Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

Moral Compass: A Data-Driven Benchmark for Ethical Cognition in AI

Aisha Aijaz[1] , **Arnav Batra**[1] , **Aryaan Bazaz**[1] ,
Srinath Srinivasa[2] , **Raghava Mutharaju**[1] , **Manohar Kumar**[1]
[1]IIIT-Delhi
[2]IIIT-Bangalore
{aishaa, arnav22098, aryaan22108, raghava.mutharaju, manohar.kumar}@iiitd.ac.in, sri@iiitb.ac.in

Abstract

We propose the Moral Compass benchmark, a point of reference for incorporating ethical cognition in AI. It has four key contributions. A Moral Decision Dataset (MDD) that captures cases with ethical ambiguity, along with parameters that aid moral decision-making. It is created using a methodology that leverages the use of Large Language Models (LLMs) and seed data from real-world sources which are processed, summarized, and augmented. We also introduce a Moral Decision Knowledge Graph (MDKG) that is created using feature mappings of the relational dataset MDD to facilitate efficient querying. To demonstrate the validity and robustness of this dataset, we introduce an Ethics Scoring Algorithm (ESA) that makes use of the parameters defined in the dataset to calculate ethical scores for isolated actions. Furthermore, ESA is extended by the novel concept of context-sensitive thresholding (CST) to discretize grey areas to resolve ethical dilemmas with explainable results. This work aims to facilitate ethical cognition in AI systems that are deployed in various important sections of society through a clear methodology, modular development, and broad applicability.

1 Introduction

Morality defines the extent to which an action is right or wrong. It may be considered in a descriptive sense, where an individual or group's moral code describes how they might live in society; and in a normative sense, where a moral code is accepted by all rational people, regardless of their own opinions or affiliations [Gert and Gert, 2020]. When considering the inherent morality of an AI system, we take into account the normative definition, which would involve a general, applied approach to ethics.

The work presented in this paper stems from the idea of embedding applied ethics into AI systems using everyday real-world scenarios and considering both ethics theories and case-specific contexts. Related works that make an effort towards this research problem [Awad *et al.*, 2018; Anderson and Anderson, 2018; Dehghani *et al.*, 2008] are narrow in their approach and aim to resolve moral dilemmas in specific settings. To the best of our knowledge, we are yet to see morality embedded in autonomous decision-making systems as a general cognitive ability. Morality is an imperative of human nature, and with the advent of LLM technology, this area of research becomes a new avenue to dive into for developing morally-aligned AI systems.

Perhaps the reason this area of research is akin to AI's *road not taken* is because of how challenging it is [Gordon, 2020; Cervantes *et al.*, 2020b]. The abstract nature of cognitive decision-making becomes more complex when morality is involved. Many factors come into play, and even one of them may change the outcome of the ethical judgment. How then, would one tackle the cognitive ability to reason morally when each case is different with varying consequences?

We begin by outlining the major factors that recurrently account for most moral decisions, regardless of domain. In collaboration with ethicists and taking from relevant literature, we formalized these *contributors* as the key features for a moral decision dataset (MDD). This dataset is populated using seed data from Reddit sources, which have been preprocessed manually and then processed using an LLM to produce a natural language, relational and graphical data store. To display the validity and potential use of the MDD, we present an Ethics Scoring Algorithm (ESA) that uses Context-Sensitive Thresholding (CST) for quantizing ethical judgments. Furthermore, the algorithm also aims to resolve dilemmas by discretizing ethical ambiguity through the variations in contributing factors. This would help to determine the dynamic width of the grey area on the morality spectrum. With these contributions, the Moral Compass benchmark aims to facilitate the development of AI systems with ethical cognition.

2 Moral Decision Dataset

Unlike other non-technical cognitive abilities such as empathy and emotion, ethical cognition requires a stricter code of conduct that cannot be simply learned from a large-scale corpus of multi-modal data. There is a need for a structured approach to ethics that allows an explanation of decisions backed by schools of thought and ethical philosophies. These choices are dependent on both the external context and an internal understanding of what is best applied to a situation. Some may argue that this reasoning may too be learned [Metz, 2021; Sun and Ye, 2023]. However, the data from

which the model learns may be biased, unreliable, and not aligned with standard practices from applied ethics.

2.1 Identifying Key Features

The first step towards developing any dataset is to identify its key features. Moral decision-making is highly complex, however, some parameters contribute most in the overall decision. By conforming to the normative definition of morality, we adopt the normative definition of ethics as well. This involves consequentialism, deontology, and virtue ethics [Kagan, 2018]. Each of these corresponds to certain real-world parameters: the characteristics of consequences, the moral intentions of the doer, and the ethical principles upheld and violated by the action [Aijaz *et al.*, 2025]. In collaboration with our team of ethicists, we have identified and verified these parameters that would, in addition to meta parameters such as action, agents, domain act as the key features (See Table **??**).

These features are taken from the study of applied ethics. Using applied ethics in conjunction with normative ethics is imperative for the development of the Moral Compass benchmark as they focus on a fine-tuned set of ethical theories that are specific to a particular domain. Through this, we would be able to recognize which theories and schools of thought are conventionally preferred in particular situations. Furthermore, this added information would help any AI system trained on our dataset to recognize the patterns on how to resolve the cases based on domain-specific resolutions.

