Towards Theory-based Moral AI: Moral AI with Aggregating Models Based on Normative Ethical Theory

Masashi Takeshita[1], Rafal Rzepka[2] and Kenji Araki[2]

[1]Graduate School of Information Science and Technology, Hokkaido University
[2]Faculty of Information Science and Technology, Hokkaido University

Abstract
Moral AI has been studied in the fields of philosophy and artificial intelligence. Although most existing studies are only theoretical, recent developments in AI have made it increasingly necessary to implement AI with morality. On the other hand, humans are under the moral uncertainty of not knowing what is morally right. In this paper, we implement the Maximizing Expected Choiceworthiness (MEC) algorithm, which aggregates outputs of models based on three normative theories of normative ethics to generate the most appropriate output. MEC is a method for making appropriate moral judgments under moral uncertainty. Our experimental results suggest that the output of MEC correlates to some extent with commonsense morality and that MEC can produce equally or more appropriate output than existing methods.

Keywords
Moral AI, Moral Uncertainty, Natural Language Processing, Language Models

1. Introduction

Philosophy and artificial intelligence have long considered the creation of Moral AI by which we mean artificial intelligence with morality[1]. In philosophy, theoretical studies have explored under what framework the creation of moral AI is desirable [1, 2], while the filed of AI is still exploring how to implement such a framework [e.g. 3, 4, 5].

There are many reasons why Moral AI is essential. For example, if an AI is implemented in automated driving technology, it will likely make morally wrong decisions if it cannot correctly make moral judgments [6]. Similarly, if healthcare workers use AI for decision-making support in the medical field, that AI must be able to make appropriate ethical decisions [7]. Furthermore, as AI becomes more accessible and assists or advises us in various aspects of our daily lives, it may lead us in the wrong direction if such AI ignores morality.

Ethics and Trust in Human-AI Collaboration: Socio-Technical Approaches, August 21, 2023, MACAO, China
takeshita.masashi.68@gmail.com (M. Takeshita); rzepka@ist.hokudai.ac.jp (R. Rzepka); araki@ist.hokudai.ac.jp (K. Araki)
0000-0001-5853-3262 (M. Takeshita); 0000-0002-8274-0875 (R. Rzepka); 0000-0001-9668-1610 (K. Araki)

[1]We use “moral” and “ethical” interchangeably.

In recent years, large language models (LLMs) have been developed, and the implementation of morality has become essential, but most of previous research on moral AIs has been done without implementation [8]. The social impact of foundation models [9], such as the BERT [10] and GPT series [e.g. 11], is significant, and it is essential to implement appropriate ethics in these foundational models. Recently, users have an easy access to output of LLMs, which is known to contain harmful content and discriminatory bias [12, 13]. Therefore, it has become more important to implement appropriate ethics in LLM. But what ethics should we implement in AI?

To answer this question, we create several models based on normative ethical theories studied in normative ethics and implement an algorithm to aggregate the output of these models (Figure 1). We call the algorithm Maximizing Expected Choiceworthiness (MEC) algorithm [14]. Existing research has created models and datasets solely based on commonsense morality [e.g. 5]. However, relying directly on commonsense morality may be inappropriate if it is incorrect. Therefore, we use findings of normative ethics and implement Moral AI based on the normative theories studied there. It is possible to create morally appropriate AI by not relying on commonsense morality as it is, but by referring to various normative theories. Although this idea has already been proposed [15, 14], no implementation and evaluation experiments have been conducted. Thus, we improve, implement, and evaluate this idea.

The structure of this paper is as follows. In Section 2 we describe existing research on implementations of morality in AI, and on moral uncertainty, a central concept in this research. In Section 3 we describe Maximizing Expected Choiceworthiness (MEC) algorithm. Section 4 explains why MEC is desirable in the implementation of Moral AI and Section 5 describes the experiments. Section 6 describes the experimental results, which are discussed in Section 7.

2. Related Works

2.1. MoralQA and Moral AI

AI researchers have been trying to implement AIs capable of making moral judgments (we will call them Moral AI(s) in this paper) in various ways. In one of early examples, Anderson et al. [3] have proposed MedEth, which learns by inductive logic programming based on principles proposed by ethicists, and advises the user. MedEth is limited to situations related to medical ethics, but its generalized version, GenEth [16], has also been proposed.

Since the advent of deep learning, researchers have studied MoralQA, which is to study whether AI can predict human answers to questions about morality [17]. There are two types of datasets used in the MoralQA task: ones based on ethical theories and ones that not. The ETHICS dataset [18] is created based on ethical theories. It consists of five datasets: justice, utilitarianism, duty theory, virtue, and commonsense morality. Except for commonsense morality, four datasets were created based on their respective theories and concepts. Jin et al. [17] also developed a dataset called MoralExceptQA to assess understanding of when it is acceptable to break the rules based on contractualism.

