Extracting Reasons for Moral Judgments Under Various Ethical Principles

Felix Lindner and Katrin Möllney

University of Freiburg
Computer Science Department, Foundations of Artificial Intelligence
79110 Freiburg im Breisgau, Germany
{lindner,moellnek}@informatik.uni-freiburg.de

Abstract. We present an approach to the computational extraction of reasons for the sake of explaining moral judgments in the context of an hybrid ethical reasoning agent (HERA). The HERA agent employs logical representations of ethical principles to make judgments about the moral permissibility or impermissbility of actions, and uses the same logical formulae to come up with reasons for these judgments. We motivate the distinction between sufficient reasons, necessary reasons, and necessary parts of sufficient reasons yielding different types of explanations, and we provide algorithms to extract these reasons.

Keywords: Machine Ethics · Explainable AI · Reasons.

1 Introduction

Artificial Intelligence technology is currently making huge impact on society. Many important questions arise on how we want to design technology to the benefit of humans, how we can build systems that are in line with our ethical values, and how we can build systems that we can trust. The approach taken by the machine-ethics community [5, 19, 18] to building systems that align with ethical values is to represent these values formally within these systems and thus enable artificial systems to explicitly take ethical values into account during reasoning and decision making. One such attempt to explicitly formalize ethics is undertaken in the HERA (Hybrid Ethical Reasoning Agents)[1] project [2]. The HERA software library currently provides a suite of philosophically founded and practically usable machine ethics tools for implementation in physical and virtual moral agents such as social robots [1]. Until recently, it was not possible to ask the HERA agent for the reasons why a situation is judged morally permissible or impermissible according to a given ethical principle. It has recently been argued that the capability to explain decisions to humans is an important ingredient for human-robot interaction to ensure trust and transparency [6], and for AI in general [7, 12]. Earlier versions of HERA could not address this requirement in a satisfying way. In this article, we report how, once a moral judgment has

[1] www.hera-project.com

been computed, reasons can be extracted based on the logical representations of ethical principles.

The paper is structured as follows: First, the moral-judgment component of HERA is briefly reviewed. We then propose an explanation component which relies on computing sufficient and necessary reasons that explain a moral judgment. We relate the problem of computing sufficient and necessary reasons to the problem of computing prime implicants and prime implicates of a Boolean formula. We discuss a connection to the INUS condition [9] and problematize cases of overdetermination. We then point out connections to related work in the eXplainable AI community (XAI).

2 Hybrid Ethical Reasoning Agents

2.1 Causal Agency Models

Causal agency models were introduced by Lindner, Bentzen and Nebel [2] as a variant of causal models in the tradition of Pearl and Halpern [3]. Causal agency models are particularly designed to capture ethically relevant aspects of a situation. These include the set of actions available to the agent in that situation, the causal chains of consequences of each action, the intended consequences of each action, as well as a utility function which assigns a numeric value to actions and consequences representing how good or bad that action or consequence is. Definition 1 introduces causal agency models formally.

Definition 1 (Causal Agency Model). *A* causal agency model M *is a tuple* (A, C, F, I, u, W)*, where* A *is the set of* action variables*,* C *is a set of* consequence variables*,* F *is a set of modifiable* Boolean structural equations*,* $I = (I_1, ..., I_n)$ *is a list of sets of intentions (one for each action),* $u : A \cup C \to \mathbb{Z}$ *is a mapping from actions and consequences to their individual* utilities*, and* W *is a set of* Boolean interpretations *of* A*.*

A pair $\langle M, w_\alpha \rangle$ with $w_\alpha \in W$ constitutes the *situation* which results from performing action α according to model M. Intuitively, each Boolean interpretation $w \in W$ of the variables in A corresponds to an *option* available to the agent. The interpretation w_α denotes the interpretation where action α has been chosen, i.e., α has the Boolean value $True$. By assumption, all other actions get the value $False$. Given some w_α, the value of each of the variables in C can be uniquely determined as long as the dependence graph induced by F in situation $\langle M, w_\alpha \rangle$ is recursive (cycle-free), cf., [3]. Subsequently, we will assume that this is always the case.

Symmetric Trolley Problem As a running example throughout this paper we consider a symmetric trolley problem: A trolley has gone out of control and threatens to kill a person (called "person 1"). However, a bystander has the chance to pull a lever and thereby direct the trolley onto the second track. Unfortunately, there is a second person (called "person 2") standing on the second track and who will die in case the lever gets pulled.