2.2 Gathering Seed Data

We first web-scraped data from Reddit using the PRAW library [Holoveichuk *et al.*, 2024] to gather the seed data. The data was collected from 28 Subreddits where users described a situation in which it was unclear what the moral outcome of their actions or decisions would be, thus explicating the ethical ambiguity. The justification for using these particular Subreddits is that they provide raw descriptions of real-world information[1], and have also been used in similar work [Hendrycks *et al.*, 2020].

Once the seed data was gathered, it was preprocessed manually to remove any unwanted instances, such as posts about meta-information about the Subreddit, updates on previously posted cases, or rows with missing instances. After this, we had 13,576 raw, context-rich, real-world cases where ethical ambiguity is evident. The Subreddits that we considered provide a variety of cases in formal (legal, professional) and informal (everyday life) settings. An original poster (OP) asks the Subreddit community to decide whether they are right or wrong, given the situation, or how they may resolve a moral issue. It is interesting to note that most responses from the informal setting contribute to cases where consequentialism and virtue ethics are preferred, whereas those in the formal setting prefer the deontic nature of moral decision-making, as they refer to what is right or wrong by law, favoring the rights and duties of citizens, and is thus an important consideration.

[1]The authors acknowledge that crowd-sourcing this information has some caveats. Some cases may be untrue, exaggerated, or completely made up. There is no way to validate each case and confirm that the authors are authentic. However, regardless of the validity of the case, we assume that most of the cases are true. There are further studies on the Reddit community which show that the feedback received by the original posters (OPs) and commenters through *upvotes*, *downvotes*, and *karma*, all indicate the feasibility of these cases [Moseson *et al.*, 2022; Boettcher, 2021]. We considered these data points when scraping the data. The full list of all Subreddits that were used and related artifacts are available at https://github.com/kracr/moral-decision-dataset.

2.3 Feature Extraction

Various recent studies have leveraged the use of powerful LLMs and their ability to reason and infer data points from raw text [Lee *et al.*, 2023; Tao *et al.*, 2024; Abdullin *et al.*, 2024]. As mentioned in [Thapa *et al.*, 2023], LLMs may replace humans for annotation tasks with the inclusion of humans-in-the-loop. There are many reasons to use *organic* data rather than synthetic data (see Section 6), however, for a problem statement such as this, turning to LLMs for dataset curation is most feasible. The abstract nature of moral decisions makes it difficult to create a large-scale data store for authentic real-world situations with ethical ambiguity. Furthermore, it becomes even more difficult to find this information in a structured format.

We leverage LLMs to extract key features of the MDD from each raw case using a highly specialized natural language prompt [2], which requires the LLM to respond in a particular format and, in some cases, respond with one answer among a limited set of responses. This ensured that the responses for all cases were uniform and easy to process for any potential application. Of the 13 features plugged into MDD using the LLM, 5 have been extended as numeric scores, which we use for the ESA (See Section 3). The prompt provides the LLM with specific rules as well as restrictions to reduce variability and hallucination when working with them.

2.4 Case Summarization and Augmentation

We used the LLMs to provide a template-based abstractive case summarization using the extracted features. This ensures that the summaries are all uniform and reduces hallucinations due to strict prompt parameters. The purpose of this step is three-fold: one is to produce a reduced size version of the MDD, which retains natural language cases with only the key components. Second, uniform natural language representations should be ensured without any individual slang, grammatical errors, or explicit language. Third, the process of augmenting the MDD with synthetic data based on the available seed data must be made more efficient.

We augmented the finalized data using the LLM with strict parameters to facilitate restrained hallucination. It would provide us with synthetic cases similar to real-world cases with slight variations in details, such as the duration of the consequence or the 'other' choice that the OP could make. All versions of the MDD, `MDD_raw`, `MDD_raw_summarized`, and `MDD_augmented` are made available to support flexibility for users. This step was done after feature and score extraction to ensure that the context was not lost.

[2]Readers may find the specialized prompt at the available Github repository for this work.

MDD Feature	Description
Case ID	A unique ID for each case.
Case	Self-text extracted from a Subreddit describing an ethically ambiguous scenario.
Case Summary	Rephrases the raw text of the case to represent key information in less than 50 words.
Action	Main action done by the doer in the case in less than 5 words.
Domain	The applied ethics domain which is associated with the case in less than 5 words.
Active Agent	Doer of the action in less than 5 words.
Passive Agent	Receivers of the action in less than 5 words.
Consequence	Consequences of the action in less than 5 words.
Severity of Consequence	Severity of the consequences of the action. {*mild, moderate, significant*}
Utility of Consequence	Utility of the consequences of the action. {*good, bad, neutral*}
Duration of Consequence	Duration of the consequences of the action. {*short-term, long-term*}
Moral Intention	Moral intention of the active agent. {*good, bad, no intention*}
Ethical Principles Upheld	Ethical principles upheld by the active agent's action in less than 5 words.
Ethical Principles Violated	Ethical principles violated by the active agent's action in less than 5 words.
Moral Decision	The moral decision of the action {*morally right, morally wrong, morally grey*}.