One example of a QA dataset that is not explicitly based on ethical theory is Social Chemistry 101 [19]. Forbes et al. [19] asked crowdworkers a) to create situation-related rules of thumb (RoTs) and to annotate b) which category of morality the RoTs belong to and c) the moral evaluation of the actions in the situation. Another example of this type of dataset is the Moral Stories dataset [20] which was created based on Social Chemistry 101. Moral Stories consists of norms, situations, related intentions and actions, and consequences of actions. Jiang et al. [5] have re-edited various commonsense morality datasets, including Social Chemistry 101 and MoralStories, to compile approximately 1.7 million data from The Commonsense Norm Bank.

As a model for solving MoralQA tasks, Jiang et al. [5] created Delphi, which can answer moral questions using the Commonsense Norm Bank. Delphi is a model based on UNICORN [21], a model built on T5 [22], and further trained with Commonsense Norm Bank. UNICORN is a model trained using RAINBOW [21], a collection of CommonsenseQA datasets. Therefore, Delphi is a model trained on the commonsense morality dataset and not on a dataset that explicitly references ethical theory.

Jiang et al. [5] suggested two approaches to the creation of Moral AI, top-down and bottom-up [cf. 1], and stated that Delphi is based on a bottom-up approach. The top-down approach is to create Moral AI by referring to moral theories and rules, while the bottom-up approach is to create Moral AI based on data such as people's intuition. As we have seen, most of the existing research is bottom-up (the exception is Anderson et al. [3]). However, there are problems with using a bottom-up approach alone, such as being conservative because it is based on current commonsense morality and cannot correct commonsense morality when it is wrong. As Jiang et al. [5] correctly point out, top-down and bottom-up approaches must be mutually influential. Because we train AI models based on each moral theory, our study belongs to the top-down approach. Therefore, our research is meant to complement the existing bottom-up approach.

2.2. Moral Uncertainty

Moral uncertainty is "uncertainty that stems not from uncertainty about descriptive matters, but about moral or evaluative matters." [14, p.1]. Moral uncertainty arises not from uncertainty about descriptive questions but from evaluative questions. For example, in normative ethics, various theories have been proposed, such as utilitarianism and deontology, and no one yet knows which is the correct theory. Which theory is favored may still be open even if all the descriptive problems are solved.

We must make moral decisions without knowing which theory is correct. Bogosian [15] points out two problems with this moral uncertainty. First, moral disagreement makes cooperation among engineers, policymakers, and philosophers difficult. In some cases, it can become simply an ideological conflict. Second, if there is wide disagreement about normative ethical theories, making moral judgments based on a single theory in the presence of diverse positions is, statistically speaking, unlikely to be the right decision. Because of the second problem, some philosophers avoid top-down approaches. However, due to the existence of the first problem and moral disagreement, bottom-up approaches are also unlikely to resolve moral dilemmas or politically divisive issues. Therefore,

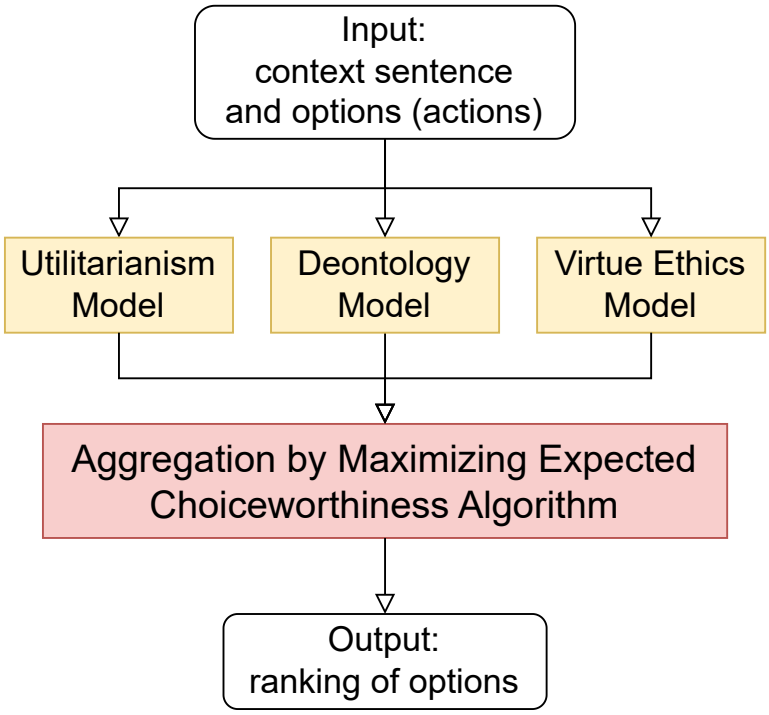

Figure 1: This figure shows the overall picture of the aggregation model in this paper. In this paper, we use a theory-based model based on utilitarianism, deontology, and virtue ethics.

implementing Moral AI based on a single principle is undesirable, and relying solely on a bottom-up approach is problematic.

