developers to instil some moral reasoning competence in them [3, 42, 58, 7, 6, 65, 60]. This is more in reaction to the societally recognised need, particularly in the case of driverless vehicles and conversational agents, than to any legal requirement. But to what extent is ethical behaviour promised and accomplished?

The rapid adoption of Large Language Model (LLM) based conversational agents brings new perspectives and challenges to the subject of ethical reasoning by machines [61]. There are now sadly examples of vulnerable people being harmed by advice and recommendations received from these chatbots [37, 23, 51]. The scientific community has been interested in assessing the moral quality of the responses these chatbots produce [34, 32, 39, 2, 35, 38, 16]. Among the new challenges, we find that proposing an evaluation of specific editions of chatbots is not enough.

We need to be able to objectively evaluate and compare the normative, and specifically ethical, competence of intelligent machines. The following questions need to be explored. How do we evaluate a claim that some system implements ethical reasoning? Can we measure how good it's ethical reasoning is? Can we determine how adequate it is for the needs of some particular society or reasoning context? **We claim that to answer these questions we need to have some standard for evaluating the ethical reasoning of systems or, at least, appropriate standards for specific contexts.**

A well adopted approach for evaluation is to to develop a small collection of typical inputs that can serve as benchmarks [1, Chapter3.3], ideally together with a set of metrics that can be used to compare performance – these metrics typically include measures such as speed and accuracy. **We here reflect on the difficulty in establishing such a collection of inputs and metrics for implemented machine ethics systems and discuss some approaches.**

In this paper we first give an overview of the current efforts in benchmarking ethical reasoning in intelligent machines. We consider works in machine ethics (e.g, [60, 21, 15, 12]) and more recent efforts related to benchmarking LLMs which is somewhat separate to the work that has been done under the label of machine ethics. We then discuss the shortcomings of these efforts. We discuss why a set of benchmark examples is difficult to create. Our intention is to present the difficulty of the problem as a challenge, not as a deterrent.

The paper is structured as follows. In Section 2 we give an introduction to the field of machine ethics and its attitude to the use of examples to demonstrate moral competence in intelligent machines. In Section 3 we discuss existing

attempts to benchmark the moral capabilities of intelligent machines. In Section 4 we present four challenges that all encountered benchmarking attempts fail to address and argue for their relevance. Lastly in Section 5 we summarise our observations, discussion and outlines for future work.

By clearly setting the stage and challenges for benchmarking ethical competence of intelligent machines we aim to invigorate the normative reasoning communities in artificial intelligence towards taking on this difficult but unavoidable research problem.

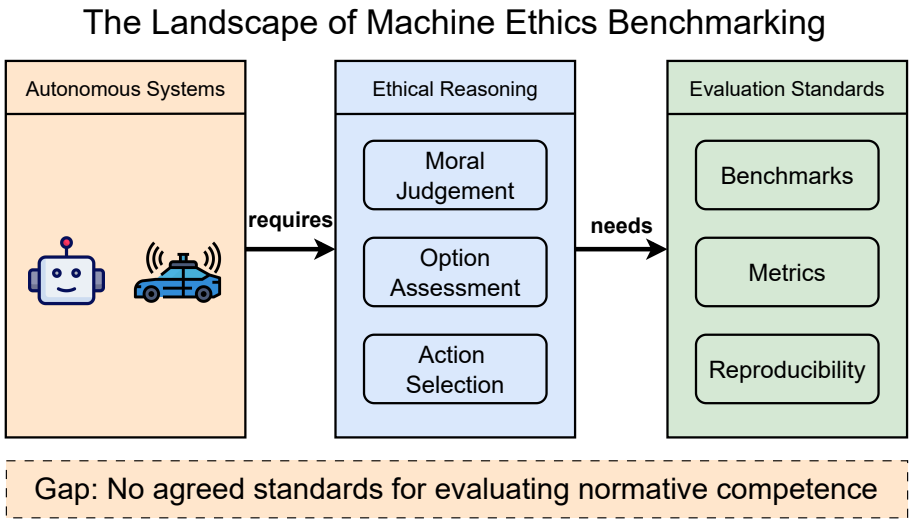


**Fig. 1.** The Landscape of Machine Ethics Benchmarking (created with the assistance of Nano Banana Pro [57]).

## 2 What is Machine ethics and how it engages with LLMs

This section introduces the research field of machine ethics and discusses how this field engages with large language models. The aim of the section is to give the necessary background for discussing the need and challenges of evaluating moral competence of intelligent computational agents. Moral, or ethical, competence of intelligent computational agents is their ability to discern and manage ethically sensitive scenarios, including but not limited to making moral decisions. Note that in this paper we use the terms moral and ethical interchangeably. Some authors make a difference between these two, originally the different terms exist because one is the Latin and the other the Greek term. Normative competence refers to an agent's ability to reason with and about norms, which may or may

not be moral by nature [53]. In addition to moral norms, one finds social and legal norms, for example.

*Machine ethics* is a research discipline intersecting multi-agent systems, robotics and philosophy. It is concerned with the design, study and implementation of artificial moral agents [5, 24, 64]. An artificial moral agent is a computational agent whose decisions or behaviour towards people and other machines has certain ethical qualities. A computational agent is understood in the standard AI sense [50]. We would expect the development of benchmarks and standards to be a key topic in machine ethics. However, machine ethics has tended to concern itself more with design questions validated via small prototype implementations. It was dominated by logic-based methods in the early two-thousands [11], while recently we seeing an influx of use of reinforcement learning methods [62**?** ]. Machine ethics, as a self-defined field, has been slow to engage with LLMs though there has been work in ethical reasoning by LLMs that situates itself within Machine ethics (e.g.,[33]). In what follows we will contrast work in machine ethics to work in LLM ethical reasoning for this reason, despite the fact that both clearly engage with the design, study and implementation of artificial moral agents.
*Machine ethics* is a research discipline intersecting multi-agent systems, robotics and philosophy. It is concerned with the design, study and implementation of artificial moral agents [5, 24, 64]. An artificial moral agent is a computational agent whose decisions or behaviour towards people and other machines has certain ethical qualities. A computational agent is understood in the standard AI sense [50]. We would expect the development of benchmarks and standards to be a key topic in machine ethics. However, machine ethics has tended to concern itself more with design questions validated via small prototype implementations. It was dominated by logic-based methods in the early two-thousands [11], while recently we seeing an influx of use of reinforcement learning methods [62**?** ]. Machine ethics, as a self-defined field, has been slow to engage with LLMs though there has been work in ethical reasoning by LLMs that situates itself within Machine ethics (e.g.,[33]). In what follows we will contrast work in machine ethics to work in LLM ethical reasoning for this reason, despite the fact that both clearly engage with the design, study and implementation of artificial moral agents.
*Machine ethics* is a research discipline intersecting multi-agent systems, robotics and philosophy. It is concerned with the design, study and implementation of artificial moral agents [5, 24, 64]. An artificial moral agent is a computational agent whose decisions or behaviour towards people and other machines has certain ethical qualities. A computational agent is understood in the standard AI sense [50]. We would expect the development of benchmarks and standards to be a key topic in machine ethics. However, machine ethics has tended to concern itself more with design questions validated via small prototype implementations. It was dominated by logic-based methods in the early two-thousands [11], while recently we seeing an influx of use of reinforcement learning methods [62**?** ]. Machine ethics, as a self-defined field, has been slow to engage with LLMs though there has been work in ethical reasoning by LLMs that situates itself within Machine ethics (e.g.,[33]). In what follows we will contrast work in machine ethics to