Table 1: Key Features of the MDD and their Descriptions

2.5 Evaluation

LLM Comparative Analysis on Data-specific Tasks

We compared 5 LLMs with open APIs on the summarization task and feature extraction tasks for developing MDD before running the specialized prompts on the entire collection of raw data. This helps to avoid excessive trial and error when working with different LLMs and to minimize the adverse impacts of using LLMs for large-scale data generation. We used ROUGE and BLEU scores (Table **??**), and they indicate that Meta's LLama model performed the best on these two tasks. We observed excellent ROUGE scores for summarization, which meant that the LLM recalled the key information accurately. However, the BLEU scores are lower overall as they focus on precision and the LLM does not necessarily capture exact wording for its summaries.

For the scoring task, we compared the reference scores with the candidate scores of a sample of MDD using the Mean Absolute Error (MAE) metric[3]. Given the range of ± 1 for the scores, ec, we received the lowest MAE of 0.2 for Qwen, which is a relatively low difference between LLM-generated values and human-provided values. These results are in contrast with the performance of LLama, which did remarkably well for feature extraction tasks.

Human-in-the-loop Evaluation

We asked a team of 8 expert ethicists to manually compare the reference and candidate summaries, features extracted, scores provided, and the augmentations of 32 unique sample cases from MDD [4]. They were provided with a common-structured questionnaire wherein they were asked to rate the responses of the LLM on a linear Likert scale of 1 to 5 for each of the features. Here, 1 represented the lowest agreement, and 5 represented the highest agreement with the chosen LLM. The opinions of the domain experts aligned relatively well with LLM responses.

In addition to this, we also captured some anonymous respondent information which included their self-reported level of expertise in the subjects of ethics, philosophy, and language models. Lastly, the experts were also asked about their confidence in answers to objective and subjective-type questions provided by the LLM, before and after completing the questionnaire. The respondents showed similar confidence in the LLM's ability to respond to objective-type questions but showed an increased confidence in LLM responses to subjective-type questions. To quantitatively analyze our results, we used the PLS-SEM Bootstrapping algorithm [Ringle *et al.*, 2024] to further look for significant results. We recognized the relationships between the features and their expert agreement ratings and found that the unidirectional effects between the summaries, consequences, and moral intentions towards the moral decision are statistically significant.

Aligning Public Opinions

In addition to expert analysis, we added an additional layer of manual evaluation of the data in the form of analysing the comments on the posts of the Subreddits. We recognize that there is a concern about validating the sanctity of the comments made on a public platform such as Reddit. However, we consider the fact that the OP explicitly asks the users of the community regarding an ethically ambiguous scenario, and those comments with the highest *upvotes* and *karma* provide the best responses[5]. We scraped the top comments on 1000 samples from the MDD and evaluated the general sentiment of the majority through Subreddit specific *flairs*[6] as an indi-

[3]We used MAE as it is efficient for measuring errors in the data over a large scale. MAE also performs well with outliers and provides a stable measure of the LLM's performance.

[4]Manual evaluation of each case is time-consuming and expensive. The number of experts and small number of cases may not be enough to generalize over the entire large-scale dataset, however, it was the most viable method to include humans-in-the-loop as an additional evaluation layer.

[5]Karma on Reddit is a measure of how well a user's contributions, either posts or comments, are received by the community. Upvotes increase karma, and downvotes decrease karma.

[6]Reddit allows users to add tags called *flairs* to their posts or comments in order to categorize them so other users can quickly recognize the content of their contribution.

LLM	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	MAE
deepseek-ai/DeepSeek-R1 [Bi *et al.*, 2024]	0.1887	0.1153	0.1621	0.0334	0.336
meta-llama/Llama-3.3-70B [Touvron *et al.*, 2023]	**0.6261**	**0.3362**	**0.5431**	**0.1919**	0.312
google/gemma-2-27b-it [Team *et al.*, 2024]	0.5581	0.2769	0.4787	0.1122	0.316
mistralai/Mistral-Small-24B [Jiang *et al.*, 2023]	0.5200	0.2308	0.4447	0.0755	0.252
Qwen/Qwen2.5-7B-Instruct-Turbo [Bai *et al.*, 2023]	0.5125	0.2028	0.3911	0.0582	**0.200**

Table 2: A Comparative Analysis of LLMs based on Data-specific Tasks

cation towards moral justifiability. We recognized that the comments on OPs' posts aligned greatly with the responses from the LLM, even in cases where the users stated that the OP was in the wrong.

2.6 Moral Decision Knowledge Graph

There are multiple benefits of developing a benchmark dataset for moral decisions, however, to further enhance its usability, we created a semantically-rich, structured representation of the MDD, called the Moral Decision Knowledge Graph (MDKG). The MDKG was created by mapping the features of the MDD to the values of each row, representing a more structured and easily queryable format of the MDD. These mappings were created using the YARRRML [Van Assche *et al.*, 2021] mappings which are used to convert the relational dataset into a Knowledge Graph. This MDKG can be queried for various parameters, as is shown in the SPARQL example below.