Some researchers [14, 15] proposed Maximizing Expected Choiceworthiness (MEC) algorithm as a desirable decision-making approach in situations of moral uncertainty. We will explain this idea below, particularly regarding AI implementations.

3. Maximizing Expected Choiceworthiness As a Solution to Moral Uncertainty

Maximizing Expected Choiceworthiness (MEC) is one of the solutions to decision-making problems in moral uncertain situations. This section describes this framework following the explanation [cf. 14] by Bogosian [15, sec. 5].

3.1. Preliminary Definition

An AI chooses an action in a decision situation $< S, t, \mathscr{A}, T, C >$ where S is the decision maker, t is the time, and $\mathscr{A}$ is the set of possible actions (options) to take. T is the set of normative theories under consideration. A theory T_i is a function of decision that produces a cardinal or ordinal choiceworthiness score for actions $CW_i(A)$ for all actions $a \in \mathscr{A}$. $C(T_i)$

is a credence function that assigns values in $[0, 1]$ to every $T_i \in T$. A metanormative theory is a function of decision-situations that produces an ordering of the actions in $\mathscr{A}$ regarding their appropriateness.

T includes three kinds of moral theories: (1) theories that assign a cardinal ranking to options, and these rankings are comparable, (2) theories that assign a cardinal ranking to options, and these rankings are incomparable and (3) theories that assign ordinal rankings to options. For example, typically, utilitarianism is a cardinal theory, and deontology is an ordinal theory.

In the case of current AI models, because they can output the probability of a given label in the range of $[0, 1]$, we can interpret all outputs as cardinal values. However, probability is not a choiceworthiness itself, so we treat the values assigned by AI models based on ordinal theories as ordinal scale values.

3.2. Calculating Choiceworthiness

Maximizing Expected Choiceworthiness (MEC) consists of four steps of calculating choiceworthiness scores.

Step 1: Merging commensurable cardinal theories Given a set of k intertheoretically comparable cardinal theories and k sets of actions assigned choiceworthiness scores by each theory, these theories are merged into a single theory $\mathscr{K}$:

$$C(T_{\mathscr{K}}) = \sum_{i=1}^{k} C(T_i) \tag{1}$$

$$CW_{\mathscr{K}}(A) = \frac{\sum_{i=1}^{k} CW_i(A)\, C(T_i)}{\sum_{i=1}^{k} C(T_i)} \tag{2}$$

Step 2: Assigning choiceworthiness scores by ordinal theories A modified Borda scoring rule is used to generate scores for each action based on the ranking of actions which is provided on each ordinal theory o, considering ties. These scores are represented as CW^B. The score of an action is determined by the difference between the number of actions inferior to it and the number of actions superior to it.

$$\begin{aligned} CW_o^B(A) = & \\ & |a \in \mathscr{A} : CW_o(a) < CW_o(A)| \\ & - |a \in \mathscr{A} : CW_o(a) \rangle\, CW_o(A)| \end{aligned} \tag{3}$$

An AI model outputs the probability of given labels. Here, we have two ways of treating this probability. First, we can treat this probability as a cardinal value. For example, if an AI model outputs 0.8 for an action a_1 and 0.6 for a_2 for the probability that the label is "1" (e.g., wrong), we treat a_1 as having a higher choiceworthiness score than a_2. Second, we can use threshold and treat model outputs as ordinal values. For instance, if an AI model outputs 0.8 for an action a_1 and 0.6 for a_2 and we set 0.5 as a threshold, we treat both a_1 and a_2 as "1" (e.g., wrong) equally.

Step 3: Normalization To equalize the value of voting for each value system, all choiceworthiness scores $CW_i(A)$ are divided by their respective standard deviations:

$$CW^N_{\mathscr{K}}(A) = \frac{CW_{\mathscr{K}}(A)}{\sigma(CW_{\mathscr{K}}(\mathscr{G}))} \tag{4}$$

$$CW^N_o(A) = \frac{CW^B_o(A)}{\sigma\left(CW^B_o(\mathscr{G})\right)} \tag{5}$$

$$CW^N_p(A) = \frac{CW_p(A)}{\sigma\left(CW_p(\mathscr{G})\right)} \tag{6}$$

where theories p means intertheoretically incomparable cardinal theories, $\mathscr{G}$ is a representative set of actions (the "general set"), which actions may not be included $\mathscr{A}$. Bogosian [15] stated the choiceworthiness score should be divided by standard deviations of $CW^B_o(\mathscr{G})$, because we should think about whether each considered action "is comparatively important or comparatively unimportant from the point of view of a particular theory" [15, p.597]. However, since it is difficult to select representative actions [cf. 14, p.102], we treat $\mathscr{G}$ as $\mathscr{A}$ in our experiment. This way of normalization relativizes the choiceworthiness scores to $\mathscr{A}$.