The situation is represented as a causal agency model M like this:

$$A = \{a_1, a_2\}$$
$$C = \{d_1, d_2\}$$
$$F = \{f_{d_1} := \neg a_1, f_{d_2} := a_1\}$$
$$I = (I_{a_1} = \{a_1, \neg d_1\}, I_{a_2} = \{a_2\})$$
$$u(a_1) = u(a_2) = 0, u(d_1) = u(d_2) = -1, u(\neg d_1) = u(\neg d_2) = 1$$
$$W = \{\{a_1 \to T, a_2 \to F\}, \{a_1 \to F, a_2 \to T\}\}$$

The action a_1 represents the pulling of the lever, action a_2 is an extra action variable representing refraining from action. This special action will never appear in structural equations, hence, refraining never causes anything. This way, causal agency models can express the distinction between causing and letting happen (cf., [2]). Consequence variables d_1, d_2 represent the deaths of person 1 and person 2, respectively. The structural equation f_{d_1} models that in case of not pulling the lever, person 1 will die. The structural equation f_{d_2} models that the effect of pulling the lever is that person 2 will die. The set I captures that by pulling the lever the agent intends to actually pull the lever (i.e., pulling is a voluntary action) and that the agent intends to rescue person 1's life. In case of refraining, only the refraining itself is supposed to be intended. The death of either of the two persons is considered a bad consequence, their survival is considered good. This is represented by the utility function u.

The symmetric trolley problem so defined does not give rise to the usual tension between utilitarian and non-utilitarian reasoning. However, it still constitutes a case of choosing between causing harm and letting harm happen, and thus different ethical theories will yield different judgments. The overall approach to ethical reasoning presented in [2] can also handle utilitarian reasoning in the classical trolley dilemma (5 persons versus 1 person). For the principles considered in this paper, more persons on the track would not make any difference and would not contribute to the demonstration of the new explainability feature.

For the remainder of the paper, we use $\langle M, w_{a_1} \rangle$ to refer to the situation of the symmetric trolley problem where the agent pulls the lever, and $\langle M, w_{a_2} \rangle$ for the situation where the agent refrains from pulling the lever.

2.2 Causal Agency Logic

A logical language is defined to talk about causal agency models. Particularly, the logic is employed for the specification of moral permissibility according to various ethical principles.

Language The language L of causal agency logic is recursively defined as follows:

- Let $Lit_p = \{p, \neg p \,|\, p \in A \cup C\} \subset L$ be the set of propositional variables denoting actions and consequences and their negations.
- For all actions or consequences $p, q \in Lit_p$ formula $Causes(p, q)$ is in L.

- For all $p \in Lit_p$ formulae $Good(p), Bad(p), Neutral(p) \in L$.
- For all $p \in Lit_p$ formula $I(p) \in L$.
- If $\phi, \psi \in L$, then $\neg\phi, \phi \wedge \psi, \phi \vee \psi, \phi \rightarrow \psi \in L$.

The language is kept simple in various ways: First, it only allows to talk about causation between literals. A more expressive logic would allow to also speak about combinations of actions and consequences being the cause of some consequence. Second, the logic can express that some consequence or action is good, bad, or neutral, but one cannot arbitrarily compare the utility of consequences and actions.

Semantics The semantics of L is defined over situations $\langle M, w_\alpha \rangle$ as follows:

- $\langle M, w_\alpha \rangle \models p$ iff p is an action and $p = \alpha$, or if p is a consequence and the structural equation f_p evaluates to True under $\langle M, w_\alpha \rangle$.
- $\langle M, w_\alpha \rangle \models Causes(p, q)$ iff $\langle M, w_\alpha \rangle \models p \wedge q$ and $\langle M_{\neg p}, w_\alpha \rangle \models \neg q$, where $M_{\neg p}$ is the model where the structural equation of p is substituted by the complement of the truth value that p has in $\langle M, w_\alpha \rangle$. This is in accordance with the but-for definition of causality [3].
- $\langle M, w_\alpha \rangle \models Good(p)$ iff $u(p) > 0$.
- $\langle M, w_\alpha \rangle \models Bad(p)$ iff $0 < u(p)$.
- $\langle M, w_\alpha \rangle \models Neutral(p)$ iff $0 = u(p)$.
- $\langle M, w_\alpha \rangle \models I(p)$ iff $p \in I_\alpha$.
- $\langle M, w_\alpha \rangle \models \neg\phi$ iff $\langle M, w_\alpha \rangle \not\models \phi$.
- $\langle M, w_\alpha \rangle \models \phi \wedge \psi$ iff $\langle M, w_\alpha \rangle \models \phi$ and $\langle M, w_\alpha \rangle \models \psi$.
- $\langle M, w_\alpha \rangle \models \phi \vee \psi$ iff $\langle M, w_\alpha \rangle \models \phi$ or $\langle M, w_\alpha \rangle \models \psi$.
- $\langle M, w_\alpha \rangle \models \phi \rightarrow \psi$ iff $\langle M, w_\alpha \rangle \not\models \phi$ or $\langle M, w_\alpha \rangle \models \psi$.