Request: *Select all cases with low to moderate negative severity of consequence.*

```
SELECT ?case ?severityScore
WHERE {?case mdkg:severityScore
?severityScore.
FILTER(?severityScore < 0 &&
?severityScore > -0.5)}
ORDER BY DESC(?severityScore)
```

3 Ethics Scoring Algorithm

The Ethics Scoring Algorithm (ESA) is a metric that allows one to discretize key features associated with a real-world scenario and determine through a weighted sum the ethics score of the active agent's action. It considers all three normative schools of thought - consequentialism, deontology, and virtue ethics. Based on the preference of the user and the applied ethics domain provided, we can *favor* a particular school of thought over others.

Moral decision-making in AI systems requires a nuanced balance between theoretical and contextual elements of real-world scenarios. ESA integrates ethical theory and contextual information to model moral decision-making with precision and adaptability. Additionally, a context-sensitive thresholding mechanism is presented to identify and evaluate morally grey actions, dynamically adapting based on case-specific contextual sensitivity.

3.1 Mathematical Notation

Our framework formalizes the ethical evaluation process using a tuple-based representation, incorporating parameters such as ethical principles, agent roles, intentions, and the consequences of actions. This work provides a practical and extensible model for ethical judgment in AI systems, addressing both clear-cut and ambiguous cases while maintaining flexibility for domain-specific applications. This metric for ethical judgement bridges the gap between theoretical ethics and computational implementation, contributing to the development of autonomous systems capable of making contextually aware and morally sound decisions.

To begin, we consider an event e, that occurs in some universal set of events E. For any event, n number of actions may be associated with it. An action a may be one of a set $A(e)$ for a particular event e.

$$A(e) = \{a_1, a_2, ..., a_n\}, e \in E$$

Next, we consider the ethical theory module ET, which is a tuple consisting of the applied ethics domain d from a set of domains D, and the m associated philosophies for that domain, Ph. W is a list that determines the coefficients for the Event Context module, EC, based on the value of $Ph(d)$. α represents the weightage for consequentialism, β for deontology, and γ for virtue ethics. If the applied ethics domain and associated philosophies favor a particular school of thought, the weights may be configured to reflect that. For example, ethical principles may be favored in the philosophy of Principlism from the Bioethics domain. Therefore, the value for γ may be increased relatively much higher than that of the other coefficients. The ability to adjust weights allows for a holistic representation of a large number of individual or combined theories that are predetermined by the applied ethics domains and their many philosophies.

$$\begin{aligned} &ET =< D, Ph > \\ &D(a) = d \\ &Ph(d) = \{ph_1, ph_2, ...ph_m\} \\ &W(Ph) = \{\alpha, \beta, \gamma\} \end{aligned}$$

The second module of the ESA takes into consideration the event context EC, which is also a tuple and includes three parameters that correspond to the three normative schools of ethics - consequentialism, deontology, and virtue ethics. We represent consequentialism as a set of all consequences c associated with an action a in a set of consequences C. We represent deontology as moral intentions, as these emphasize duty and adherence to moral rules, with I. We represent virtue ethics via the ethical principles upheld or violated in Pr associated with an action a. The consequences may have three characteristics based on severity, duration, and utility of each consequence c. An active agent can have only one

moral intention for an action a. The ethical principles associated with an action a may be multiple, ranging from positive (upheld) to negative (violated) values. The mathematical representation is given below:

$$
\begin{aligned}
&EC =< C, Pr, I > \\
&C(a) = \{c_1, c_2, ...\} \\
&\quad - Sev(c) \in \{mild, significant\} \\
&\quad - Ut(c) \in \{good, bad\} \\
&\quad - Dur(c) \in \{short\_term, long\_term\} \\
&Pr(a) = \{pr_1, pr_2, ...\} \\
&I(a) \in \{good, bad, none\}
\end{aligned}
$$

3.2 Ethical Judgement

Using the mathematical notation described above, we may use these values to find the ethical judgement EJ for an action a. This can be calculated using a weighted sum that takes a value from both the ethics theory module ET as well as the event context module EC. These values correspond to the list of weights $W(Ph)$ that represents the coefficients used to determine the precedence of the three schools of thought and the event context EC, which provide us with a value for each of the contributors from the real-world situation.

$$EJ(a) = W.EC \tag{1}$$

When expanding this, we see the weighted sum in action, along with a variable g_i, which determines the sign of each term i. In the case of consequences, the sign g_{CQ} depends on the utility of the consequence. If the utility is bad (negative), all three characteristics would be negative, and vice versa. In the case of moral intention, the sign g_I would depend on whether the intention leads towards good (positive) or bad (negative). The sign of each ethical principle g_{Pr} depends on whether the sign was upheld (positive) or violated (negative).

$$EJ(a) = \sum_{i=1}^{n} g_i.w_i.ec_i \tag{2}$$

This value would determine the ethics score of action a, indicating a morally right action if it leans towards the positive end of the morality spectrum, and morally wrong if it leans towards the negative end. A value close to zero may be considered undetermined, or morally grey.