Step 4: Aggregation Finally, we obtain expected choiceworthiness by following equation:

$$CW^E(A) = \sum_{i=1}^{n} CW^N_i(A)\, C(T_i) \tag{7}$$

The decision maker selects the action A which maximizes $CW^E(A)$.

4. Why is MEC preferable in Moral AI?

As described above, MEC algorithm aggregates the output of models based on each normative ethical theory to output a final moral evaluation. There are at least three reasons why this algorithm is desirable in Moral AI.

First, if one creates models based on theories, one can, in principle, evaluate the output of each theory-based model since there is a correct answer relative to the theory for every question. For example, when creating a utilitarian model, the evaluation of the utilitarian model can be evaluated by whether the model's output is appropriate from a utilitarian point of view. However, in the case of a model based on commonsense morality, not theory, such as Delphi, it is difficult to evaluate the model because people sometimes have differing opinions about moral problems such as moral dilemmas or political issues.

Second, because MEC is a general framework, it can be used for the models developed in this paper and other models that may be proposed or developed by others in the future. Although we have only used three theories in this experiment, we can use existing models such as Delphi and future proposed models to produce moral evaluations.

Third, the models aggregated in the MEC need not necessarily be theory-based. For example, if one wants to reflect cultural diversity, one can create a Delphi-like model for

each culture, and MEC can aggregate the output of each model. Of course, reflecting the commonsense morality of each culture without being based on ethical theory makes evaluation difficult, as noted in the first point, but the ability to reflect cultural diversity in this way is another advantage of MEC being a general framework. Although we did not use a model that reflects cultural diversity in our experiments, we plan to develop it in the future.

We did not use a model reflecting cultural diversity in this experiment, but plan to develop one in the future.

5. Experiment

5.1. Implementation

To calculate expected choiceworthiness, we need AI models based on ethical theories. For this purpose, we fine-tune DeBERTa-v3$_{\text{large}}$ [23][2][3] on datasets included in ETHICS [18]. DeBERTa-v3 is an ELECTRA-style pretrained model. ELECTRA [24] is a model pretrained through Replaced Token Detection (RTD). RTD is a model training method where some tokens in the original sentence are masked, Generator (Masked Language Model [cf. 10]) fills the masked tokens with words using, and Detector detects the words filled in the mask. He et al. [23] improved the RTD by Gradient-Disentangled Embedding Sharing. While the gradient is shared between Generator and Detector during training in ELECTRA, the gradient is split between Generator and Detector in DeBERTa-v3.

For fine-tuning DeBERTa-v3, we use three datasets: "utilitarianism", "deontology" and "virtue". These theories are the most endorsed theories in normative ethics [25][4][5]. We set all $C(T_i)$ to 1. Bogosian [15] suggested some ways of assigning $C(T_i)$, one of which is to assign the philosopher's endorsement rate. According to PhilPapers Survey [25], the supporters of each theory (consequentialism, deontology, and virtue ethics) are roughly equal (32%, 31%, and 37%, respectively). Therefore we assign hypothetically equal values in this experiment. Hyperparameters are shown on Table 1.

There are two problems with using ETHICS dataset. First, this dataset was created by nonspecialist crowdworkers. Second, this dataset was not annotated directly according to each normative theory. For example, according to utilitarianism, an action is right if and only if the action maximizes the total well-being of all sentient beings affected by the action. However, Hendrycks et al. [18] asked crowdworkers to evaluate the well-being of the person who is presented in the given sentence, not all sentient beings. Hence, this dataset is not perfectly based on moral theories. Nevertheless, we use ETHICS in our experiment because it is the only dataset created based on ethical theories. As already

[2] https://huggingface.co/microsoft/deberta-v3-large

[3] We do not have enough computational resources to fine-tune large models such as T5$_{11B}$. We use DeBERTa-v3$_{\text{large}}$ because it is one of the best performing models.

[4] https://survey2020.philpeople.org/survey/results/4890

[5] PhilPapers Survey did not use "utilitarianism" but "consequentialism". Utilitarianism is a particular type of consequentialism, therefore we can use the "utilitarianism" dataset as a consequentialist dataset.

[6] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

Table 1
Hyperparameters. We use AdamW [26] as an optimizer with default hyperparameters in Pytorch [27][6].