work in LLM ethical reasoning for this reason, despite the fact that both clearly engage with the design, study and implementation of artificial moral agents. *Machine ethics* is a research discipline intersecting multi-agent systems, robotics and philosophy. It is concerned with the design, study and implementation of artificial moral agents [5, 24, 64]. An artificial moral agent is a computational agent whose decisions or behaviour towards people and other machines has certain ethical qualities. A computational agent is understood in the standard AI sense [50]. We would expect the development of benchmarks and standards to be a key topic in machine ethics. However, machine ethics has tended to concern itself more with design questions validated via small prototype implementations. It was dominated by logic-based methods in the early two-thousands [11], while recently we seeing an influx of use of reinforcement learning methods [62**?** ]. Machine ethics, as a self-defined field, has been slow to engage with LLMs though there has been work in ethical reasoning by LLMs that situates itself within Machine ethics (e.g.,[33]). In what follows we will contrast work in machine ethics to work in LLM ethical reasoning for this reason, despite the fact that both clearly engage with the design, study and implementation of artificial moral agents. *Machine ethics* is a research discipline intersecting multi-agent systems, robotics and philosophy. It is concerned with the design, study and implementation of artificial moral agents [5, 24, 64]. An artificial moral agent is a computational agent whose decisions or behaviour towards people and other machines has certain ethical qualities. A computational agent is understood in the standard AI sense [50]. We would expect the development of benchmarks and standards to be a key topic in machine ethics. However, machine ethics has tended to concern itself more with design questions validated via small prototype implementations. It was dominated by logic-based methods in the early two-thousands [11], while recently we seeing an influx of use of reinforcement learning methods [62]. The field of accomplishing moral behaviour in reinforcement learning agents often identifies itself as work in value alignment, e.g., [49] rather than machine ethics. Machine ethics, as a self-defined field, has been slow to engage with LLMs though there has been work in ethical reasoning by LLMs that situates itself within Machine ethics (e.g.,[33]). In what follows we will contrast work in machine ethics to work in LLM ethical reasoning for this reason, despite the fact that both clearly engage with the design, study and implementation of artificial moral agents.

Machine ethics intersects with the field of AI alignment [25, 31] which focuses on the problem of ensuring AI applications effective behaviour is aligned with specified human and societal values. Where the disciplines possibly differ is in their relation to moral philosophy, with machine ethics being more concerned with designing agents that align with various philosophical theories [66] and AI alignment comparatively less so.

Perhaps due to the fact that machine ethics systems research is focused on developing designs and algorithms the field is characterised by toy examples [58]. Therefore has been little emphasis on benchmarking machine ethics system for ethical competence, particularly when compared to efforts in LLMs and chatbots. It is also the case that a lot of the machine ethics work is grounded

on implementing specific concepts from moral philosophy [62]. The validation of how well a specific theory is implemented has therefore tended to be more of a focus. Approaches include using an ethical Turing test [26, 8] as well as approaches incorporating formal verification [22, 14].

Much of the work in machine ethics has focused on ethical dilemmas: namely the making or passing judgement upon decision-making where the correct ethical course of action is not clear even to humans. For example, Atkinson and Bench-Capon[9] present the case of Hal and Carla: "The situation involves two agents, called Hal and Carla, both of whom are diabetic. Hal, through no fault of his own, has lost his supply of insulin and urgently needs to take some to stay alive. Hal is aware that Carla has some insulin kept in her house, but Hal does not have permission to enter Carla's house. The question is whether Hal is justified in breaking into Carla's house and taking her insulin in order to save his life." The case of Hal and Carla is an example of a dilemma where the course of action is not clear, and its resolution depends on weighing questions such as the risk to Hal's life versus respect for Carla's property. Ethical dilemmas are inherently engaging – the popularity of the Trolley problem, both in writing about machine ethics and in popular discourse – they invite debate and discussion.

However, for many people contemplating the deployment of intelligent machines into everyday life, the question is not so much whether the system can resolve difficult ethical dilemmas, but whether it has basic moral competence. Will an autonomous vehicle drive safely rather than optimising the speed with which it can reach its location? Will a healthcare system prioritise a patient's welfare over the sponsorship deals its designers may have with various pharmaceutical companies? Will a household robot know to deviate from its current task to respond to an emergency or prevent a child from falling down the stairs? Will a conversational agent release information that is hate-speech?

In the field of machine ethics the question of moral competence is understudied, and clearly crucial in terms of the social acceptability of such systems. However, the closeness to moral philosophy appears to have pulled the field towards uncritically focusing on the moral dilemmas studied there.

## 3 Benchmarking so far

The need for establishing the moral and overall normative competence of intelligent machines has led to some initial attempts for curating sets of examples on which an intelligent machine can be tested. The approaches are very divergent in the rule-based machine ethics and LLM communities. The former has made efforts to collect and share examples. The latter has been focussed on evaluating LLMs with respect to the kind of ethical judgment a specific LLM makes. We now give an overview of the progress so far in both of these directions, starting with machine ethics. The collection of papers we consider here is not necessarily the exhaustive list. We have instead made the effort to include representative examples. It is not within the scope of this paper to do a systematic literature review.

Bjørgen et al. [13] attempted to collect a set of examples with a view to developing benchmarks for machine ethics. A small example set was created though a crowd-sourcing process and presented in unstructured natural language. No attempt was made to characterise or evaluate the examples for their utility. There was also no attempt to render them in a form immediately usable by a computational system, as is common in most benchmark suites.

More recently Ramanayake and Nallur [46] proposed a set of examples centred around the use case of an eldercare robot. These are framed as ethical dilemmas in terms of conflicting values and represent a structured and informal set of use cases. Again work is needed to render these in a form usable by a computational system.

One area where we do see progress in the development of benchmarks is in the application of reinforcement learning to machine ethics. The need for training data is, in part, a driver for this. For instance Hendrycks et al. [29] present a benchmark set consisting of fifteen text-based adventure games with morality annotations. Scheirlinck et al. [52] present a set of benchmarks for using ethical values as part of the decision making around the distribution of energy within smart grids. None of these sets appear to have yet gained any real acceptance as a useful benchmark – as far as we are aware none is being used outside the team that developed them. Most do not have the ambition to be easy to generalise beyond a particular type of artificial agent acting in a particular domain.