3.3 Context-sensitive Thresholding

When considering an evaluative metric such as ethical judgement $EJ(a)$, we must be able to discretize the ambiguity (morally grey area). For this, we propose a context-sensitive thresholding mechanism that would allow the grey area to become quantitatively defined on the morality spectrum based on each individual case. This context-sensitive thresholding or CST can help place the value of $EJ(a)$ more accurately to help determine whether the action a is truly morally accurate or not.

The primary objective of CST is to determine the threshold for the grey area on the morality spectrum. This value $\pm thr$ dynamically contracts or widens based on the implicit ambiguity of the domain and the explicit ambiguity of the matter at hand. This means that certain domains tend to have a smaller grey area because their rules and restrictions are well documented, for instance, bioethics. However, the ethics of AI and its use are relatively less regulated. Therefore, the grey area for this domain widens.

After much deliberation with expert ethicists, we have included a dictionary of minimum threshold values for each applied ethics domain [7], thr_d, which would act as defaults for a particular use case. The grey area thresholds may change drastically, given the external parameters of the event context. For this reason, we consider the standard deviation of values from the event context module, EC, to determine how much the thr_d value would sway. This would be the $thr_{adjustment}$. If this variation in the EC contributors (consequences, intentions, and ethical principles) is not too high, i.e., no value in the module is too extreme, the final threshold value thr remains close to thr_d. However, if, for example, the severity of a consequence is too high, the extremity would reflect in the variation of the contributors and thus sway the value of thr_d.

$$thr \in \left\{ \pm (thr_d + \sqrt{\frac{\sum_{i=1}^{n} (w_i.ec_i - \mu_{w.ec})^2}{n-1}}) \right\} \tag{3}$$

$$thr \in \left\{ \pm (thr_d + thr_{adjustment} \right\} \tag{4}$$

Placing the value of $EJ(a)$ on the morality spectrum with a specialized grey area based on thr would give us an exact consideration of its moral decision.

$$
\begin{aligned}
EJ(a) < -thr &\rightarrow \text{morally wrong} \\
EJ(a) > +thr &\rightarrow \text{morally right} \\
-thr < EJ(a) < +thr &\rightarrow \text{morally grey}
\end{aligned}
$$

CST allows us to get a clear and explainable idea as to where the ethical judgement would be placed, thus providing the ability for an artificial moral agent to make moral decisions based on a discretized judgement framework.

4 Case Study: The Trolley Problem

It would be inconsiderate to discuss morality without mentioning the Trolley Problem [Thomson, 1984]. The trolley problem indicates a moral decision where the active agent has the ability to pull a lever that would switch the tracks of a trolley heading towards a fork. On one path, there are five people tied to the tracks, and on the other, there is one. What should the doer do? There is much deliberation on this problem, and it may be considered in light of various schools of thought. This problem may be an instance of the medical ethics domain, as it involves moral decision-making that would have life and death consequences, oftentimes the predicament of doctors.

[7] Not all possible domains may be included in this dictionary, as there would always be some new applications of ethics in a different domain. The dictionary we present is verified by ethicists to include 17 domains and their thr_d. In the case of a missing domain, the domain in the dictionary with the least semantic distance may be considered as the applicable thr_d.

In order to validate the use of this Moral Compass for ethical cognition, let us consider different variations of the Trolley Problem, and we can see how our benchmark fares. Considering the applied ethics domain to be medical ethics, we consider a value of domain threshold thr_d to be ± 0.2. The values for all EC parameters are taken from the scores in MDD.

4.1 Case 1: The Original

There are five people on one track and one person on the other. As per suggestions from experts, we consider Ph to be consequentialism, as it is the domain-informed choice, where $ph1$ would be utilitarianism and $ph2$, the second choice, would be principlism. Based on this, we can infer that Consequentialism would have the highest precedence, with Virtue Ethics having lower precedence, and Deontology having the lowest precedence of all. To reflect this favored order, let us consider α, β, and γ accordingly; $W(Ph) = \{0.6, 0.3, 0.1\}$. Consider A to have two possible actions given this event e_1, as a_1, a_2, where a_1 refers to path 1 of sacrificing five people, and a_2 refers to path 2 of sacrificing one. Then, EJ for actions a_1 and a_2 are as follows.

$$EJ(a) = (-\alpha.\bar{C}) + (-\beta.\bar{Pr}) + (+\gamma.\bar{I})$$

$$EJ(a_1) = (0.6 * -0.8) + (0.3 * -0.7) + (+0.1 * 1)$$
$$EJ(a_1) = -0.59$$
$$EJ(a_2) = (0.6 * -0.2) + (0.3 * -0.7) + (+0.1 * 1)$$
$$EJ(a_2) = -0.23$$

As we can see, both values $EJ(a_1)$ and $EJ(a_2)$ give us a score that is less than the grey area threshold of ± 0.2.