2.3 Making Moral Judgments

The moral-judgment component of HERA employs model checking: Ethical principles are formulae written in the causal agency logic introduced above. A causal agency model together with an interpretation which sets one action (the performed action) to true is a representation of the situation to be judged from the perspective of a particular ethical principle. The performed action is permissible according to the ethical principle if and only if the situation satisfies the ethical principle.

For brevity we will only introduce the deontological principle and the do-no-harm principle, but many other ethical principles can be formulated (cf., [2]) and handled likewise. The deontological principle is a non-consequentialist ethical principle. Accordingly, all that matters is the intrinsic value of an action rather than the consequences it will bring about.

Definition 2 (Deontological Principle). *Action α in situation $\langle M, w_\alpha \rangle$ is* morally permissible according to the deontological principle *if and only if the action α is morally good or neutral, i.e.,$\langle M, w_\alpha \rangle \models \neg Bad(\alpha)$.*

Thus, in the trolley problem modeled earlier in the text, both pulling the lever and refraining from doing so are permissible from the perspective of the deontological principle, because their intrinsic values are neutral. To verify this, the formulae $\phi_{deon}^{\langle M,w_{a_1}\rangle} = \neg Bad(a_1)$ and $\phi_{deon}^{\langle M,w_{a_2}\rangle} = \neg Bad(a_2)$ have to be checked for truth in the situations $\langle M, w_{a_1}\rangle$ and $\langle M, w_{a_2}\rangle$, respectively.

We note that the resulting judgment is not in line with many textbooks that claim deontology forbids pulling the lever in the trolley problem. This judgment could be reproduced in our model by assigning negative utility to the pull action. Doing so is justified if the modeler advocates the moral view that pulling the lever and causing the death of the one person is actually the same and should not be distinguished from each other. This shows that, generally, moral judgment is a matter of both moral principles and conceptualizations and (mental) models of moral situations.

The do-no-harm principle is a consequentialist principle. Consequentialists do not believe that actions bear intrinsic value which cannot be reduced to the consequences they bring about. The do-no-harm principle renders exactly those actions morally permissible which do not cause harmful consequences. That is, it may be acceptable that harm exists in the situation, however, this harm should not be due to the action performed by the agent. Definition 3 captures the do-no-harm principle formally.

Definition 3 (Do-No-Harm Principle). *An action α in situation $\langle M, w_\alpha\rangle$ is* morally permissible according to the do-no-harm principle *if and only if none of the bad consequences is caused. Formally, $\langle M, w_\alpha\rangle \models \bigwedge_c (Bad(c) \rightarrow \neg Causes(\alpha, c))$.*

Unlike the deontological principle, the do-no-harm principle forbids pulling the lever. This is, because in the situation resulting from pulling the lever person 2 dies, and person 2 would not have died if the lever had not been pulled, i.e., the pulling is a but-for cause for the person's death. Therefore, both $Bad(d_2)$ and $Causes(a_1, d_2)$ hold in situation $\langle M, w_{a_1}\rangle$. However, refraining is permitted by the do-no-harm principle: It is true that person 1 dies when a_2 is set to $True$, but setting a_2 to $False$ does not help person 1. Only in the case of pulling the lever, the harm can be avoided by doing less, and thus the harm counts as caused.

We refer to [2] for a more complete presentation of the HERA approach to ethical reasoning. Next, we make a new contribution by outlining an approach to computing explanations for moral judgments.

