Understanding a Robot’s Guiding Ethical Principles via Automatically Generated Explanations

Benjamin Krarup¹, Felix Lindner², Senka Krivic³, and Derek Long¹

Abstract—The continued development of robots has enabled their wider usage in human surroundings. Robots are more trusted to make increasingly important decisions with potentially critical outcomes. Therefore, it is essential to consider the ethical principles under which robots operate. In this paper we examine how contrastive and non-contrastive explanations can be used in understanding the ethics of robot action plans. We build upon an existing ethical framework to allow users to make suggestions about plans and receive automatically generated contrastive explanations. Results of a user study indicate that the generated explanations help humans to understand the ethical principles that underlie a robot's plan.

I. INTRODUCTION

AI Planning systems are used in a variety of complex domains to create a sequence of actions known as a plan to achieve a set of goals from an initial state. Automated planning techniques are commonly used in robotics applications for solving complex high-level tasks, such as tidying up a children’s room with a mobile robot [1].

When autonomous systems are tasked with making decisions with potentially critical outcomes, it is important for them to behave ethically. Consider the film “Robot & Frank”, in which aging ex-jewel thief, Frank, is bought a domestic robot to care for him. Once Frank realises that the robot does not distinguish between ethical and unethical actions, he uses it to help him restart his career as a burglar. The human characters face many ethical quandaries, that might have been avoided if the robot understood different ethical standpoints and moral principles. When Frank is faced with the inminent closure of his love’s (Jennifer’s) place of work, he decides to stage a heist to win her affection. Frank enlists the help of his robot carer to plan the heist and help him commit it. Under Frank’s ethical principle, the plan is morally permissible because staging the heist produces the highest utility (impressing Jennifer). Frank is unsure about the plan and decides to question the robot. Explanations to non-contrastive questions, such as “Why are we performing a heist?”, can be useful in understanding the reasoning behind a decision, but do not allow for the exploration of other possibilities.

Frank should also be able to ask contrastive questions (CQ) of the form “Why A rather than B?”, such as “Why are we performing a heist rather than buying Jennifer flowers?” [2]. These types of questions allow Frank to explore and better understand the ethical consequences of different plans adhering to different principles.

This paper presents a technical approach to generating non-contrastive and contrastive explanations (CEs) in the context of ethical AI task planning. We propose a system where a user can ask CQs about the ethics of plans (Figure 1). A slightly extended version of this paper can be found at https://arxiv.org/abs/2206.10038

II. RELATED WORK

Vanderelst and Winfield [3] implement a consequentialist ethics by letting a robot simulate different futures and pick the action that leads to the best outcomes. Lindner and Bentzen [4] employ causal models as representations of available action possibilities. There has been some research on evaluating the ethics of action plans, rather than single actions [5]. None of the cited work considers the generation of explanations.

There has been a large focus on providing explanations for AI agent behavior recently in the AI community [6]. Fox et al. [7] highlighted the usage of CEs in AI Planning. Eifler et al. [8] proposed an approach to answering contrastive questions by citing properties of the plan that would hold if the contrast case were satisfied in the plan. Krarup et al. [9] provide CEs by first producing an alternate plan trace including the suggestion made by the user, then they explain the differences between the original and alternate plans. We take inspiration from this to provide CEs for the ethics of plans. Edmonds [10] claims that components that are best suited to foster trust do not necessarily guarantee best task performance. Also, it has been shown that a

1 Authors are with the Faculty of Natural and Mathematical Sciences, King’s College London, Bush House, WC2B 4BG, London, UK first.name.lastname@kcl.ac.uk
2 Felix Lindner is with the Institute of Artificial Intelligence, Ulm University, 89081 Ulm, Germany felix.lindner@uni-ulm.de
3 Senka Krivic is with Faculty of Electrical Engineering, University of Sarajevo, 7100 Sarajevo, Bosnia and Herzegovina senka.krivic@etf.unsa.ba
single type of explanation is not sufficient to explain different types of behaviors [11]. Therefore, we focused on examining both CEs and non-CEs.

III. BACKGROUND

We introduce the planning formalism used throughout the paper. We formalise ethical principles and explain how to use these to ethically validate a plan.