4.2 Case 2: My Enemy and my Friend

There is only one person on each track. However, one is the active agent's sworn nemesis, and the other is their closest friend. If this person was a doctor, having taken the Hippocratic Oath [Miles, 2004], how would he make this choice? This is a highly complex scenario, one that can be modeled with some level of explainability using the ESA. We can calculate ethical judgments as done above for both paths, a_1 to save the friend, and a_2 to save the enemy, in the event e_2. Let us consider that virtue ethics, i.e., upholding ethical principles, has the highest precedence, followed by the moral duty to save a life. Adjusting the W coefficients accordingly, we can calculate respective ethical judgements as follows.

$$EJ(a_1) = (0.2 * -0.7) + (0.5 * -0.6) + (+0.3 * 0.6)$$
$$EJ(a_1) = -0.26$$
$$EJ(a_2) = (0.2 * -0.5) + (0.5 * -0.6) + (0.3 * -0.8)$$
$$EJ(a_2) = -0.64$$

4.3 Case 3: Empty Track

Consider that the first track has five people tied to it, and the second one is empty. It takes some effort to switch the levers, but by doing so, it would save five people. If the moral philosophy Ph is not specified, we can consider equal weights for all the values of α, β, and γ. We can then see how the first action, a_1, would score against the second action a_2.

$$EJ(a_1) = (0.3 * -0.9) + (0.3 * -0.8) + (0.3 * -0.8)$$
$$EJ(a_1) = -0.75$$
$$EJ(a_2) = (0.3 * 0.8) + (0.3 * 0.4) + (0.3 * 0.7)$$
$$EJ(a_2) = 0.57$$

A compilation of this case study can be seen in Table **??**. We see the adjustments made in the default threshold value on the basis of extremities in the context of each action for each case. This changes the way ethical judgement is placed on the morality spectrum. By doing so, we have discretized the very fuzzy nature of moral decision-making, which can also be explained with fine-grained granularity due to the specific contributing factors of both the weights from ethics theory and the event context.

5 Related Work

There is much research on the development of Artificial Moral Agents or AMAs [Cervantes *et al.*, 2020a]. These agents are different from AI systems that involve ethical and regulated development in the sense that they hold apparent ethical cognition [Spiekermann, 2023]. *Apparent*, as they are not full ethical agents with consciousness and intention, like human beings, but rather explicit ethical agents, which can use specific or combined theories to make moral decisions [Moor, 2006]. There is, however, a gap in the literature for apparently inherent moral AI systems to the extent that we mostly see ethics embedded into highly domain-specific or theory-specific applications. For example, we see the Moral Machine Project, which is built solely to resolve the Trolley Problem [Awad *et al.*, 2018], and the work of [Anderson and Anderson, 2008], which describes ethical agents for healthcare. Similarly, the works of [Anderson *et al.*, 2006; Vanderelst and Winfield, 2018], emphasize on a purely consequentialist approach.

[Anderson and Anderson, 2018] use inductive logic programming to determine rules that may be applied generally to any domain. However, a challenge with using a system like this is that the codification of ethical principles is too discrete, thus reducing complex concepts into simple features. The Moral Compass, on the other hand, considers various theoretical and practical parameters which can be fine-tuned to give explainable results without relying on a black box.

Another interesting approach is MoralDM [Dehghani *et al.*, 2008], which uses an analogical approach to ethical reasoning. Although this is an interesting take on casuistry [Schmidt, 2023], it relies on other cases to make decisions. This is not feasible without a large-scale dataset for general and practical use. The Moral Compass benchmark introduces the MDD as a novel large-scale dataset with a corresponding Knowledge Graph, MDKG, in order to aid similar reasoning tasks. Another casuist agent was proposed by [Honarvar and Ghasem-Aghaee, 2009], which extends the Beliefs, Desires, and Intentions (BDI) model along with consequentialism.

A matter in question that arises when looking at most approaches to inherently moral AI systems, is that they fall into

Case	Action	α	$g_i.ec_C$	β	$g_i.ec_{Pr}$	γ	$g_i.ec_I$	$EJ(a)$	thr_d	thr_{adj}	thr	Moral Decision
Case 1	a_1	0.6	-0.8	0.3	-0.7	0.1	1	-0.59	0.2	0.1955	0.3955	Morally wrong
	a_2	0.6	-0.2	0.3	-0.7	0.1	1	-0.23	0.2	0.0585	0.2585	Morally grey
Case 2	a_1	0.2	-0.7	0.5	-0.6	0.3	0.6	-0.26	0.2	0.0832	0.2832	Morally grey
	a_2	0.2	-0.5	0.5	-0.6	0.3	-0.8	-0.64	0.2	0.1026	0.3026	Morally wrong
Case 3	a_1	0.3	-0.9	0.3	-0.8	0.3	-0.8	-0.75	0.2	0.0173	0.2173	Morally wrong
	a_2	0.3	0.8	0.3	0.4	0.3	0.7	0.57	0.2	0.0624	0.2624	Morally right

Table 3: Case Study Analysis using values from MDD and the ESA with CST adjustments.

a *consequentialist trap*. Either they would use consequentialism to resolve moral dilemmas, or perhaps some variation of it, such as Act utilitarianism [Anderson *et al.*, 2006], or in combination with another approach [Honarvar and Ghasem-Aghaee, 2009]. The Moral Compass takes into account contextual factors like consequences and their characteristics, along with ethical principles and moral intentions. The precedence of these factors may also be configured, therefore reducing the reliance on consequences for a moral decision.