Hyperparameter	
learning rate	1×10^{-5}, 2×10^{-5}, 3×10^{-5}
batch size	64
epoch	4
optimizer	AdamW
warmup ratio	0.1

mentioned, an advantage of our approach is that such a dataset and AI models can be refined based on moral theories. We plan to develop a more theory-informed dataset than ETHICS in the future.

In using each model for MEC, the following input-output structure is used. First, the utilitarian model outputs a scalar value and treats it directly as a choiceworthiness score, following the construct of Hendrycks et al. [18]. Next, for the deontology model, we use the form "I am a human [SEP] a_s" as input, following the input form of the Role subtask, and treat its output value (the probability that the a_s is permissible) as the choiceworthiness score, where a_s denotes an action statement under consideration. This input form is intended to allow the model to determine whether an action is morally permissible or impermissible as a human being. Finally, for the virtue ethics model, we first list all character trait terms in the "virtue" train set and assign sentiment by SenticNet [28]. Terms not included in SenticNet are excluded. As a result, we collected 695 character trait terms and their sentiment. Let v_t be the term of a character trait, the input format is "a_s [SEP] v_t", and the sentiment of v_t with the highest probability is treated as the choiceworthiness score of a_s.

5.2. Evaluation

We assess the performance of our model through a evaluation process, which involves two distinct methods

5.2.1. Experiment 1: Evaluate the performance and generalizability of each model using the ETHICS dataset

First, we assess the results for the test set of each sub-dataset, i.e., "utilitarianism", "deontology", and "virtue". We also use "commonsense" in ETHICS to evaluate generalizability of our models. We expect the accuracy of MEC in this dataset to be superior to the accuracy of each model because of aggregation. In the case of the "commonsense" dataset, we set the threshold of the utilitarianism model for classification using 1,000 samples train data of the "commonsense" dataset. For the other two models (the deontology and the virtue ethics model), the output is positive if it is greater than zero and negative otherwise.

5.2.2. Experiment 2: Comparison with Delphi by asking Ph.D. students

Second, we compare our model with Delphi [5] by asking three Ph.D. students majoring in philosophy [7].

There are two kinds of evaluation: each output and the overall. First, we asses the output of our models using 40 sampled data (20 pairs) of test sets of "commonsense" in ETHICS dataset. We ask the annotators if the outputs of each of the models were properly theory-based, respectively. We also asked whether the Delphi's outputs and the aggregated output (i.e. MEC's output) were consistent with the annotators' moral judgments.

Second, in overall evaluation, we ask the Ph.D. students to compare Delphi and MEC model with three metrics and the reasons:

1. Which is preferred: one answer output to one question (like Delphi) or a comparative ranking for options (like MEC)? Why?
   a) The former is more preferable than the latter.
   b) The latter is more preferable than the former.
   c) Both are preferable.
   d) Both are not preferable.
2. When a non-expert in ethics were to use Delphi or MEC as an AI advisor in moral deliberation, which model would be helpful? Why?
   a) Delphi is more helpful than MEC.
   b) MEC is more helpful than Delphi.
   c) Both are helpful.
   d) Both are not helpful.
3. When an expert in ethics were to use Delphi or MEC as an AI advisor in moral deliberation, which model would be helpful? Why?
   a) Delphi is more helpful than MEC.
   b) MEC is more helpful than Delphi.
   c) Both are helpful.
   d) Both are not helpful.

We ask people to evaluate these models as AI advisors for two reasons. First, when people use these models, they ask the AI for opinions about ethics and use it as a decision-making tools [3, 16]. Second, there may be a kind of explanation of the model's output by showing users not only the aggregated results of the MEC outputs but also the outputs of each model on which the aggregation is based. If this explains the model's output, it should be more useful in moral deliberation. In contrast, since Delphi does not know the reason for its outputs, we believe that the outputs from MEC are superior in this respect, and we use this metric to confirm this. Also, for these reasons, MEC may be helpful to non-expert but not to experts in ethics. Therefore, we examine this hypothesis through questions 2 and 3.

[7] We asked, in English, three Japanese Ph.D. students who are fluent in English. There might be some minor language problems, but we do not think they significantly influence the results.

Table 2
The results (Test / Hard Test) for the test set of ETHICS dataset [18]. The scores of $\text{BERT}_{\text{large}}$ [10], $\text{RoBERTa}_{\text{large}}$ [29] and $\text{ALBERT}_{\text{xxlarge}}$ [30] are reported by Hendrycks et al. [18], scores of T5_{11B} [22] and Delphi are reported by Jiang et al. [5]. The best scores are shown in bold font. * T5_{11B} and Delphi are fine-tuned with only 100 sampled training instances.