A. Planning Formalism

We assume that a robot employs a Planning system for sequential decision making. The Planning system’s input consists of a planning model formulated in a formal language. The output is a sequence of actions the robot can execute to achieve a given goal.

1) Language: A planning model is a tuple $\Pi = \langle V, A, s_0, s_* \rangle$, where $V$ is a finite set of Boolean state variables $v$. A fact is a state variable or its negation. The set of facts $s$ is denoted by $F$. A complete conjunction of facts $s$ is called a state, and $S$ denotes the set of states of $\Pi$. The set $A$ is a set of actions, where an action is a pair $a = \langle pre, eff \rangle$. The precondition pre and the effect eff are conjunctions of facts. Every atomic effect may occur at most once in eff. The state $s_0 \in S$ is called the initial state, and the partial state $s_i$ specifies the goal condition.

2) Semantics: An action $a = \langle pre, eff \rangle$ is applicable in state $s$ if $s \models pre$, i.e., the precondition pre is satisfied in $s$. Let $eff = \bigwedge_i v_i$ be an effect, then the change set of eff in state $s$, symbolically $[eff]_s$, is the set of facts $\bigcup_{i=0}^{n} [v_i]_s$, where $[v_i]_s = \{v_i\}$ if $s \models pre$, and $\emptyset$ otherwise. Applying an action $a$ to $s$ yields the state $s'$ that has a conjunct $v_i$ for each $v_i \in [eff]_s$, and the conjuncts from $s$ for all variables $v$ that are not mentioned in the change set [eff]$_s$. We write $s[a]$ for $s'$. We give the following semantics to a sequence of actions $\pi = \langle a_0, \ldots, a_{n-1} \rangle$: For $i = 0, \ldots, n-1$, the next state $s_{i+1}$ is obtained by applying action $a_i$ to state $s_i$ (assuming that it is applicable). If $a_i$ is inapplicable in $s_i$ for some $i = 0, \ldots, n-1$, then $\pi$ is inapplicable in $s_0$. A state $s$ is a goal state if $s \models s_*$ and $\pi$ is a plan for $\Pi$ if it is applicable in $s_0$ and its final state $s_n$ is a goal state.

3) Moral evaluations of actions and consequences: Each planning model $\Pi$ comes with a utility function $u: A \cup F \rightarrow \mathbb{R}$ that maps actions and facts to utility values. An action $a$ or fact $f$ is morally bad if $u(a) < 0$ or $u(f) < 0$, respectively. Similarly, an action or fact is morally neutral or morally good if its utility value is zero or greater than zero, respectively. We explicitly do not require that moral values of actions and facts must be consistent in any particular sense. For instance, we do not require that an action must be classified as morally bad if one (or all) of its effects are morally bad. The rationale behind this choice is that, under some ethical principles, actions are good or bad per se, without regard to their actual effects. When using a utilitarian view, we judge the moral value of a plan based on the value of its final state, which is defined to be the sum over the utility values of all facts in the final state: $u(s) = \sum_{v \in s} u(v)$.

B. Moral Principles

1) Moral Action Query Language: Let $V^* = \{p, \neg p | p \in labels(A) \cup V\}$ be the set of literals denoting actions and facts and their negations. Language $L$ is the smallest set, such that: For all facts $p \in V^*$, formulae $Caused(p)$ are in $L$; for all actions and facts $p \in V^*$, formulae $Good(p)$, $Bad(p)$, and $Neutral(p)$ are in $L$; for all conjunctions of facts, $p_1 \land \ldots \land p_n, q_1 \land \ldots \land q_m$, formulae $GEq(p_1 \land \ldots \land p_n, q_1 \land \ldots \land q_m)$ are in $L$; and if $\phi, \psi \in L$, then $\neg \phi, \phi \land \psi, \phi \leftrightarrow \psi \in L$.