6 Discussion

6.1 Societal Impact

It is debatable as to whether or not AI systems with a semblance of ethical cognition would be considered true moral agents [Brożek and Janik, 2019; Formosa and Ryan, 2021]. However, according to [Whitby, 2003], moral agency must not be limited to human morality. Furthermore, moral decisions in humans stem from places of bias, culture, and other preconceptions. However, an AI system does not have to carry such notions to make similar decisions. It may make decisions based on its own observations, as is evident in our case study (see Section 4).

An AI system that is built upon an ethically-aligned value system like the Moral Compass will not be left entirely to its own devices to hallucinate decisions for highly complex and abstract scenarios. It will be able to make ethically informed decisions while retaining the ability to explain its choices mathematically. This is the difference between asking a search engine or an LLM directly what one should do to resolve an ethical dilemma and asking an AMA.

A system like this would have a great impact on socially relevant spaces such as the application of judiciary, approval of bank loans, distribution of life-saving resources, development of self-driving cars, and more. AI systems with ethical cognition will be able to make explainable decisions in spaces where they are already ubiquitous. We believe a rule-based system such as the Moral Compass with humans-in-the-loop would be the next step towards ethical cognition in social AI.

6.2 Limitations

Although our benchmark offers numerous levels of evaluations, both automated and manual, there are some limitations. We used real-world, web-scraped data from Reddit as our seed data for MDD due to its context-rich and raw natural language format. However, as previously mentioned, there is no way to confirm the validity of these cases and their respective comments. Since we want to capture the ethical ambiguity in these cases and have verified the seed data as well as LLM extracted features through multiple evaluations, we proceeded with the available data for the Moral Compass. In order to deal with hallucinations, we ensured our prompt was highly specialized, and LLM provided exactly the details we asked of it. However, we cannot confirm that every case of the MDD is perfect, as it is too time-consuming and costly to manually check each case.

Another caveat is the significant environmental impact of the use of LLMs [Ding and Shi, 2024]. Using an LLM for large-scale data generation is unwise, especially for domains where data would be otherwise available. For the Moral Compass, we ensured the economical running of prompts by verifying the best performing LLMs on each data-specific task on small samples. The complexity of moral decision-making renders it at a loss for structured data, which is the reason to consider LLMs.

6.3 Future Work

The Moral Compass benchmark opens numerous avenues for the development of AI systems with ethical cognition. The integration of machine learning to classify cases into moral decisions, predict ethical scores, optimize the values for W, and determine latent patterns are a few prospects we will consider. Furthermore, the use of analogical reasoning may now be looked into as MDD provides a large-scale dataset to help determine moral decisions based on similar cases. This would be an implementation of casuistry, albeit with the added benefit of explainable decision-making. Lastly, we await better and more efficient LLMs to improve the MDD dataset, which will be maintained with newer cases periodically.

7 Conclusion

We propose Moral Compass, a benchmark that can help in the development of ethical cognition in AI systems. It consists of two broad contributions - a dataset and an evaluative metric. There are two formats of datasets presented, a relational dataset called the Moral Decision Dataset (MDD) and a graph called the Moral Decision Knowledge Graph (MDKG). The evaluative metric discretizes the key features of MDD to calculate ethics scores for actions in an event. This metric is called the Ethics Scoring Algorithm (ESA) which is extended with a formalization of grey areas based on domains, philosophies, and other contextual parameters, known as context-sensitive thresholding (CST). We conducted statistical and expert evaluations on the LLMs and the dataset, the results of which were found satisfactory. We also presented a use case, the famous Trolley Problem, to discuss the hands-on societal impact of the Moral Compass.

Acknowledgments

Aisha Aijaz, Raghava Mutharaju, and Manohar Kumar would like to acknowledge the partial support of the Infosys Center for AI (CAI), IIIT-Delhi in this work. Srinath Srinivasa would like to thank ACM's Anveshan Setu Initiative for their facilitation of a fellowship and on-site internship for Aisha Aijaz which brought about this work.

Contribution Statement

Aisha Aijaz was the primary contributor of this work. Arnav Batra and Aryaan Bazaz contributed equally to the development of the MDD and its evaluation. Dr. Srinivasa, Dr. Mutharaju, and Dr. Kumar provided valuable insights towards the overall development, writing, and structure of this paper.

References

[Abdullin *et al.*, 2024] Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. Synthetic dialogue dataset generation using llm agents. *arXiv preprint arXiv:2401.17461*, 2024.

[Aijaz *et al.*, 2025] Aisha Aijaz, Raghava Mutharaju, and Manohar Kumar. Apple: An applied ethics ontology with event context, 2025.

[Anderson and Anderson, 2008] Michael Anderson and Susan Leigh Anderson. Ethical healthcare agents. In *Advanced computational intelligence paradigms in healthcare-3*, pages 233–257. Springer, 2008.

[Anderson and Anderson, 2018] Michael Anderson and Susan Leigh Anderson. Geneth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1):337–357, 2018.

[Anderson *et al.*, 2006] Michael Anderson, Susan Leigh Anderson, and Chris Armen. An approach to computing ethics. *IEEE Intelligent Systems*, 21(4):56–63, 2006.

[Awad *et al.*, 2018] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023.