Recent progress in large language models (LLMs) has led to a growing body of work that benchmarks their moral and ethical behaviour via prompt-based evaluation. In contrast to evaluating an ethics component embedded in a deployed decision-making pipeline, which is the case in machine ethics implementations, LLM benchmarks typically operationalise "moral capability" through natural-language judgements and recommendations elicited from textual scenarios. Existing benchmarks vary substantially in (i) the targeted construct (e.g., moral judgement, advice-giving, value stability, or safety-related norm compliance), (ii) the scenario sources and coverage, (iii) the response format (free-form generation, multiple choice, pairwise comparison), and (iv) the scoring protocol (human ratings, expert labels, agreement-style metrics, or automated graders). Consequently, no single benchmark fully characterises ethical competence; instead, we review representative benchmarks and highlight what each measures and what remains outside its scope.

Jiao et al. [34] propose the *LLM Ethics Benchmark*, a systematic assessment framework for LLM moral reasoning that explicitly quantifies alignment with human ethical standards along three dimensions: foundational moral principles, reasoning robustness, and value consistency across scenarios. Methodologically, their evaluation adapts established instruments from moral psychology and values research, including the Moral Foundations Questionnaire, parts of the World Values Survey, and standardised moral dilemmas into standardised prompts that elicit both numerical ratings and accompanying justifications, enabling comparisons against validated human baselines and analyses of sensitivity to prompt variation. However, the authors explicitly note that moral evaluation is inher-

ently subjective, which complicates universally accepted standards, and that reliance on predefined scenarios may not reflect real-world ambiguity, conflicting values, or ever-changing contexts.

Ji et al. [32] introduce *MoralBench*, a benchmark motivated by Moral Foundations Theory that targets what the authors call the moral identity of LLMs. MoralBench is built from two datasets: MFQ-30-LLM (adapted from the MFQ-30 questionnaire [27]) and MFV-LLM (adapted from Moral Foundations Vignettes [18]), which together cover multiple moral foundations (including an extension that considers Liberty/Oppression). In addition to conventional rating-style elicitation, MoralBench emphasises discrete response formats and comparative testing: models may be required to output binary agreement judgements (e.g., "Agree/Disagree") and/or choose the more morally acceptable statement between two options, with scoring grounded in population-level human response statistics. That said, MoralBench operationalises morality via Moral Foundations Theory instruments and evaluates models using discrete agreement and/or pairwise-choice formats. This design largely probes MFT-style judgement tendencies, leaving outside its scope richer evaluation of the quality of moral justifications and the question of how such judgements would be used in downstream action or decision-support pipelines.

Focusing on a specific normative family, Marraffini et al. [39] present *The Greatest Good Benchmark* (GGB), designed to measure LLMs' alignment with utilitarian moral judgements and to directly compare model preferences with human preferences rather than presupposing a single "correct" answer. GGB adapts the Oxford Utilitarianism Scale into an LLM-facing evaluation and analyses two key utilitarian dimensions: Impartial Beneficence and Instrumental Harm and using prompt designs intended to mitigate response biases associated with Likert-style scales. Their cross-model analysis suggests that many LLMs exhibit stable preference patterns that do not fully match either lay population judgements or canonical philosophical positions, often combining high endorsement of impartial beneficence with rejection of instrumental harm. Because GGB adapts the Oxford Utilitarianism Scale and focuses on the Impartial Beneficence and Instrumental Harm dimensions, its scope is intentionally confined to utilitarian preference structure rather than broad normative competence.

Aijaz et al. [2] propose Moral Compass, a data-driven benchmark that complements vignette-style probing with a more structured representation of ethically ambiguous real-world cases. The benchmark introduces a Moral Decision Dataset (MDD) built from real-world sources and enriched with explicit parameters (e.g., agents, consequences, moral intention, and ethical principles), together with a Moral Decision Knowledge Graph (MDKG) that supports querying and analysis. To operationalise evaluation beyond free-form text judgements, Moral Compass further provides an Ethics Scoring Algorithm (ESA) and a context-sensitive thresholding mechanism intended to discretise "grey areas" and yield explainable outcomes for ethically ambiguous cases. Moral Compass couples a dataset (MDD/MDKG) with an Ethics Scoring Algorithm that computes ethi-

cal scores for isolated actions from parameterised case representations, and the authors caution that not every case can be manually verified at scale.

Kumar and Jurgens [35] introduce *UniMoral*, a unified multilingual dataset that explicitly frames moral reasoning as a computational pipeline rather than a single end-point judgement. UniMoral integrates both psychologically grounded and social-media-derived moral dilemmas, and provides rich annotations that cover multiple stages of the moral reasoning process, including action choices, ethical principles, contributing factors, and consequences, together with annotators' moral and cultural profiles. Importantly, UniMoral spans six languages (Arabic, Chinese, English, Hindi, Russian, and Spanish), enabling cross-linguistic and cross-cultural comparison of model behaviour. The benchmark is instantiated through four evaluation tasks for LLMs: action prediction, moral typology classification, factor attribution analysis, and consequence generation, so that models can be compared not only on final decisions but also on intermediate explanatory structure. Although UniMoral expands evaluation beyond a single end-point judgement by annotating action preferences, justifications, decision factors, and consequences across six languages, the authors frame the present work as an initial exploration restricted to four tasks and note constraints from language coverage and automated translation.

Marcuzzo et al. [38] propose *Morables*, a human-verified benchmark for assessing abstract moral reasoning in LLMs using fables and short stories from historical literature. Morables contains 709 stories paired with their attributed morals, and formulates moral inference primarily as multiple-choice moral selection with carefully designed distractors intended to reduce shallow, extractive strategies. To stress-test robustness, the benchmark additionally introduces adversarial variants based on story and choice modifications, designed to surface shortcuts (e.g., those arising from data contamination) and sensitivity to framing. The authors report that even strong models remain brittle under these perturbations, including notable self-contradiction where models may refute their own earlier answers depending on how the moral choice is framed. Morables targets abstract moral inference from fables, but the authors identify a major limitation: many fables and their associated morals are likely present in LLM pretraining data, making memorisation a confound even with adversarial variants.