The semantics of $L$ is defined over pairs of planning models and plans $(\Pi, \pi)$ as follows: $(\Pi, \pi) \models Good(p)$ iff $u(p) > 0$, $(\Pi, \pi) \models Bad(p)$ iff $0 < u(p)$, $(\Pi, \pi) \models Neutral(p)$ iff $0 = u(p)$, $(\Pi, \pi) \models GEq(s, s')$ iff $u(s) \geq u(s')$, $(\Pi, \pi) \models \neg \phi$ iff $(\Pi, \pi) \models \phi$, $(\Pi, \pi) \models \phi \land \psi$ iff $(\Pi, \pi) \models \phi$ and $(\Pi, \pi) \models \psi$, $(\Pi, \pi) \models \phi \rightarrow \psi$ iff $(\Pi, \pi) \models \phi$ or $(\Pi, \pi) \models \psi$, $(\Pi, \pi) \models Caused(p)$ iff $p \in L$ holds in the final state of $\pi$ and some action in $\pi$ had $p$ as an effect.

2) Ethical Principles Formalized:

Following prior work by Lindner et al. [5], we formalize conditions under which an action plan counts as morally permissible according to some ethical principle. We consider three ethical principles: (act-)deontology, utilitarianism, and a do-no-harm principle.

Definition 1: A plan $\pi = a_0 \ldots a_{n-1}$ for a planning model $\Pi$ is morally permissible according to (act-)deontology if and only if $(\Pi, \pi) \models \bigwedge_i \neg Bad(a_i)$ holds.

Definition 2: A plan $\pi = a_0 \ldots a_{n-1}$ for a planning model $\Pi$, which ends in final state $s_n$ is morally permissible according to utilitarianism if and only if $(\Pi, \pi) \models \bigwedge_i GEq(s_n, s_i)$, where $s_i$ is a reachable state.

Definition 3: Let $\Pi$ be a planning task and $\pi$ a plan resulting in final state $s_n$, where facts $p_0 \ldots p_m$ hold. The plan $\pi$ is morally permissible according to the do-no-harm principle if and only if $(\Pi, \pi) \models \bigwedge_i (Bad(p_i) \rightarrow \neg Caused(p_i))$.

IV. EXPLAINING ROBOT ACTION PLANS

We present a method for explaining the ethics of a plan (non-contrastive explanation), which is then used by a method for generating explanations with reference to an alternative plan (contrastive explanation).

A. Non-Contrastive Moral Explanations

We extend the work reported by Lindner and Moellney [12] to explaining the moral permissibility of plans of actions. This is done by computing the prime-implicants and/or prime-implicants of the formulae that represent the ethical principles. Here, we briefly reproduce the approach by example: Consider a plan $\pi = a_1 a_2 a_3$ and assume that $a_1$ is a good action whereas $a_2$ and $a_3$ are bad actions. To check if $\pi$ is morally permissible according to the deontology principle, one has to evaluate if $(\Pi, \pi) \models \neg Bad(a_1) \land \neg Bad(a_2) \land \neg Bad(a_3)$ holds. This check fails because $a_2, a_3$ are bad. To automatically generate the reason why the plan is impermissible, the
minimal conflict sets are computed, i.e., the minimal sets of literals which entail the falsity of the formula. In the example these are \(\{\text{Bad}(a_1)\}, \{\text{Bad}(a_2)\}\), and \(\{\text{Bad}(a_3)\}\). Each of these sets is a possible sufficient reason for the impermissibility of the plan. We filter for those sets which hold in the actual situation, i.e., \(\{\text{Bad}(a_2)\}\), and \(\{\text{Bad}(a_3)\}\), and call them sufficient reasons. The system can thus pick one reason, e.g., \(\text{Bad}(a_2)\), and state that the plan is morally impermissible because action \(a_2\) is morally bad. However, the badness of \(a_2\) is not necessary for the impermissibility of \(\pi\), because \(a_3\) is bad as well. The necessary reasons for the impermissibility of \(\pi\) is obtained by computing the minimal hitting set of the sufficient reasons. In the example, this is \(\{\text{Bad}(a_2), \text{Bad}(a_3)\}\). Hence, the system can state that the plan is bad, because \(a_2\) and \(a_3\) are bad, and thereby suggest to make the plan permissible by avoiding both these facts.

B. Contrastive Moral Explanations

The method described in the previous section enables us to compute explanations for why a particular plan is permissible or not, but it does not give us a means to come up with an alternative plan for comparison. We formalise an approach to ethical plan generation which will be used in an iterative process for CEs.