3 Generating Sufficient and Necessary Reasons

Given a judgment about the moral (im-)permissibility of some action in a situation, we want to compute reasons that explain why the judgment was made. One naïve way of doing so would consist in just citing the whole formula that was proven to be true in the given situation and which thus is necessary and sufficient for the permissibility judgment, and therefore explains it. For the symmetric trolley dilemma, one could just cite Formula (1) stating the whole necessary and

sufficient condition for the permissibility of refraining in situation $\langle M, w_2 \rangle$.

$$\phi_{\text{DoNoHarm}}^{\langle M, w_{a_2} \rangle} = \tag{1}$$
$$(Bad(d_1) \rightarrow \neg Causes(a_2, d_1)) \wedge (Bad(d_2) \rightarrow \neg Causes(a_2, d_2)) \wedge$$
$$(Bad(\neg d_1) \rightarrow \neg Causes(a_2, \neg d_1)) \wedge (Bad(\neg d_2) \rightarrow \neg Causes(a_2, \neg d_2))$$

Hence, there are eight different literals and their logical connectives to be reported. As the models grow bigger, the formulae representing permissibility also grow in size. There is hope that not the whole of these formulae has to be verbalized to produce a comprehensible explanation. We therefore now turn to the problem of pinpointing subformulae of the ethical principles that were responsible for the moral permissibility judgment in a particular situation.

3.1 Preliminaries

We briefly recall some basic terminology of propositional logic. Every propositional variable and its negation is called a *literal*. An *interpretation* w assigns a Boolean value to every propositional variable of a formula. If the formula is true under w, then w is called a *model*. We also think of a model as the set of literals true under w. A conjunction of literals m (a *monomial*) is called a *prime implicant* of formula ϕ if and only if m entails ϕ, and no proper part of m entails ϕ. A disjunction of literals (a *clause*) c is called a *prime implicate* of ϕ if and only if ϕ entails c, and no proper part of c is already entailed by ϕ.

As an example consider formula $\phi = (x_1 \wedge x_2) \vee (x_1 \wedge x_2 \wedge x_3) \vee (x_1 \wedge x_3)$. The monomial $m_1 = (x_1 \wedge x_2)$ is a prime implicant of ϕ, because the truth of m_1 implies the truth of ϕ and no subformula of m_1 does. The monomial $m_2 = (x_1 \wedge x_2 \wedge x_3)$ is not a prime implicant of ϕ, because removing x_3 results in m_1, which already is a prime implicant. The clause x_1 is a prime implicate of ϕ, because a model that satisfies ϕ will also make x_1 true. The other prime implicate is $x_2 \vee x_3$.

Regarding causal agency logic, not every Boolean model corresponds to a causal agency model. For example, consider $Good(a) \wedge Bad(a)$, which has the propositional model $\{Good(a), Bad(a)\}$ while it is unsatisfiable in causal agency logic. We therefore implemented a theory solver which filters those models that are actual models of causal agency logic respecting the logic's specific constraints: Given a Boolean model w, then if $Good(x) \in w$, then $Neutral(x) \notin w$ and $Bad(x) \notin w$ (and analog constraints for $Neutral$ and Bad); if $Causes(x, y) \in w$, then $\neg x, \neg y \notin w$, $Causes(x, \neg y) \notin w$, $Causes(\neg x, y) \notin w$, and $Causes(y, x) \notin w$ (if $x \neq y$); and $\neg Causes(x, x) \notin w$; if $I(x) \in w$, then $I(\neg x) \notin w$.

3.2 How Principles Give Rise to Reasons

As a starting point of our analysis we take a deeper look at the properties of the formulae which represent the ethical principles. Each such formula is grounded in a particular situation. This means they are built from the action and consequence

variables specified in the given situation via the causal agency model. Thus, the domain of quantification is fixed. Formula (1) is such a formula grounded in situation obtained from pulling the lever in the causal agency model defined in Subsect. 2.1. Usually, the permissibility judgment depends on further properties of actions and consequences—such as being bad or being caused. Some combinations of such properties already entail the permissibility judgment. For instance, in the trolley dilemma from the introduction, nothing being caused by refraining entails the formula that represents the do-no-harm principle: Because nothing is caused, all other properties (being good or bad) have no impact on the judgment. Hence, nothing being caused is a *sufficient reason* for the permissibility of refraining. Counterfactually, had refraining caused the death of person 1, then refraining would have been judged impermissible. Hence, the fact that person 1's death is not caused is a *necessary reason* for the permissibility of refraining.