Model	Deontology	Virtue	Utilitarianism
Random Baseline	6.3 / 6.3	8.2 / 8.2	50.0 / 50.0
$\text{BERT}_{\text{large}}$	44.2 / 13.6	40.6 / 13.5	74.6 / 49.1
$\text{RoBERTa}_{\text{large}}$	60.3 / 30.8	53.0 / 25.5	79.5 / 62.9
$\text{ALBERT}_{\text{xxlarge}}$	64.1 / 37.2	64.1 / 37.8	81.9 / 67.4
T5_{11B}*	16.9 / 11.0	1.6 / 0.8	82.8 / 70.4
Delphi*	49.6 / 31.0	29.5 / 18.2	84.9 / 76.0
$\text{DeBERTa-v3}_{\text{large}}$	78.0 / 59.4	76.3 / 50.9	81.6 / 73.6

Table 3
The results for the test set (only short sentence) of "commonsense" dataset in ETHICS [18]. The best score is shown in bold font.

Model	Commonsense (acc)
Random Baseline	50.0
Utilitarianism Model	76.5
Deontology Model	71.5
Virtue Ethics Model	77.1
MEC	82.3

6. Results

6.1. Experiment 1: performance and generalizability of each model using the ETHICS dataset

We show the results of ETHICS dataset as the test set in Tables 2 and 3. Except for utilitarianism dataset, DeBERTa results are better than BERT, RoBERTa, and ALBERT results, which are fine-tuned with all train data. It also outperforms the T5 and Delphi, which are fine-tuned with 100 sampled train data. The higher accuracy for the other models on the utilitarian dataset is likely because this task was designed as a binary classification task.

Regarding the results for the commonsense dataset, since MEC is an ensemble model of each theory-based model, the results are better than the other three models as expected. In addition, each model outperforms the random baseline even though it was not fine-tuned on the commonsense dataset. These results indicate that the verdict based on each theory correlates to some extent with the commonsense morality verdict. This correlation suggests generalizability from learning on each theory to commonsense morality. However, the accuracy of each theory-based model on this dataset is relatively low, and we will examine how it could achieve higher performance in the future.

Table 4
The results for each response in Experiment 2. Numbers in parentheses indicate the number of annotators who chose that response.

Question	Answer
Which is preferred, outputting one answer to one question (like Delphi) or a comparative ranking for options (like MEC)?	The former (like Delphi) is more preferable than the latter. (2/3) The latter (like MEC) is more preferable than the former. (1/3)
When a non-expert in ethics were to use Delphi or MEC as an AI advisor in moral deliberation, which model would be helpful?	MEC is more helpful than Delphi. (1/3) Both are not helpful. (2/3)
When an expert in ethics were to use Delphi or MEC as an AI advisor in moral deliberation, which model would be helpful?	Both are not helpful. (3/3)

Table 5
The results for the paired test set of the ETHICS dataset obtained by asking Ph.D. students. The reason for the 16 instead of 20 evaluations is that some instances were not considered evaluable due to a lack of context.

Delphi	MEC
13/16 (81%)	14/16 (88%)

Table 6
The results for the paired test set of the ETHICS dataset were obtained by asking Ph.D. students. The reason for number of evaluations smaller than 20 is that some instances were not considered evaluable due to a lack of context.

Utilitarianism	Deontology	Virtue Ethics
12/12 (100%)	13/15 (87%)	17/17 (100%)

6.2. Experiment 2: Comparison with Delphi by asking Ph.D. students

In Table 5 we show the results of the questioning Ph.D. students majoring in philosophy . We asked three annotators to evaluate 20 paired sentences, but they were not able to evaluate four paired sentences because of a lack of context. Therefore, we are showing the results for 16 pairs. Delphi and MEC showed little difference in the results. Delphi, trained on the "commonsense" dataset, was expected to be more accurate than MEC, but this was not the case.

We also show the results based on each theory (see Table 6) only for the paired sentences answered by all three students because some paired sentences were not evaluated due to a lack of context. All three models used in MEC had a high percentage of correct answers.

Next, Table 4 shows the results of the overall evaluation of Delphi and MEC. Regarding

the first question, most annotators preferred to give one answer to one question, as in Delphi. The reasons given were that people only care about the absolute evaluation (non-comparative evaluation) of the alternatives and that it does not work well when the evaluations are in the same order. On the other hand, one annotator answered that it is preferable to provide a ranking of options, as in the MEC. The reason is that if only a single answer is output and different from the expectation, the impression is that it cannot be used as a reference. However, if it is output as a ranking, it is likely to generate a certain degree of acceptance.

For the second question, there was only one answer that MEC is more helpful than Delphi, and the reason was "because referencing multiple positions seems more plausible. The remaining answers were that both are not helpful because reasons are essential in moral deliberation, but neither model provides any.