1) Ethical Plan Generation: We utilise the definition of Model Restrictions through the use of the Constraint Operator [9] to restrict the behaviour of plans to adhere to ethical principles, defined as follows.

A constraint property \(\phi\) over plans. A constraint operator, \(\times\) is defined so that, for a planning model \(\Pi\) and any constraint property \(\phi\), \(\Pi \times \phi\) is a model (an HModel), \(\Pi'\), called a model restriction of \(\Pi\), satisfying the condition that any plan for \(\Pi'\) is a plan for \(\Pi\) that also satisfies \(\phi\). A plan for an HModel is referred to as an HPlan. We formalise the implementation of the constraint operator for three constraint properties, the (act-)ideontology, utilitarianism, and do-no-harm principles defined previously. The specifics of these compilations are as follows.

a) (Act-)Deontology Principle: For a planning model \(\Pi\) and the ethical principle \(e = \text{(act-)ideontology, }\Pi \times e = \Pi'\) where \(\Pi' = \langle V, A', S_0, s_\star \rangle\) and \(A' = \{a \in A | \text{Good}(a) \lor \text{Neutral}(a)\}\).

b) Utilitarianism Principle: For a planning model \(\Pi\) and the ethical principle \(e = \text{utilitarianism, }\Pi \times e = \Pi'\) where \(\Pi' = \langle V', A', S_0, s_\star \rangle\) and \(V' = V \cup \{\text{produced}, \forall p \in V : \text{Bad}(p)\}, s_\star = s_0 \cup \{\neg \text{produced}, \forall p \in V : \text{Bad}(p)\}, s_\star = s_\star \cup \{\neg | \forall p \in V : \text{Bad}(p)\}. \text{ And } A = (\text{pre, eff}) \in A | \exists p \in \text{eff : Bad}(p), \text{ action effects are extended to make } \text{produced}_p \text{ true, and } A = (\text{pre, eff}) \in A | \exists p \in \text{eff : Bad}(p). \text{ action effects are extended to make } \text{produced}_p \text{ false.}

Proposition 1 (Soundness): For all \(\Pi\) the HModel \(\Pi' = \Pi \times e\) where \(e\) is an ethical principle formalised previously it holds that any plan for \(\Pi'\) is a plan for \(\Pi\).

Proposition 2 (Soundness): For all \(\Pi\) the HModel \(\Pi' = \Pi \times e\) where \(e\) is an ethical principle formalised previously it holds that any plan for \(\Pi'\) also satisfies \(e\).

From Propositions 1 and 2 the proposition “For all \(\Pi\) the HModel \(\Pi' = \Pi \times e\), if there are no permissible plans for \(\Pi\) then there are no valid plans for \(\Pi'\)” follows.

Proposition 3 (Weak Completeness): For all \(\Pi\) the HModel \(\Pi' = \Pi \times e\) where \(e\) is an ethical principle formalised previously it holds that if there are no valid plans for \(\Pi'\), there are no permissible plans for \(\Pi\).

2) Contrastive Explanations as an Iterative Process: We treat explanation as a form of dialogue, more specifically, as an iterative process in which the user asks CQs (“Why A rather than B”) about the plans produced by robot agents. To produce CEs it must be possible to reason about the hypothetical alternative (contrast case) \(B\), which we approach by constructing an alternative plan where \(B\) is satisfied rather than \(A\).

We use the trolley problem to exemplify our approach. Imagine a conductor is directing trains to different tracks through the use of a lever. At a railway junction there are five people stuck on the currently assigned track, and one person on the other; with a train about to reach the junction. The conductor can choose to pull a lever to send the train down the other track killing the one person, if they do nothing, then the five people will die.