We anticipate that sufficient reasons give a good idea about the regularities that underlie a judgment, while necessary reasons give an idea about what should have been different to prevent that judgment. Hence, an agent can learn from necessary reasons for future actions. As will become apparent by the end of this section, there are also reasons that are both sufficient and necessary. These types of reasons are often very concise and straight to the point.

Sufficient Reasons We take a *sufficient reason* to be a minimal conjunctive term which entails the permissibility judgment. More formally, a conjunctive term ψ is a sufficient reason for the permissibility of α in model M according to principle P iff $\langle M, w_\alpha\rangle \models \phi_P^{\langle M,w_\alpha\rangle} \wedge \psi$ (actuality), $\psi \models \phi_P^{\langle M,w_\alpha\rangle}$ (sufficiency), and no sub-term of ψ is already sufficient. Hence, asking for a sufficient reason is the same as asking for a prime implicant of the (grounded) ethical principle formula. To compute all prime implicants of $\phi_P^{\langle M,w_\alpha\rangle}$, HERA employs a SAT solver [15] to compute all models of $\phi_P^{\langle M,w_\alpha\rangle}$, and a theory solver to pick those models which are also models of causal agency logic (see note in Subsect. 3.1). Each of the models so found is an implicant. To obtain prime implicants we search for inclusion-minimal parts of the implicants, *sub*, already sufficient for the truth of $\phi_P^{\langle M,w_\alpha\rangle}$, i.e., for which $sub \rightarrow \phi_P^{\langle M,w_\alpha\rangle}$ is a tautology. All inclusion-minimal parts sufficient for the truth of $\phi_P^{\langle M,w_\alpha\rangle}$ are kept.

The set of prime implicants of Formula (1), $\phi_{\text{DoNoHarm}}^{\langle M,w_{a_2}\rangle}$, is listed as Formulae (2–5) below.

$$\neg Causes(a_2, d_1) \wedge \neg Causes(a_2, \neg d_2) \tag{2}$$

$$\neg Causes(a_2, d_1) \wedge \neg Bad(\neg d_2) \tag{3}$$

$$\neg Causes(a_2, \neg d_2) \wedge \neg Bad(\neg d_1) \tag{4}$$

$$\neg Bad(d_1) \wedge \neg Bad(\neg d_1) \tag{5}$$

Thus, if we only knew the HERA agent used formula $\phi_{\text{DoNoHarm}}^{\langle M,w_{a_2}\rangle}$ to evaluate the situation and that the agent came to the judgment that the situation was permissible, then we can conclude that the HERA agent believes in at least

one of the four formulae (2) to (5). In the depicted case, we learn that to be permissible, the situation has to be such that either the action does not cause any consequences (2), or no consequence is bad (5), or one consequence is not caused (so it does not matter if it is bad) and the other is not bad (so it does not matter if it is caused).

Next, the HERA agent can state its beliefs about the causal relationships in the situation and its beliefs about moral badness or goodness by citing those prime implicants that are consistent with these beliefs as an explanation for its judgment. For the case of refraining from pulling the lever in the trolley problem (Subsect. 2.1), the agent can thus cite formulae (2) and (3) as sufficient reasons:"Refraining is permissible, because the death of person 1 is not caused nor is the survival of person 2." and "Refraining is permissible, because the death of person 1 is not caused, and the survival of person 2 is not bad."

Necessary Reasons We take a *necessary reason* for a permissibility judgment to be a minimal property whose negation would result in an impermissibility judgment, thus, literally, the truth of this property is necessary for the permissibility. Thus, a necessary reason for the truth of $\phi_p^{\langle M,w_\alpha\rangle}$ is a minimal formula ψ such that the falsehood of ψ implies the falsehood of $\phi_p^{\langle M,w_\alpha\rangle}$. That is, ψ is a necessary reason for the truth of $\phi_p^{\langle M,w_\alpha\rangle}$ (and therefore for the permissibility of α according to principle p), iff $\models \neg\psi \rightarrow \neg\phi_p^{\langle M,w_\alpha\rangle}$ holds. This is equal to requiring that $\models \phi_p^{\langle M,w_\alpha\rangle} \rightarrow \psi$ holds. Hence, ψ is a necessary reason iff ψ is a prime implicate of $\phi_p^{\langle M,w_\alpha\rangle}$. For the computation of prime implicates, we make use of the relationship between prime implicates and prime implicants [16]: The prime implicates of a formula ϕ are just the negations of the prime implicants of $\neg\phi$; and we have already seen above how prime implicants can be computed. The prime implicates of $\neg\phi_{\text{DoNoHarm}}^{\langle M,w_{a_2}\rangle}$ are given in Equations (6) and (7).