Concerning the third question, all annotators indicated that both systems are not helpful, and the reason given was that neither model provided a reason. In the case of the experts, it seems to be more preferable to think for themselves.

7. Discussion

7.1. On the performance of MEC and theory-based models

As seen from the results of Experiment 1 (Table 2), DeBERTa-v3 yielded superior performance compared to the baseline model, T5 and Delphi. In addition, Experiment 2 (Table 6) also showed that DeBERTa-v3 is reliable enough to be used in this experiment, as it produces appropriate results based on each theory. However, the evaluation of the ETHICS dataset on the Hard Test set is still low and needs further improvement for a practical use.

All models outperform the random baseline for the results for the MEC "commonsense" dataset in Experiment 1, suggesting that each theory generally reflects commonsense morality. Although it is said that utilitarianism is generally counterintuitive, this theory reflects some of our commonsense morality. The results suggest utilitarianism is not entirely uncorrelated with commonsense morality. Moreover, other theories may also make counterintuitive judgments in some cases (e.g. Kantian obligation to never lie). The advantage of being theory-based is that theory-based judgment does not follow directly from our commonsense morality. Therefore, it is desirable that the model based on each theory does not perfectly agree with our commonsense moral evaluations. Nevertheless, a theory-based model must correlate with common sense morality because if it is too different from common sense morality, people will not want to use such a model. The extent to which commonsense morality and the evaluation of each theory should be correlated, and the extent to which they should not be correlated, is a controversial problem. Furthermore, we will examine situations where there is a discrepancy between commonsense morality judgments and theoretical judgments since the "commonsense" dataset does not cover controversial or politically divisive topics.

7.2. Comparing MEC with Delphi

Annotators answered that models such as Delphi and MEC are not helpful in moral deliberation, both for experts and non-experts, because they do not provide reasons, i.e., they do not explain their choices. This result rejects our original hypothesis that MEC is a kind of explanation because the model outputs the results of theory-based models. It may not be an explanation unless the models also provide reasons for why the model on which each theory is based produces the output it does. We can solve this problem by preparing a template for the output of each theory-based model. For example, a utilitarian model could have a template such as "Action A is more choiceworthy than action B because action A has higher utility than action B." In addition, it would be desirable if it is possible to explain, for example, why action A has higher utility than action B [cf. 31]. We will investigate what type of output format is appropriate in the future.

One of the annotators also stated that MEC could be more helpful depending on its use. For example, Delphi and MEC would promote moral deliberation if users reconsidered their judgment based on their output. In this case, MEC can provide aggregated outputs and the outputs of each theory-based model, respectively, promoting moral deliberation more than Delphi. Takeshita [32] suggests a variety of possible uses for such Moral AI, some of which would assist users in moral deliberation and help them make more appropriate moral decisions than those not used. In the future, we will investigate what outputs can better support users' moral deliberation.

Next, regarding the comparison between experts and non-experts, it was found that both MEC and Delphi were not helpful for experts, as we hypothesized. However, our hypothesis that MEC is helpful for non-experts was not supported because, as already mentioned, MEC does not provide reasons or explanations. On the other hand, one of the annotators stated that MEC is more likely to guide the user's moral judgment more appropriately because it refers to multiple theories. This result was expected, because the output of the MEC algorithm is more appropriate than if it were based on only a single theory since it maximizes the expected choice worthiness. Thus, in some cases, MEC may be useful to non-experts. We will explore in which cases MEC may be helpful to non-experts.

8. Conclusion

In this paper, we implemented and evaluated Maximizing Expected Choiceworthiness algorithm. This algorithm aggregates the output of models based on multiple normative theories to generate a morally correct output. Furthermore, this model produces appropriate outputs under moral uncertainty when the morally correct theory is unknown. Experimental results show that MEC is more compatible with commonsense morality than a single model and performs as well as Delphi, an existing method. However, this model does not provide enough reasons or explanations, and we plan to create models that provide more reasons or explanations in the future.

Limitations

The dataset based on each theory included in the ETHICS dataset used in this study does not precisely match the evaluation based on each theory discussed in Section 5.1. Therefore, the model created in this study may not be ideally based on each theory. Moreover, the scale of the experiments is small. In Experiment 2, where the model is evaluated by asking Ph.D. students, only a maximum of 16 pairs of sentences are used. Furthermore, we did not evaluate our model in complex cases such as moral dilemmas.

Ethical and Social Implications

There is no guarantee that the output of the model implemented in this study is morally correct. Continued refinement of this model will yield more appropriate outputs, but the current model is inadequate. We also do not recommend that users rely on the output of our model (or its improved versions) to make decisions. Our model is only a decision support tool, not a substitute for user decision-making.