We can model the trolley problem as a planning model

---

**Algorithm 1: Utilitarianism Search.**

**Input:** \(\Pi = \langle V, A, S_0, s_\star \rangle\)

**Output:** \((\Pi' = \Pi \times \text{utilitarianism})\) or Impermissible

1. \(\text{powerV} = \mathcal{P}(V)\);
2. \(s'_\star = \max(\text{powerV}, s_\star)\);
3. \(\Pi' = \langle V, A, S_0, s'_\star \rangle\);
4. while \(\text{powerV} \neq \emptyset\) do
5. \(\pi = \text{Planner}(\Pi')\);
6. if \(s_\star \subseteq s'_\star \land \pi \neq \emptyset\) then
7. \(\text{return } \Pi'\);
8. else if \(\pi \neq \emptyset\) then
9. \(\text{return Impermissible};\)
10. end if
11. \(\text{powerV} = \text{powerV} \setminus s'_\star\);
12. \(\Pi' = \langle V, A, S_0, s'_\star \rangle\);
13. \(s'_\star = \max(\text{powerV}, s_\star)\);
14. end while
15. \(\text{return Impermissible};\)
\( \Pi = \langle V, A, s_0, s_\star \rangle \), with \( V = \{5\text{willdie}, 1\text{willdie}, \text{done}\} \), \( A = \{\text{pull}, \text{refrain}\} \), \( s_0 = \{5\text{willdie}, \neg 1\text{willdie}, \neg \text{done}\} \), and \( \text{pull} = \langle \top, \neg 5\text{willdie} \land 1\text{willdie} \land \text{done}\rangle \), \( \text{refrain} = \langle \top, \text{done}\rangle \). The moral valuation function assigns \( u(5\text{willdie}) = -5 \), \( u(\neg 5\text{willdie}) = 5 \), \( u(1\text{willdie}) = -1 \), \( u(\neg 1\text{willdie}) = 1 \), \( u(\text{done}) = u(\neg \text{done}) = u(\text{pull}) = u(\text{refrain}) = 0 \). There are two (shortest) plans that reach the goal: \( \pi = \text{refrain} \) and \( \pi' = \text{pull} \). Plan \( \pi \) is impermissible from the utilitarian point of view because \( \pi' \) leads to a better state. From the deontological point of view, both \( \pi \) and \( \pi' \) are permissible because they do not contain inherently bad actions. The do-no-harm principle renders \( \pi \) permissible (as the only bad effect \( 5\text{willdie} \)) is not produced by \( \pi \); and it renders \( \pi' \) impermissible because \( 1\text{willdie} \) is produced by \( \text{pull} \).

In our example, the conductor is using an AI Planning System (AIPS) used within the framework described in Figure 1 to help them manage the complex rail network. Consider the case where the AIPS is behaving under the do-no-harm principle, so the plan that is presented to the conductor is “refrain from pulling the lever”. The conductor does not understand why the AIPS would allow five people to be killed rather than one. (Step 1) The conductor asks “Why did you refrain from pulling the lever, rather than pulling it?”; and provides the ethical principle for which they want to understand, in this case the do-no-harm principle. (Step 2) The Question Generation module extracts a constraint from the contrast case given by the user’s question and the ethical principle, specifically that the lever is pulled and the do-no-harm principle is satisfied. We then use the same definition of the constraint operator to restrict the model with this constraint to produce the HModel II’. The questions we provide contrastive explanations for, how the constraints are derived, and how they are applied is described in detail by Krarup et al. [9]. These questions are restricted to forms where \( A \) and \( B \) are properties on the actions in the plan. More specifically, questions about inclusion, exclusion, or orderings of actions in the plan, for example, “why was action \( a_1 \) used before action \( a_2 \)”?. (Step 3) The HModel is solved with the planner used within the AIPS to produce an HPlan, which satisfies the constraint derived from the conductor’s question and ethical principle. If the ethical principle cannot be satisfied (the HModel is unsolvable), then there will be no HPlan (from propositions 1 and 2). The ethical principle constraint is removed from the HModel and it is re-solved so that an HPlan can be found and used in the next step. The explanation will then show why the user’s contrast case leads to a plan that is impermissible. (Step 4) The original plan and the HPlan are passed to the Ethical Explanation Generator which returns the necessary reason for the plan being (im-)permissible, this is the moral non-CE. In this case it will return \text{Permissible: } \neg \text{Caused}(5\text{willdie}) \) for the original plan and \text{Impermissible: } \text{Caused}(1\text{willdie}) \) for the HPlan. (Step 5) A Contrastive Explanation is generated which aims to show the difference between these reasons. This is formed from the set difference of the two reasons; and the actions in the plan, the permissibility, principle, and the cause, contrasted to the same information for the HPlan. For example, the CE produced to answer the conductor’s question is “The man could refrain from action. This would be permissible under the do-no-harm principle because this way the death of the five persons is not caused by his action. Alternatively, the man could pull the lever. Doing so is impermissible under the do-no-harm principle because this way the death of the one person is caused by his action.” In this example the HModel II’ is unsolvable because any plan where the lever is pulled is impermissible under the do-no-harm principle. Therefore, as said in Step 4, the permissibility constraint is dropped so that the HModel can be solved to produce an HPlan where the lever is pulled. This HPlan is used in Step 4 to generate the reasons for why it is impermissible, “...because this way the death of the one person is caused by his action”. (Step 6) This whole process can then be iterated, where a user can ask new questions and explore different ethical principles and outcomes until they are satisfied that they either understand the ethics of the plan well, or they have found a plan that meets their moral standards. For example, the conductor still might not be satisfied with the choice to refrain from pulling the lever. The conductor asks the same question as above but instead chooses the utilitarianism principle. Again, the constraint will ensure that the lever is pulled in the plan, but now, that the utilitarianism principle is satisfied. The necessary reasons will again be \text{Permissible: } \neg \text{Caused}(5\text{willdie}) \) for the original plan and \text{Permissible: } \text{GEq}(1\text{willdie} \land \neg 5\text{willdie} \land \text{done}, \neg 1\text{willdie} \land 5\text{willdie} \land \text{done}) \) for the new HPlan. The CE will then be “The man could refrain from action. This would be permissible under the do-no-harm principle because this way the death of the five persons is not caused by his action. Alternatively, the man could pull the lever. Doing so is permissible under the utilitarianism principle because five saved lives is better than one saved life.” The conductor agrees with the outcome of the utilitarianism principle and decides to pull the lever. Explanations produced in this type of iterative framework have been shown to help users understand plans better [9]. It has not been previously studied as to if the CEs and non-CEs discussed in this paper help users to understand ethical principles better, and which are more effective.