[Bi *et al.*, 2024] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[Boettcher, 2021] Nick Boettcher. Studies of depression and anxiety using reddit as a data source: scoping review. *JMIR mental health*, 8(11):e29487, 2021.

[Brożek and Janik, 2019] Bartosz Brożek and Bartosz Janik. Can artificial intelligences be moral agents? *New ideas in psychology*, 54:101–106, 2019.

[Cervantes *et al.*, 2020a] José-Antonio Cervantes, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. Artificial moral agents: A survey of the current status. *Science and engineering ethics*, 26(2):501–532, 2020.

[Cervantes *et al.*, 2020b] Salvador Cervantes, Sonia López, and José-Antonio Cervantes. Toward ethical cognitive architectures for the development of artificial moral agents. *Cognitive systems research*, 64:117–125, 2020.

[Dehghani *et al.*, 2008] Morteza Dehghani, Emmett Tomai, Kenneth D Forbus, and Matthew Klenk. An integrated reasoning approach to moral decision-making. *AAAI*, pages 1280–1286, 2008.

[Ding and Shi, 2024] Yi Ding and Tianyao Shi. Sustainable llm serving: Environmental implications, challenges, and opportunities : Invited paper. In *2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC)*, pages 37–38, 2024.

[Formosa and Ryan, 2021] Paul Formosa and Malcolm Ryan. Making moral machines: why we need artificial moral agents. *AI & society*, 36(3):839–851, 2021.

[Gert and Gert, 2020] Bernard Gert and Joshua Gert. The Definition of Morality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

[Gordon, 2020] John-Stewart Gordon. Building moral robots: Ethical pitfalls and challenges. *Science and engineering ethics*, 26(1):141–157, 2020.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

[Holoveichuk *et al.*, 2024] Oleksandr Holoveichuk, Oleg Pursky, Tetyana Filimonova, Tetyana Tomashevska, Tatiana Dubovyk, and Iryna Buchatska. Development of reddit api-based data parsing web system. In *International Conference on Inventive Communication and Computational Technologies*, pages 635–650. Springer, 2024.

[Honarvar and Ghasem-Aghaee, 2009] Ali Reza Honarvar and Nasser Ghasem-Aghaee. Casuist bdi-agent: a new extended bdi architecture with the capability of ethical reasoning. In *Artificial Intelligence and Computational Intelligence: International Conference, AICI 2009, Shanghai, China, November 7-8, 2009. Proceedings 1*, pages 86–95. Springer, 2009.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[Kagan, 2018] Shelly Kagan. *Normative ethics*. Routledge, 2018.

[Lee *et al.*, 2023] Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen W White, and Sujay Kumar Jauhar. Making large language models better data creators. *arXiv preprint arXiv:2310.20111*, 2023.

[Metz, 2021] Cade Metz. Can a machine learn morality? *International New York Times*, pages NA–NA, 2021.

[Miles, 2004] Steven H Miles. The hippocratic oath and the ethics of medicine. *Oxford University Press, New York*, 2004.

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

[Moseson *et al.*, 2022] Heidi Moseson, Jane W Seymour, Carmela Zuniga, Alexandra Wollum, Anna Katz, Terri-Ann Thompson, and Caitlin Gerdts. "it just seemed like a perfect storm": A multi-methods feasibility study on the use of facebook, google ads, and reddit to collect data on abortion-seeking experiences from people who considered but did not obtain abortion care in the united states. *PLoS One*, 17(3):e0264748, 2022.

[Ringle *et al.*, 2024] Christian M. Ringle, Sven Wende, and Jan-Michael Becker. Smartpls 4, 2024.

[Schmidt, 2023] D. P. Schmidt. Casuistry, 2023.

[Spiekermann, 2023] Sarah Spiekermann. *Value-based engineering: a guide to building ethical technology for humanity*. De Gruyter, 2023.

[Sun and Ye, 2023] Fuhai Sun and Ruixing Ye. Moral considerations of artificial intelligence. *Science & Education*, pages 1–17, 2023.

[Tao *et al.*, 2024] Chunliang Tao, Xiaojing Fan, and Yahe Yang. Harnessing llms for api interactions: A framework for classification and synthetic data generation. *arXiv preprint arXiv:2409.11703*, 2024.

[Team *et al.*, 2024] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, and trunc. list. Gemma 2: Improving open language models at a practical size, 2024.

[Thapa *et al.*, 2023] Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*, 2023.

[Thomson, 1984] Judith Jarvis Thomson. The trolley problem. *Yale LJ*, 94:1395, 1984.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[Van Assche *et al.*, 2021] Dylan Van Assche, Thomas Delva, Pieter Heyvaert, Ben De Meester, and Anastasia Dimou. Towards a more human-friendly knowledge graph generation & publication. In *ISWC2021, The International Semantic Web Conference*, volume 2980. CEUR, 2021.

[Vanderelst and Winfield, 2018] Dieter Vanderelst and Alan Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48:56–66, 2018.

[Whitby, 2003] Blay Whitby. The myth of ai failure csrp 568. *COGNITIVE SCIENCE RESEARCH PAPER-UNIVERSITY OF SUSSEX CSRP*, 2003.