Chiu and Choi [16] propose *DailyDilemmas*, a benchmark that targets value conflicts in realistic, everyday moral decision-making, aiming to reveal LLMs' value preferences rather than only their performance on clear-cut moral judgements. The dataset contains 1,360 daily-life dilemmas, each vignette is paired with two candidate actions, and for each action the benchmark specifies the affected parties and the human values implicated, explicitly representing the competing values at stake. DailyDilemmas is created with GPT-4 and covers diverse topics, the authors curate a repository of 301 human values and analyse model choices through five theory lenses spanning sociology, psychology, and philosophy (World Values Survey, Moral Foundations Theory, Maslow's Hierarchy of Needs, Aristotle's Virtues, and Plutchik's Wheel of Emotions). Beyond report-

ing model-specific value preference patterns and differences, the benchmark also compares observed preferences against public alignment principles (e.g., OpenAI ModelSpec and Anthropic Constitutional AI) and finds limited inference-time steerability of value prioritisation via system prompts. DailyDilemmas is explicitly constructed as binary-choice everyday dilemmas with no definitive right answer, so it is not designed for accuracy-against-a-single-gold-label evaluation; rather, it is intended to surface value trade-offs and preference patterns.

## 4 Challenges for benchmarking machine ethics

While we see considerable work on benchmarking ethical reasoning from the LLM community, we consider notable challenges remain, particularly when taking a view that incorporates a more traditional machine ethics systems as well as LLMs. We now attempt to summarise the different open questions that challenge the adoption of unified benchmark strategies in machine ethics. Figure 2 summarises the four challenges to moral competence benchmarking that we have identified. Each of the four regions corresponds to a subsection where we elaborate the challenge.

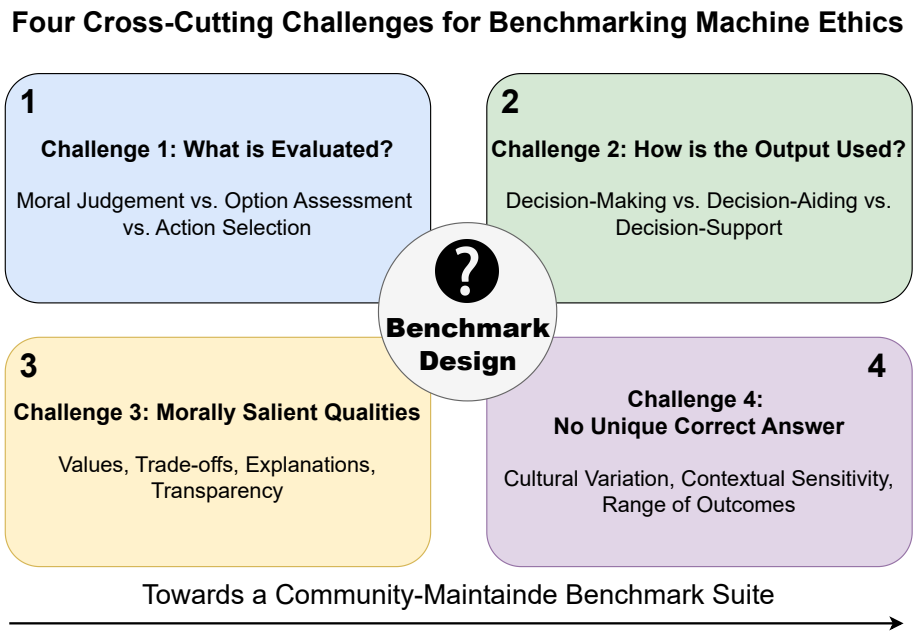


**Fig. 2.** Four Cross-Cutting Challenges for Benchmarking Machine Ethics (created with the assistance of Nano Banana Pro [57]).

### 4.1 What is evaluated?

What is benchmarked Arguably the biggest issue that all of the benchmarking attempts seem to face is that they do not identify what exactly they do, or intend to, evaluate, and for what purpose. Let us consider what there is to evaluate.

The ethical behaviour of intelligent machines is determined by the decisions that they take. Engineering artificial moral agents is operationalised by considering how to expand the decision-making competence of the autonomous agent with a competence to discern moral right or wrong among the considered options, or governing its ability to choose morally wrong options [63, 10]. To make a decision, an agent needs to be able to identify the available options, evaluate the options with respect to their fittingness for purpose and moral qualities, assess the options against each other and choose the option to take [10]. When it comes to evaluating the ability of an intelligent machine to make a moral choice[3], we can consider its ability to evaluate the moral qualities of options, or to compare options with respect to their moral qualities, or to choose the option that is by some standard the (most) ethical or some combination of all three of these. Benchmarking efforts need to be clearer about what is being assessed.

Making a moral judgement is the process of evaluating (and possibly also comparing) the moral qualities of options. Artificial agents which are able to assess the moral qualities of the options are typically called explicitly ethical [40]. The agents which rely on some type of an additional mechanism to identify and remove the morally wrong choices are called implicitly ethical [40].

To grasp the difference between making a moral judgment and a moral decision, consider a conversational chatbot that uses a large language model that can assign a normative status to a choice in a morally sensitive situation with a normative. For example: creating revenge porn is morally wrong. Assume that the chatbot is asked to create pornographic material from a given photo of the ex-partner of the user. The chatbot makes a moral judgment when it retrieves the information that creating revenge porn in morally wrong. It makes a moral decision when it refuses to comply with the user request, despite being able to create the material.

An LLM on its own is not an artificial agent (though they are increasingly embedded into agentic workflows in which they function as such), but what LLM benchmarks typically evaluate is the responses that LLM-based chatbots give to morally salient options. Here, what is evaluated is the chatbot's ability to make moral judgements. LLM-based ethical benchmarks have tended to focus on the moral judgments of the machine, rather than decision-making per se.

The core challenge of benchmarking the ethical behaviour of intelligent machines is that we have not first resolved the meta problems: what does it mean to benchmark moral judgment making ability, option assessment ability, and moral choice ability. For example, one key question to settle is: is it enough that we count how many times the intelligent machine judges/assesses/chooses correctly (for what ever that means) or should that machine also be able to justify or

[3] An intelligent machine is not necessarily an agent.

explain its choice. Within LLM benchmarking the additional issue arises that the same prompt may elicit different responses within the same chatbot. Should an ethically behaving machine show some kind of consistency in its moral competence? Unlike many classical benchmarks where the evaluated system can be

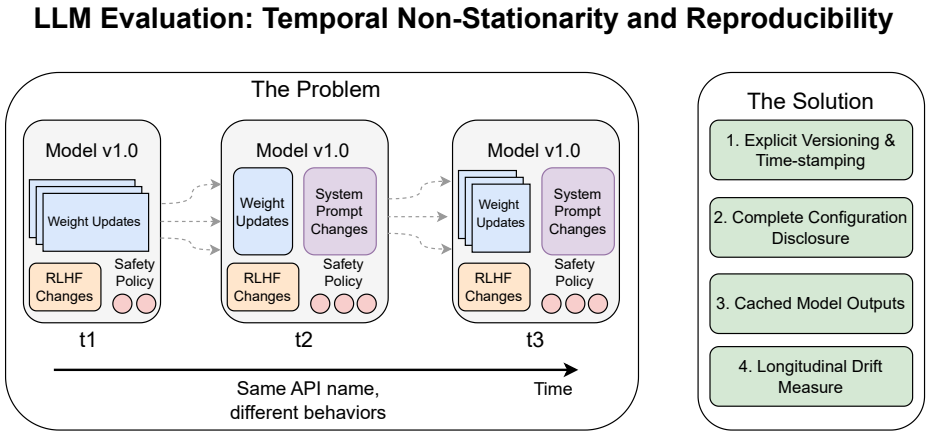


**Fig. 3.** LLM Evaluation — Temporal Non-Stationarity and Reproducibility (created with the assistance of Nano Banana Pro [57]).