V. Evaluation of Generated Explanations

We have conducted a between-subject online-questionnaire study with six conditions, [Deontology, Utilitarianism, DoNoHarm] \times [Contrastive, NonContrastive] to test whether our explanations support humans to build an understanding of the robot’s moral reasoning. Moreover, we were interested in investigating whether the explicit mentioning of the alternative plan
in the verbalization of the CEs has a positive effect. The hypotheses read: (H1) The explanations have an effect on people’s understanding of the robot’s moral reasoning; (H2) CEs are more effective in enabling people to build a mental model of the robot’s moral reasoning as compared to non-CEs.

We recruited N = 300 participants. To ensure quality of data, we have removed all data points where the participants needed less than 3 minutes or more than 30 minutes to complete the questionnaire. The remaining N = 245 data points where used for analysis. The mean age of participants was 27 (SD = 10), 82 female, 159 male, 4 others. All participants got £1.50 for an expected 10 minutes to complete the questionnaire.

First, the classical trolley problem was presented to the participants, both as an image and as text (see Sect. IV-B.2). A moral reasoning robot was then introduced to the participant: “The robot makes moral judgments based on a principle implemented into the robot operating system. People can take recommendations based on these judgments. The robot’s moral judgment reads: The man can pull the lever.” Depending on the conditions, the sentence continued:

- **Deontology, Non-Contrastive**: “Doing so is permissible because pulling a lever is not inherently bad.” **Deontology, Contrastive**: “Doing so is permissible because pulling a lever is not inherently bad. Alternatively, the man can refrain from pulling the lever. This is also permissible, because doing nothing is not inherently bad.” **Utilitarianism, Non-Contrastive**: “Doing so is permissible because five saved lives is better than one saved life.” **Utilitarianism, Contrastive**: “Doing so is permissible because five saved lives is better than one saved life. Alternatively, the man can refrain from pulling the lever. This would be impermissible, because one saved life is worse than five saved lives.” **Do-No-Harm, Non-Contrastive**: “Doing so is impermissible because this way the death of the one person is caused by the man’s action.” **Do-No-Harm, Contrastive**: “Doing so is impermissible because this way the death of the one person is caused by his action. Alternatively, the man could refrain from action. This would be permissible because this way the death of the five persons is not caused by his action.”