$$\neg Causes(a_2, d_1) \vee \neg Bad(d_1) \tag{6}$$

$$\neg Causes(a_2, \neg d_2) \vee \neg Bad(\neg d_2) \tag{7}$$

Consequently, the permissibility of refraining implies that both (6) and (7) are true: Either the death of person 1 is not true or it is not bad, and either the death of survival is not caused or it is not bad. Hence, if the negation of any of the prime implicates (6) or (7), viz., (8) or (9), were satisified in the situation, refraining would be impermissible according to the do-no-harm principle.

$$Causes(a_2, d_1) \wedge Bad(d_1) \tag{8}$$

$$Causes(a_2, \neg d_2) \wedge Bad(\neg d_2) \tag{9}$$

Some of the conjuncts of the negated prime implicates may already be satisfied in the given situation and therefore be no convincing reasons when talking about the concrete situation. For example, $Bad(d_1)$ is true in the causal agency model representing the symmetric trolley problem, and it may sound strange to state that had the death of person 1 been bad and caused, then refraining would

have been impermissible (although this is, of course, correct). Therefore, we decide that $Bad(d_1)$ can be removed leaving $Causes(a_2, d_1)$ as the interesting part of this implicate. Indeed, if it were (additionally) the case that $Causes(a_2, d_1)$ were true in the situation, then the action would be impermissible. A second way of altering the judgment would require two changes to the situation: the survival of person 2 must be caused and its survival must be morally bad. We hence end up with Formulae (10) and (11).

$$Causes(a_2, d_1) \tag{10}$$

$$Causes(a_2, \neg d_2) \wedge Bad(\neg d_2) \tag{11}$$

In accordance to necessity, we take the perspective that conditions (10) and (11) to *not* be true was necessary for the permissibility judgment, i.e., $\neg Causes(a_2, d_1)$ and $\neg(Causes(a_2, \neg d_2) \wedge Bad(\neg d_2))$ are necessary reasons for the permissibility of action a_2. This leads to (12) and (13).

$$\neg Causes(a_2, d_1) \tag{12}$$

$$\neg Causes(a_2, \neg d_2) \vee \neg Bad(\neg d_2) \tag{13}$$

Formulae (12) and (13) correspond to the actual output of the HERA agent: For refraining to be permissible, it was necessary that the death of person 1 was not caused, and it was necessary that it was not the case that the survival of person 2 was bad and caused.

Necessary Parts of Sufficient Reasons Mackie [9] has proposed the INUS condition, according to which causal explanations are *Insufficient but Necessary parts of a condition which is itself Unnecessary but Sufficient.*

Indeed, for the symmetric trolley problem, we find exactly one such INUS reason, viz., $\neg Causes(a_2, d_1)$. However, for the deontology principle, there is no INUS reason, because the only fact that explains the permissibility of the action is its not being bad, and this reason is both sufficient and necessary. We could either decide that no INUS reasons exists in this case, or we can decide to weaken the INUS condition a bit and identify those reasons which are necessary parts of sufficient reasons. Under this condition, we still get only $\neg Causes(a_2, d_1)$ as the reason for the permissibility of refraining in the symmetric trolley problem under the do-no-harm principle, and we also get $\neg Bad(a_2)$ as the reason under the deontology principle. In the actual HERA implementation, we have decided to go for the weakened version of the INUS condition.

The computation of INUS reasons is straightforward: For each necessary reason c (a clause), we check if there is a sufficient reason m (a monomial), such that every literal in c is also a literal in m.

Using the concept of a necessary part of a sufficient reason, the HERA agent is able to say:"Refraining is permissible (according to the do-no-harm principle), because refraining does not cause the death of person 1." and "Refraining is permissible (according to deontology), because refraining is not morally bad."