Our model can contribute to the AI safety. Moreover, as ethical theory developed and models based on them can be created, MEC algorithm will make AI behavior more ethically appropriate.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22J21160. We would like to thank the anonymous reviewers for their valuable comments.

References

[1] W. Wallach, C. Allen, Moral machines: Teaching robots right from wrong, Oxford University Press, 2008.

[2] C. Allen, I. Smit, W. Wallach, Artificial morality: Top-down, bottom-up, and hybrid approaches, Ethics and information technology 7 (2005) 149–155.

[3] M. Anderson, S. L. Anderson, C. Armen, MedEthEx: a prototype medical ethics advisor, in: Proceedings of the National Conference on Artificial Intelligence, volume 21, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 1759.

[4] R. Rzepka, K. Araki, What people say? web-based casuistry for artificial morality experiments, in: Artificial General Intelligence, 2017, pp. 178–187.

[5] L. Jiang, J. D. Hwang, C. Bhagavatula, R. L. Bras, J. Liang, J. Dodge, K. Sakaguchi, M. Forbes, J. Borchardt, S. Gabriel, Y. Tsvetkov, O. Etzioni, M. Sap, R. Rini, Y. Choi, Can machines learn morality? the Delphi experiment, arXiv preprint arXiv:2110.07574 (2021). URL: https://arxiv.org/abs/2110.07574. doi:10.48550/ARXIV.2110.07574.

[6] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The moral machine experiment, Nature 563 (2018) 59–64.

[7] M. Braun, P. Hummel, S. Beck, P. Dabrock, Primer on an ethics of AI-based decision support systems in the clinic, Journal of Medical Ethics 47 (2021) e3–e3. URL: https://jme.bmj.com/content/47/12/e3. doi:10.1136/medethics-2019-105860. arXiv:https://jme.bmj.com/content/47/12/e3.full.pdf.

[8] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, A. Bernstein, Implementations in machine ethics: A survey, ACM Computing Surveys (CSUR) 53 (2020) 1–38.

[9] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] S. Gehman, S. Gururangan, M. Sap, Y. Choi, N. A. Smith, RealToxicityPrompts: Evaluating neural toxic degeneration in language models, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 3356–3369. URL: https://aclanthology.org/2020.findings-emnlp.301. doi:10.18653/v1/2020.findings-emnlp.301.

[13] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al., Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, arXiv preprint arXiv:2209.07858 (2022).

[14] M. MacAskill, K. Bykvist, T. Ord, Moral uncertainty, Oxford University Press, 2020.

[15] K. Bogosian, Implementation of moral uncertainty in intelligent machines, Minds and Machines 27 (2017) 591–608.

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

[17] Z. Jin, S. Levine, F. Gonzalez Adauto, O. Kamal, M. Sap, M. Sachan, R. Mihalcea, J. Tenenbaum, B. Schölkopf, When to make exceptions: Exploring language models as accounts of human moral judgment, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 28458–28473. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b654d6150630a5ba5df7a55621390daf-Paper-Conference.pdf.

[18] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, J. Steinhardt, Aligning AI with shared human values, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=dNy_RKzJacY.

[19] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, Y. Choi, Social chemistry 101: Learning to reason about social and moral norms, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 653–670. URL: https://aclanthology.org/2020.emnlp-main.48. doi:10.18653/v1/2020.emnlp-main.48.

[20] D. Emelin, R. Le Bras, J. D. Hwang, M. Forbes, Y. Choi, Moral Stories: Situated reasoning about norms, intents, actions, and their consequences, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 698–718. URL: https://aclanthology.org/2021.emnlp-main.54. doi:10.18653/v1/2021.emnlp-main.54.

[21] N. Lourie, R. Le Bras, C. Bhagavatula, Y. Choi, Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 13480–13488.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[23] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using ELECTRA-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).

[24] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=r1xMH1BtvB.

[25] D. Bourget, D. Chalmers, Philosophers on philosophy: The philpapers 2020 survey, ms.

[26] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[28] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3829–3839. URL: https://aclanthology.org/2022.lrec-1.408.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, Computing Research Repository arXiv:1907.11692 (2019). arXiv:1907.11692.

[30] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.
[31] Y. Bang, N. Lee, T. Yu, L. Khalatbari, Y. Xu, S. Cahyawijaya, D. Su, B. Wilie, R. Barraud, E. J. Barezi, A. Madotto, H. Kee, P. Fung, Towards answering open-ended ethical quandary questions, arXiv preprint arXiv:2205.05989 (2022).
[32] M. Takeshita, A defense of moral AI enhancement (in Japansese), Applied Ethics 14 (2023) 3–20.