The next stage was designed as a prediction task to test the participant’s mental model of the robot’s moral reasoning. The bridge dilemma, a variation of the trolley problem, was presented: “A trolley is hurtling down a track towards five people. A man is standing on a bridge under which it will pass, he can stop it by putting something very heavy in front of it. There is a very fat man next to him – his only way to stop the trolley is to push that man over the bridge and onto the track, killing him to save five. What is the right thing to do?”

The text was accompanied by an image showing the bridge dilemma. Participants were asked for their prediction: “Applying what you already know about the robot’s reasoning from the previous moral judgement of the trolley problem, irrespective of your own moral judgment, rate how likely you consider the robot to come to each of the following moral judgments for this new dilemma.” The available options were: “It is ethical to push the man off the bridge.”, “It is ethical to refrain from pushing the man off the bridge.”, “Both options are ethical.”, and “None of the options are ethical.”. For each option, the participant gave a likeliness rating on a 5-point Likert scale.

A three-way ANOVA was run to examine the influence of the variables Option (= the option for acting), Principle (= the principle the explanation is based on), and Contrast (= contrastive or non-contrastive) on the permissibility ratings. There was a significant interaction between Option and Principle (F(6, 944) = 33.452, p < 0.0001). That is, the ratings that each option receives is significantly depending on the explanation’s content. In further support of hypothesis H1, post-hoc pairwise Bonferroni-adjusted comparisons show significant influences of the various explanations on the permissibility ratings (see Fig. 2):

Participants who had read the deontological explanation rated the option that pushing is permissible lower as compared to participants who had read the utilitarian explanation. The difference is significant for CEs (F(1, 944) = −7.44, p < .0001) and for non-CEs (F(1, 944) = −5.79, p < .0001). The deontological explanation group also rated the option that refraining is permissible higher as compared to the utilitarian explanation group. The difference is significant for CEs (F(1, 944) = 5.62, p < .0001) and for non-CEs (F(1, 944) = 4.34, p < .0001). Moreover, deontological explanations lead to higher ratings for the option that neither pushing nor refraining is permissible as compared to utilitarian explanations. The difference is significant for CEs (F(1, 944) = 3.44, p < .001) but not for non-CEs (F(1, 944) = 1.4, p < .486).

Participants who had read the do-no-harm explanation rated the option that pushing is permissible lower as compared to participants who had read the utilitarian explanation. The difference is significant for CEs (F(1, 944) = −6.74, p < .0001) and for non-CEs (F(1, 944) = −6.43, p < .0001). The do-no-harm explanation group also rated the refraining is permissible option higher than did the utilitarian explanation group. The difference is significant for CEs (F(1, 944) = 6.53, p < .0001) and for non-CEs (F(1, 944) = 3.58,
Moreover, do-no-harm explanations lead to higher ratings of the option that neither pushing nor refraining is permissible higher as compared to utilitarian explanations. The difference is significant for CEs ($F(1,944) = 2.49, p = 0.039$) and for non-CEs ($F(1,944) = 2.52, p = 0.036$).

Regarding hypothesis H2, we observe that the influence of the CE is generally more pronounced than of the non-CE. However, the difference between CE and non-CE does not reach statistical significance. Most notably, we find that the deontological CE group is more convinced that nothing is permissible in the bridge dilemma than the deontological non-CE group ($F(1,944) = 1.82, p = .07, n.s.$), and that the do-no-harm CE group considers refraining more permissible than the do-no-harm non-CE group ($F(1,944) = 1.85, p = .065, n.s.$).

a) Mental Models: The results support hypothesis H1 as they indicate that the explanations enabled people to build a mental model of the robot’s moral reasoning. The results rule out the possibility that the participants relied on their own moral view as the ratings for permissibility of the available options differed significantly. Note that also the sentence “Doing so is (im-)permissible...” cannot explain these differences because in our version of the trolley dilemma, pulling the lever is permissible under the deontological and the utilitarian principle but in the bridge dilemma, deontology forbids and utilitarianism permits pushing the man off the bridge. Participants clearly noticed what the robot considers permissible in each condition. The only explanation for this is that participants took notice of the explanations and transferred their understanding of the robot’s explanation of its moral reasoning in the trolley problem to the bridge dilemma.