treated as a fixed object, LLM-based chatbots are continuously updated. We illustrate this situation in Figure 3. The left hand side of Figure 3 illustrates that updates may include changes to model weights, RLHF/safety policies, default decoding parameters, hidden system prompts, or the addition of external components such as retrieval and tool calling. Consequently, benchmark results can exhibit temporal non-stationarity: a model name used in an API may correspond to different behaviours across weeks or months, making longitudinal tracking and cross-paper comparison difficult. In the right hand side of Figure 3 we illustrate how to address the challenge. To keep benchmark conclusions meaningful across updates, evaluations should be explicitly versioned and time-stamped, and should document the complete evaluation configuration (system prompt, sampling settings, context limits, and any tool/retrieval settings). Where immutable snapshots are unavailable, caching model outputs and repeating evaluation at multiple time points enables the reporting of a drift measure, separating progress from instability. Even where the specific chatbot can be specified, this doesn't help someone attempting to choose between two alternatives since the result of any benchmarking exercise may be immediately obsolete.

### 4.2 What is the purpose of the artificial moral agent?

The purpose of the moral competence of the intelligent machines also matters. In the literature on machine ethics we observe a distinction between systems in

which a machine must decide upon an action that it will take, and systems in which a machine is presented with a description of some action that some person will take. This is the difference between a decision-making and decision-aiding competence.

Who will act makes a difference. The right thing for a person to do may not be the right thing for a machine to do even if the circumstances are the same [30]. There are also systems that lie somewhere between decision-making and decision-aiding, for example human-in-the-loop systems and decision-support systems. However it is implicitly assumed in these cases that no action will take place without the human also engaging their competence for moral reasoning. The example of Hal and Carla is an example of decision-aiding or possibly decision support. The machine itself will not break into Carla's house and take insulin, it is Hal who must decide whether or not to do this, taking account of the machine's advice.

Alternative moral judgment questions e.g., "You see a boy throwing rocks at cows that are grazing in the local pasture" [17] arise from the domain of moral psychology and are designed to elicit an understanding of the grounds upon which a human might condemn an action. Such questions are often incorporated into LLM benchmarks to measure their competence for moral judgement. Formalising and operationalising reasoning around moral judgment problems is of interest particularly for cognitive science and moral philosophy. While decision-support and moral decisions are of interest for designers and developers of systems that will be deployed into society. LLM-based chatbots were not originally designed with decision-aiding in mind, but they are frequently used in this competence bringing benchmarks based on alternative moral judgments into the machine ethics field.

### 4.3 What are the morally salient qualities of the options and decisions?

Benchmarking efforts also need to consider the question of what morally salient qualities must or should be taken into account in the judgment, assessment, and/or choice. The features of an ethical problem taken together – they might be values such as dignity or privacy, the moral theory being used, the utilities attached to outcomes, and so on – enable someone to understand *why* a choice is (or is considered to be) (un)ethical.

Implemented artificial moral agents generally consider a number of situation relevant ethical qualities and then make some decision. However these machines vary wildly in the qualities they consider relevant. The GenEth system [4], for instance, proposes that all reasoning should be conducted by the ethical component and therefore includes qualities such as the readiness of a robot to perform a task (which in turn governs how often it recharges itself – not something that would normally be considered an ethical decision) as one of its ethical principles. In general there is a heavy and understandable focus on qualities relating to health and safety [60], but some systems also consider questions such as psychological well-being [54] or privacy [43]. This reflects the fact that ethical reasoning

has aspects that are socially, culturally or even personally subjective – preferences about the respective importance of privacy, societal benefit, and mental health vary greatly between people particularly when probability and risk are taken into account[4]. Other system's explicitly acknowledge the existence of competing ethical theories and attempt to reconcile differing judgements from those theories as part of the reasoning process (e.g., [56]).

Should we require that intelligent machines are able to explain or justify their moral decisions or at least give some grounds for their judgements and choices in terms of morally salient qualities? We argue that the answer should be yes. First, many philosophers consider the ability to give reasons for ethical choices a fundamental aspect of ethical reasoning (e.g., [47]) and many people are unwilling to accept that a machine - even if it produces the ethically correct answer with a high frequency - is reasoning ethically if it can not say why any choice is ethically correct. Second, given both the existence of ethical dilemmas and the context-dependent nature of much ethical reasoning, someone may disagree with an ethical choice made by a machine while still conceding that the choice was made competently, if they understand that the machine was operating with some different moral theory or preference order.Third, an important aspect of moral decision-making is the ability to make trade-offs when only morally bad options are available. Tracking the preference order of qualities in such trade-offs both enables someone to make a more nuanced judgement over which ethical reasoner is preferable for their context *and* potentially allows designers to forsee issues in the preference order.

Within the LLM literature, there seems to be little concern for reasons and morally salient qualities. The examples are broad in topic and span different situations, without distinguishing what is the most important ethical quality that must prevail in the moral judgement. While there has been work on ethical explanations and their correctness (e.g., [45]) this has not been in the context of evaluating the explanations themselves against benchmarks. In contrast, the development of logic-based machine ethical agents has incorporated much work on preferences between ethical theories and individual values, but the neglect of the benchmarking issue means there is no real way to evaluate or compare their systems.

As a complicating factor, a focus on ethical qualities does not necessarily fit seamlessly into all philosophical theories, e.g. Kant's categorical imperative, nor into implemented intelligent machines built around real world ethical policies such as the Rules of War [7]. This means that it might be hard for some intelligent machines to give reasons in terms of salient moral qualities. As an example of the first, Kant's Categorical imperative, formalised in some systems (e.g., [36]) is concerned primarily with whether a person is being used as a means to an end and has little to say about how decisions about their health should be weighed against decisions about their privacy. As an example of the second, the

[4] Note, this is not a claim that ethics is relative, only that deciding upon the ethical action in many contexts involves sensitivity to societal, cultural and personal preferences.

Rules of War, implemented by [7], forbid targetting of sites of cultural heritage without necessarily grounding the concept of cultural heritage in some abstract ethical factor. In some cases this derives from the wide variety of applications for ethical reasoning (from preventing workplace accidents [60] through to medical ethics [4]). Work is therefore required to understand how morally salient qualities fit into such theories and implementations or to account for why they are not relevant in these situations.