Impermissibility We finally turn to explaining impermissibility judgments. In case of impermissibility, the formula $\phi_p^{\langle M,w_\alpha\rangle}$ is false in the given situation. For instance, pulling the lever in the symmetric trolley problem is impermissible according to the do-no-harm principle, because $\phi_{\text{DoNoHarm}}^{\langle M,w_{a_1}\rangle}$ is false in $\langle M, w_{a_1}\rangle$. To find out the reasons for the formula not to be satisfied, we reduce the computation of sufficient and necessary reasons for impermissibility judgments to reason computation for permissibility judgments. First, $\phi_p^{\langle M,w_\alpha\rangle}$ is negated, and then the necessary and sufficient reasons for the truth of $\neg\phi_p^{\langle M,w_\alpha\rangle}$ are computed just in the same way as outlined above. For the trolley problem example, we thus start with the formula $\neg\phi_{\text{DoNoHarm}}^{\langle M,w_{a_1}\rangle}$:

$$\begin{aligned}\neg\phi_{\text{DoNoHarm}}^{\langle M,w_{a_1}\rangle} = \\ (Bad(d_1) \wedge Causes(a_1, d_1)) \vee (Bad(d_2) \wedge Causes(a_1, d_2)) \vee \\ (Bad(\neg d_1) \wedge Causes(a_1, \neg d_1)) \vee (Bad(\neg d_2) \wedge Causes(a_1, \neg d_2))\end{aligned} \tag{14}$$

In this case, each conjunct is a prime implicant. One of them is true in $\langle M, w_{a_1}\rangle$:

$$Bad(d_2) \wedge Causes(a_1, d_2) \tag{15}$$

Hence, there is one sufficient reasons for the impermissibility of pulling the lever:"Pulling the lever is impermissible, because the death of person 2 is bad and pulling causes the death of person 2."

Each of the conjuncts of Formula 15 is a necessary reason. In fact, both these reasons are also reasons according to the INUS condition (and its weakened version). As they refer to the salient feature of the situation (the death of person 2), the formulations "Pulling the lever is impermissible, because the death of person 2 is bad" and "Pulling the lever is impermissible, because pulling the lever causes the death of person 2" sound reasonable.

4 Discussion

We have proposed three types of explanations: those based on sufficient reasons, those based on necessary reasons, and those based on necessary reasons that are part of a sufficient reason. One problem that all these reasons may suffer from is that they do not explicitly take the knowledge status of the hearer into account, a factor which is known to be essential for explanations to be comprehensible [8]. Consider, for example, the reason $Causes(a_1, d_2)$, which is a necessary part of a sufficient reason for the impermissibility of pulling the lever (a_1). A hearer of this explanation who is not aware of the do-no-harm principle or who does not know that d_2 is morally bad, may ask "Why is causing the death of person 2 (d_2) a reason for impermissibility?" We leave it as an open question, how such questions could be addressed by including more information or by a dialogue with the HERA agent.

Moreover, it is questionable if principle-based reasons are appropriate moral reasons in all possible application domains. Stocker [24] gives the example of

visiting a friend in a hospital. Neither "I visit you, because visiting is not bad" nor "I visit you, because doing so does not cause harm" seem appropriate—instead the explanation should cite care for that person as reason.

A more technical problem explanations have to deal with is the problem caused by overdetermination. Overdetermination occurs in theories of causation whenever two or more conditions are sufficient for one effect to occur, cf., [10]. Under such circumstances, it is not possible to point out single causes. In our case, overdetermination comes into play when more than one condition is sufficient for the (im-)permissibility judgment. Consider the situation when pulling the lever causes the death of two persons: person 1 and person 2. The causing of one of the two deaths is already sufficient for the impermissibility judgment. The two sufficient reasons are:

$$Bad(d_1) \wedge Causes(a_1, d_1) \tag{16}$$

$$Bad(d_2) \wedge Causes(a_1, d_2) \tag{17}$$

The necessary reasons are:

$$Bad(d_1) \vee Causes(a_1, d_2) \tag{18}$$

$$Bad(d_2) \vee Causes(a_1, d_1) \tag{19}$$

$$Bad(d_1) \vee Bad(d_2) \tag{20}$$

$$Causes(a_1, d_1) \vee Causes(a_1, d_2) \tag{21}$$

In this case, we cannot find any necessary reason which is part of a sufficient reason. This is because it is necessary for the action to be permissible to change conditions with respect to both caused deaths, viz., either make them morally acceptable or avoid causing them. However, the sufficient reasons only talk about individual deaths, because each of these deaths is sufficient on its own. For now, we just take this as a proof that INUS reasons do not always exist, even under the weakened definition. Thus, subsequent procedures, like natural-language generation, should be prepared to make use of the other two types of reasons in case the set of INUS reasons is empty.