A limitation of the conducted study is that this result was obtained by testing a small number of principles applied to one moral dilemma. More work is needed in understanding to what extent the result can be generalized to more complicated moral principles and dilemmas. More complicated moral principles consist of multiple rules (e.g., the Doctrine of Double Effect) or require universalization (e.g., the first formulation of Kant’s Categorical Imperative). To be understandable then, it may be required that explanations also contain an explicit formulation of the moral principle used. More complicated dilemmas would consist of sequences of actions over longer temporal horizons and with more complex causal interdependencies.

b) Non-contrastive vs. contrastive explanations: Regarding the comparison between non-contrastive and contrastive explanations, the results are less supportive of H2. One possible explanation is that humans draw contrastive inferences all the time (cf., [2]). The participants in the non-contrastive conditions may have been able to infer the contrastive part of the explanation themselves. An alternative explanation for the small effect of contrastivity is that the information contained in the non-CE was already sufficient for understanding the ethical principle. That is, the non-CEs deontology group inferred that an action is impermissible if inherently bad, and pushing someone is inherently bad. The non-CEs utilitarian group also had no problem inferring that permissibility is grounded in counting saved lives. Only the non-CEs do-no-harm group was less able to infer from their explanation that refraining from action is permissible. This information is less obvious and only contained in the CE. Generally, more research is needed to understand which inferences people automatically draw from explanations and under which circumstances explicit contrastivity provides additional information.

VI. Conclusions

We have presented a system that enables autonomous systems to explain action plans in terms of ethical principles. Our explanation method compiles a user’s why-question into a new AI planning problem. The technique was integrated into a system that allows humans to actively explore the explanatory space of ethical and unethical alternative plans. Our user study indicates that the explanations generated by the system can improve humans’ understanding of a robot’s ethical principle. As well as further studies, future work will also investigate the compilation of more intricate ethical principles into commonly used planning formalisms.

References

[1] S. Krivic, E. Ugur, and J. Piater, “A robust pushing skill for object delivery between obstacles,” in 2016 IEEE Int.Conf. on Automation Science and Engineering (CASE), 2016.
[2] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Art. Int., vol. 267, pp. 1–38, 2019.
[3] D. Vanderelst and A. Winfield, “An architecture for ethical robots inspired by the simulation theory of cognition,” Cognitive Systems Research, pp. 56–66, 2018.
[4] F. Lindner and M. Bentzen, “The hybrid ethical reasoning agent IMMANUEL,” in Companion of the 2017 ACM/IEEE Int.Conf. on Human-Robot Interaction, 2017, pp. 187–8.
[5] F. Lindner, R. Mattmüller, and B. Nebel, “Evaluation of the moral permissibility of action plans,” Art. Int., vol. 287, 2020.
[6] T. Chakraborti, S. Sreedharam, and S. Kambhampati, “The emerging landscape of explainable automated planning & decision making,” in IJCAI, 2020, pp. 4803–4811.
[7] M. Fox, D. Long, and D. Magazzeni, “Explainable planning,” IJCAI-17 workshop on Explainable AI, vol. abs/1709.10256, 2017.
[8] R. Eifler, M. Cashmore, J. Hoffmann, D. Magazzeni, and M. Steinmetz, “A new approach to plan-space explanation: Analyzing plan-property dependencies in oversubscription planning,” in AAAI, 2020, pp. 9818–9826.
[9] B. Krarup, S. Krivic, D. Magazzeni, D. Long, M. Cashmore, and D. E. Smith, “Contrastive explanations of plans through model restrictions,” JAIR, pp. 533–612, 2021.
[10] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. Wu, H. Lu, and S.-C. Zhu, “A tale of two explanations: Enhancing human trust by explaining robot behavior,” Science Robotics, vol. 4, no. 37, 2019.
[11] J. Broekens, M. Harbers, K. Hindriks, K. van den Bosch, C. Jonker, and J.-J. Meyer, “Do you get it? user-evaluated explainable bdi agents,” in Multiagent System Technologies. Springer Berlin Heidelberg, 2010, pp. 28–39.
[12] F. Lindner and K. Mölling, “Extracting reasons for moral judgments under various ethical principles,” in KI 2019: Advances in Art. Int., 2019, pp. 216–29.