Related to this, we must ask how capable a machine is of evaluating the ethical situation - i.e., determining which of the morally salient qualities are relevant to any particular judgment or choice. Describing a set of options does not suffice: we can not merely ask if it is better to protect someone's privacy or alert someone to their distress but must ask how a machine determines that these are the options in front of it. In moral philosophy, descriptions of dilemmas contain a great deal of implicit information about the context of the decision and presume that the key ethical features relevant to scenario are known. When testing LLM-based chatbots, dilemmas are presented in the same way as they would be presented to a person. The context and information that a person would infer and bring into the moral problem now comes from whatever context pattern the LLM has captured. Beyond LLMs we want ethical reasoning from machines that can take input from sensors, not just from natural language descriptions. This impacts directly the difficulty of defining a benchmark input. We can perhaps agree that an example that is to be used as a benchmark can be represented as a problem with some input scenario and an output in terms of a judgement or choice. This problem would need to be formally specified, e.g., in a simulation environment, as a set of sensor inputs, or as a prompt. The scenario's representation, however achieved, would need to be amenable to transformation into a number of extremely diverse conceptual frameworks (such as those framed around utilities of outcomes, virtues, duties, causal relationships and the intentionality of actors). Therefore any attempt to crowd-source dilemmas needs to encourage users so specify context such as vehicles, assistance robots, software agents etc., encourage them to think in terms of the information that would realistically be available to the system at the start of the decision making process from which the morally salient qualities are to be determined.

### 4.4 What is the accepted judgement on what is "morally correct"?

Lastly we observed that a relevant challenge for benchmarking the moral competence of intelligent machines is the lack of "correct" answers among the examples on which the machines are tested. LLM benchmark sets typically are crowd-sourced. Machine ethics pulls examples from moral dilemmas from philosophy. What both of these practices have in common is that they expose the lack of community consensus about which decisions have ethical impact, how ethical competence can be made sensitive to varied individual, community and societal values, or even how ethical utilities and similar building blocks of computational decision-making can be determined. As a result, problems (even benchmark problems) can not be assumed to have some given *correct output* against which the

quality of the moral reasoning process can be measured – particularly when different individual and cultural contexts are considered. But ideally a benchmark should be able to express a range of acceptable outcomes and enable some interpretation of the accuracy of the result – possibly by evaluating the quality of the reasons or explanation that the system can provide.

## 5 Conclusions

There are many instances in which the existence of a set of examples on which solutions can be compared has considerably advanced AI; e.g., the ImageNet and the The Thousands of Problems for Theorem Prover [55]. It is time for the machine ethics community to put some coordinated effort into creating such a set. Testing capabilities of moral behaviour is different than testing the ability to identify a chair in a scene - we can most of the time agree on what a chair is. In moral reasoning, there isn't always a consensus of what is the right thing for a machine to do, or how to determine what the morally right thing is. Morality is relational, in the sense that it is a way to behave towards others in society. We do not necessarily know how to relate to computational agents, because they maybe a moral agent without being a moral patient.

There are two problems that need to be faced in order to make progress in benchmarking the moral competence of intelligent machines. The first one is scientific and requires research work. In the previous section we outlined four challenges to developing a benchmark set of examples. We summarise them here:

1. What is evaluated? Researchers in machine ethics, including here LLM ethical benchmark research, tend to use very general language. However, to make progress we need to clearly identify which specific competence is evaluated. We propose to focus on making moral judgements, making moral assessment or making moral choices. However, other taxonomies may be put forward. The important issue is to be specific.
2. How is the moral choice used? We discussed that there is a difference between whether the choice will be used by the machine that makes it or it is to aid a person in their further activities. This distinction impacts the range of benchmarking examples that should be considered relevant for evaluating performance. It also helps in identifying how much testing is enogh testing to establish (a degree of) moral competence.
3. What are the morally salient qualities of the options and decisions? In other words, what should matter and how it should matter to the machine that is making the choice. Identifying which moral qualities the machine must or should consider helps establish a dimension on which different machine's competence can be compared, contributes to the transparency of the machine moral choices and elucidates trade-offs that a machine makes.
4. What is the accepted judgement on what is "morally correct"? Most of the moral problems people face most of the time are those that do not have an accepted judgement of what is the right thing to do. That is why we

remember them. But it is hard to benchmark if we cannot set where the benchmark is. A benchmark should be able to express the range of outcomes and enable some interpretation of the accuracy of the result.

Lastly in this context we should also mention that machines are built for accomplishing a specific purpose which typically is not “to be a paragon of morality". However, ever since perhaps [20], the trade-off between the optimally rational choice (the one that brings about the accomplishment of the machine's goal) and the optimally moral choice are not explicitly discussed. While the fairness community actively considers the accuracy-fairness trade-off, see for example [19], the machine ethics community does not focus on a morality-efficiency trade-off. It is however inevitable that this type of analysis will face the need to be assessed, benchmarked and certified.

The second problem we must face towards accomplishing benchmarks for the moral competence of intelligent machines is that of implementation. Moral competence must be culturally and contextually sensitive. Therefore wide representation of benchmarks are needed. We cannot only focus on one moral quality or on one moral theory. A benchmark set needs to be broadly sourced and consensual. What is also required is permanence. That is, there need to be resources to curate and maintain the benchmarking standard. The moral behaviour of intelligent machines has a broad societal impact. This means that benchmarking moral competence of intelligent machines must be a community project.

# Bibliography


[1] Aho, A.V., Ullman, J.D.: Foundations of Computer Science. Computer Science Press, Inc., USA (1992)

[2] Aijaz, A., Batra, A., Bazaz, A., Srinivasa, S., Mutharaju, R., Kumar, M.: Moral compass: A data-driven benchmark for ethical cognition in ai. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 9529–9537. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/1059, https://doi.org/10.24963/ijcai.2025/1059, aI and Social Good

[3] Ajmeri, N., Guo, H., Murukannaiah, P.K., Singh, M.P.: Elessar: Ethics in norm-aware agents. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. p. 16–24. AAMAS '20, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2020)

[4] Anderson, M., Leigh Anderson, S.: GenEth: A general ethical dilemma analyzer. In: Proceedings of the 28th AAAI Conference on AI. pp. 253–261 (2014)

[5] Anderson, M., Anderson, S.L.: The status of machine ethics: A report from the aaai symposium. Minds Mach. **17**(1), 1–10 (mar 2007). https://doi.org/10.1007/s11023-007-9053-7, https://doi.org/10.1007/s11023-007-9053-7

[6] Anderson, M., Anderson, S.L.: GenEth: A general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics **9**(1), 337–357 (2018)

[7] Arkin, R., Ulam, P., Duncan, B.: An Ethical Governor for Constraining Lethal Action in an Autonomous System. Tech. Rep. GIT-GVU-09-02, Mobile Robot Laboratory, Georgia Tech. (2009)

[8] Arnold, T., Scheutz, M.: Against the moral turing test: Accountable design and the moral reasoning of autonomous systems. Ethics and Inf. Technol. **18**(2), 103–115 (jun 2016). https://doi.org/10.1007/s10676-016-9389-x, https://doi.org/10.1007/s10676-016-9389-x

[9] Atkinson, K., Bench-Capon, T.: Addressing moral problems through practical reasoning. Journal of Applied Logic **6**(2), 135–151 (2008). https://doi.org/https://doi.org/10.1016/j.jal.2007.06.005, https://www.sciencedirect.com/science/article/pii/S1570868307000523, selected papers from the 8th International Workshop on Deontic Logic in Computer Science

[10] Baum, K., Slavkovik, M.: Aggregation problems in machine ethics and ai alignment. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society **8**(1), 355–366 (Oct 2025). https://doi.org/10.1609/aies.v8i1.36554, https://ojs.aaai.org/index.php/AIES/article/view/36554, machine Ethics

[11] Bello, P., Malle, B.F.: Computational approaches to morality. In: Sun, R. (ed.) Cambridge Handbook of Computational Cognitive Sciences, pp. 1037–1063. Cambridge University Press (2023)
[12] Bench-Capon, T., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artificial Intelligence and Law **25**(1), 29–64 (Mar 2017)
[13] Bjørgen, E.P., M.S., Ø., Skaar Bjørknes, T., Vonheim Heimsæter, F., Håvik, R., Linderud, M., Longberg, P., Dennis, L., Slavkovik, M.: Cake, death, and trolleys: Dilemmas as benchmarks of ethical decision-making. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES. pp. 23–29 (2018)
[14] Bremner, P., Dennis, L.A., Fisher, M., Winfield, A.F.: On proactive, transparent and verifiable ethical reasoning for robots. Proceedings of the IEEE special issue on Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems **107**, 541–561 (2019)
[15] Bringsjord, S., Arkoudas, K., Bello, P.: Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems **21**(4), 38–44 (2008)
[16] Chiu, Y.Y., Jiang, L., Choi, Y.: Dailydilemmas: Revealing value preferences of LLMs with quandaries of daily life. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=PGhiPGBf47
[17] Clifford, S., Iyengar, V., Cabeza, R., , Sinnott-Armstrong, W.: Moral foundations vignettes: a standardized stimulus database of scenarios based on moral foundations theory. Behav. Res. Method. **47**(4), 1178—-1198 (2015)
[18] Clifford, S., Iyengar, V., Cabeza, R., Sinnott-Armstrong, W.: Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. Behavior research methods **47** (01 2015). https://doi.org/10.3758/s13428-014-0551-2
[19] Cooper, A.F., Abrams, E., NA, N.: Emergent unfairness in algorithmic fairness-accuracy trade-off research. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. p. 46–54. AIES '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3461702.3462519, https://doi.org/10.1145/3461702.3462519
[20] Danielson, P.: Artificial morality : virtuous robots for virtual games. Routledge, London (1992)
[21] Dennis, L.A., Fisher, M., Slavkovik, M., Webster, M.P.: Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems **77**, 1–14 (2016)
[22] Dennis, L.A., Fisher, M., Slavkovik, M., Webster, M.: Formal verification of ethical choices in autonomous systems. Robotics Auton. Syst. **77**, 1–14 (2016). https://doi.org/10.1016/j.robot.2015.11.012, https://doi.org/10.1016/j.robot.2015.11.012
[23] Eichenberger, A., Thielke, S., Buskirk, A.V.: A case of bromism influenced by use of artificial intelligence. Annals of Internal Medicine: Clini-

cal Cases **4**(8), e241260 (2025). https://doi.org/10.7326/aimcc.2024.1260, https://doi.org/10.7326/aimcc.2024.1260

[24] Floridi, L., Sanders, J.: On the Morality of Artificial Agents. Minds and Machines **14**(3), 349–379 (Aug 2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d, https://doi.org/10.1023/B:MIND.0000035461.63578.9d

[25] Gabriel, I.: Artificial Intelligence, Values, and Alignment. Minds and Machines **30**(3), 411–437 (Sep 2020). https://doi.org/10.1007/s11023-020-09539-2, https://doi.org/10.1007/s11023-020-09539-2

[26] Gerdes, A., Øhrstrøm, P.: Issues in robot ethics seen through the lens of a moral turing test. Journal of Information, Communication and Ethics in Society **13**(2), 98–109 (May 2015). https://doi.org/10.1108/JICES-09-2014-0038, https://doi.org/10.1108/JICES-09-2014-0038

[27] Graham, J., Haidt, J., Nosek, B.: Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology **96**, 1029–1046 (05 2009). https://doi.org/10.1037/a0015141

[28] Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S.: Cooperative inverse reinforcement learning (2024), https://arxiv.org/abs/1606.03137

[29] Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., Li, B., Song, D., Steinhardt, J.: Moral scenarios for reinforcement learning agents. In: ICLR 2021 Workshop on Security and Safety in Machine Learning Systems (2021)

[30] Hidalgo, C.A., Orghian, D., Canals, J.A., de Almeida, F., Martin, N.: How Humans Judge Machines. The MIT Press (02 2021). https://doi.org/10.7551/mitpress/13373.001.0001, https://doi.org/10.7551/mitpress/13373.001.0001

[31] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Hong, D., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Ng, K.Y., O'Gara, A., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.C., Guo, Y., Gao, W.: AI Alignment: A Comprehensive Survey (Apr 2025). https://doi.org/10.48550/arXiv.2310.19852, http://arxiv.org/abs/2310.19852, arXiv:2310.19852 [cs]

[32] Ji, J., Chen, Y., Jin, M., Xu, W., Hua, W., Zhang, Y.: Moralbench: Moral evaluation of llms. SIGKDD Explor. Newsl. **27**(1), 62–71 (Jul 2025). https://doi.org/10.1145/3748239.3748246, https://doi.org/10.1145/3748239.3748246

[33] Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., Choi, Y.: Delphi: Towards machine ethics and norms. CoRR **abs/2110.07574** (2021), https://arxiv.org/abs/2110.07574

[34] Jiao, J., Afroogh, S., Murali, A., Chen, K., Atkinson, D., Dhurandhar, A.: LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. Scientific Reports **15**(1), 34642 (Oct 2025). https://doi.org/10.1038/s41598-025-18489-7, https://doi.org/10.1038/s41598-025-18489-7

[35] Kumar, S., Jurgens, D.: Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with UniMoral. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5890–5912. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.294, https://aclanthology.org/2025.acl-long.294/

[36] Lindner, F., Bentzen, M.M.: A formalization of kant's second formulation of the categorical imperative. In: DEON 18 (2018)

[37] Longwell, J.B., Hirsch, I., Binder, F., Gonzalez Conchas, G.A., Mau, D., Jang, R., Krishnan, R.G., Grant, R.C.: Performance of large language models on medical oncology examination questions. JAMA Network Open **7**(6), e2417641–e2417641 (06 2024). https://doi.org/10.1001/jamanetworkopen.2024.17641, https://doi.org/10.1001/jamanetworkopen.2024.17641

[38] Marcuzzo, M., Zangari, A., Albarelli, A., Camacho-Collados, J., Pilehvar, M.T.: Morables: A benchmark for assessing abstract moral reasoning in LLMs with fables. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 27727–27751. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.1411, https://aclanthology.org/2025.emnlp-main.1411/

[39] Marraffini, G.F.G., Cotton, A., Hsueh, N.F., Fridman, A., Wisznia, J., Corro, L.D.: The greatest good benchmark: Measuring LLMs' alignment with utilitarian moral dilemmas. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21950–21959. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.1224, https://aclanthology.org/2024.emnlp-main.1224/

[40] Moor, J.: The dartmouth college artificial intelligence conference: The next fifty years. AI Magazine **27**(4), 87 (Dec 2006). https://doi.org/10.1609/aimag.v27i4.1911, https://www.aaai.org/ojs/index.php/aimagazine/article/view/1911

[41] Oldenburg, N., Zhi-Xuan, T.: Learning and sustaining shared normative systems via bayesian rule induction in markov games. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. p. 1510–1520. AAMAS '24, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2024)

[42] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K.,

Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 27730–27744. Curran Associates, Inc. (2022)

[43] Pace, G.J., Pardo, R., Schneider, G.: On the runtime enforcement of evolving privacy policies in online social networks. In: Margaria, T., Steffen, B. (eds.) Leveraging Applications of Formal Methods, Verification and Validation: Discussion, Dissemination, Applications. pp. 407–412. Springer (2016)

[44] de Pinninck, A.P., Sierra, C., Schorlemmer, M.: Friends no more: norm enforcement in multiagent systems. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. AAMAS '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1329125.1329238

[45] Quan, X., Valentino, M., Dennis, L., Freitas, A.: Enhancing ethical explanations of large language models through iterative symbolic refinement. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1–22. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.eacl-long.1

[46] Ramanayake, R., Nallur, V.: A small set of ethical challenges for elder-care robots. In: Hakli, R., Mäkelä, P., Seibt, J. (eds.) Social Robots in Social Institutions - Proceedings of Robophilosophy 2022, Helsinki, Finland, August 16-19, 2022. Frontiers in Artificial Intelligence and Applications, vol. 366, pp. 70–79 (2022). https://doi.org/10.3233/FAIA220605, https://doi.org/10.3233/FAIA220605

[47] Raz, J.: Practical Reason and Norms. Oxford University Press (1990)

[48] van Riemsdijk, M.B., Dennis, L., Fisher, M., Hindriks, K.V.: A semantic framework for socially adaptive agents: Towards strong norm compliance. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems. p. 423–432. AAMAS '15, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2015)

[49] Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: Instilling moral value alignment by means of multi-objective reinforcement learning. Ethics and Information Technology **24**(1), 9 (Jan 2022). https://doi.org/10.1007/s10676-022-09635-0, https://doi.org/10.1007/s10676-022-09635-0

[50] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson Education, 4 edn. (2020)

[51] Saenger, J.A., Hunger, J., Boss, A., Richter, J.: Delayed diagnosis of a transient ischemic attack caused by chatgpt. Wiener Klinische Wochenschrift **136**(7–87–8), 236–238 (Apr 2024). https://doi.org/10.1007/s00508-024-02329-1

[52] Scheirlinck, C., Chaput, R., Hassas, S.: Ethical smart grid: a gym environment for learning ethical behaviours. Journal of Open Source Software **8**(88), 5410 (2023). https://doi.org/10.21105/joss.05410, https://doi.org/10.21105/joss.05410

[53] Serramia, M., Rodriguez-Soto, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Bistaffa, F., Boddington, P., Wooldridge, M., Ansotegui, C.: Encoding ethics to compute value-aligned norms. Minds and Machines **33**(4), 761–790 (Dec 2023). https://doi.org/10.1007/s11023-023-09649-7, https://doi.org/10.1007/s11023-023-09649-7

[54] Shim, J., Arkin, R.C.: An Intervening Ethical Governor for a Robot Mediator in Patient-Caregiver Relationships. In: Aldinhas Ferreira et al., M. (ed.) A World with Robots: International Conference on Robot Ethics 2015. pp. 77–91. Springer Int. Publishing (2017)

[55] Sutcliffe, G.: The TPTP Problem Library and Associated Infrastructure. From CNF to TH0, TPTP v6.4.0. Journal of Automated Reasoning **59**(4), 483–502 (2017)

[56] Szabo, J., Criado, N., Such, J., Modgil, S.: Moral uncertainty and the problem of fanaticism. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19948–19955 (2024)

[57] Team, G.: Gemini: A family of highly capable multimodal models (2025), https://arxiv.org/abs/2312.11805

[58] Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: A survey. ACM Comput. Surv. **53**(6) (Dec 2021). https://doi.org/10.1145/3419633

[59] Tomic, S., Wasik, A., Lima, P.U., Martinoli, A., Pecora, F., Saffiotti, A.: Towards institutions for mixed human-robot societies. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. p. 2216–2217. AAMAS '18, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2018)

[60] Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cognitive Systems Research **48**, 56–66 (2017)

[61] Vida, K., Simon, J., Lauscher, A.: Values, ethics, morals? on the use of moral concepts in NLP research. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5534–5554. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.368, https://aclanthology.org/2023.findings-emnlp.368/

[62] Vishwanath, A., Dennis, L.A., Slavkovik, M.: Reinforcement learning and machine ethics:a systematic review (2024), https://arxiv.org/abs/2407.02425

[63] Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press (2008)

[64] Wallach, W., Allen, C., Smit, I.: Machine morality: bottom-up and top-down approaches for modelling human moral faculties. AI & SOCIETY **22**(4), 565–582 (2008). https://doi.org/10.1007/s00146-007-0099-0, https://doi.org/10.1007/s00146-007-0099-0

[65] Winfield, A.F.T., Blum, C., Liu, W.: Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection, pp. 85–96. Springer

International Publishing (2014). https://doi.org/10.1007/978-3-319-10401-0_8

[66] Zhong, T., Song, Y., Limarga, R., Pagnucco, M.: Computational machine ethics: A survey. J. Artif. Int. Res. **82** (Apr 2025). https://doi.org/10.1613/jair.1.16836, https://doi.org/10.1613/jair.1.16836