The runtime performance of the current implementation is not suited for real-time use. As dilemmas become more complex or more complex principles (such as the Pareto principle [4] or the principle of double effect [2]) are used, reason generation becomes time consuming. For instance, while explaining permissibility of refraining in the symmetric trolley problem is very fast, explaining under the double effect principle already takes several minutes on a usual desktop machine. This is due to the fact that finding an prime implicant is a NP-hard problem, and our procedure enumerates all (potentially exponentially many) of them. In future, we plan to employ more sophisticated approaches to prime implicant enumeration, e.g., those described in [16, 23]. Moreover, some of our principles are actually Horn formulas or even representable as 2-CNF formulas (e.g., do-no-harm principle). Thus, in future we will exploit the complexity class of the satisfiability problem of the logical fragment actually needed for the description of the ethical principle at hand, and use more specialized algorithms for prime implicant and prime implicate generation.

5 Related Work

Explainable AI has recently gained new interest due to the broad success of AI. This section only very briefly summarizes some recent developments that cut cross statistical and logics-based approaches to AI.

Dannenhauer and colleagues [11] propose an architecture for enabling agents to explain why they chose not to adopt a goal. In their approach, when the agent rejects a goal, the agent proves that it could not find a plan that achieves that goal without violating a hard constraint. In contrast, HERA agents evaluate actions not goals. Another difference is that HERA agents employ ethical principles which do not necessarily reason about alternatives. Explanations that involve contrastive arguments referring to alternatives (e.g., the action is permissible, because the other ones are even worse) are not always what we are after. Russell [13] proposes a method for generating counterfactual explanations of outputs of arbitrary classifiers. The method solves this problem as an integer program which finds a data-point which is maximally similar to the original input but results in another classification. The difference between the original datapoint and the chosen one can be cited as a counterfactual reason. This is closely related to our necessary reasons, whose negations also denote minimal conditions under which the judgment would have been different. Shih and colleagues [14] take a similar approach in the context of Bayesian classifiers. Apart from data points which lead to changes in the classifier's output (necessary reasons), the authors also consider parts of input that always lead to the classifier's output no matter how the other parts of the input look like (sufficient reasons).

The aforementioned approaches are located outside the domain of logic-based AI. However, they are less far away than one might expect: Our setting can also be conceptualized as a classifier (viz., the formula representing the conditions for moral permissibility of the action in that situation) classifying input (viz., the situation as given as a causal agency model together with an option) as either permissible or not. By finding out which parts of the input are sufficient or necessary for the classification, we compute sufficient and necessary reasons.

Another closely related approach is presented by Baum and colleagues [18]. Here, a robot's moral decision making is modeled as a utilitarian decision function, which judges an action possibility A as more, less, or equally morally permissible than another action possibility B. The authors propose a means to algorithmically derives arguments for the robot's judgments which can be presented to humans as rationalizations of the robot's actions. Borgo and colleagues [22] propose a system for explaining action plans to users. The user of the system can propose an alternative plan to the one generated by an AI planner. The explanation module then generates an explanation by comparing features of the user's and the AI planner's suggested plans. The type of explanation generated then can be categorized as consisting of necessary reasons, i.e., "the AI planner has generated that plan, because the alternative plan is more costly." Finally, logic programming has been employed for machine ethics in [20, 19]. The authors also capture notions of causality for moral reasoning using logics. Future work

should explore how approaches to explanation generation for logic programs [21] could be applied to machine ethics.

Interestingly, we are the only ones to explicitly point out INUS reasons and the problem of overdetermination in explanation. We expect both concepts to be important aspects when computing comprehensible explanations. None of the mentioned approaches really investigates the production of linguistic explanation. In the context of a navigating robot, Rosenthal an colleagues [17] propose different types of explanations along the dimensions of abstractness and granularity. However, they do not address the distinction between sufficiency and necessity. We see it as a limitation of the current state of the art that the algorithmic problem of reason computation and linguistic aspects of explanation generation have not been considered together.

6 Conclusions

The HERA architecture got extended with a module for the extraction of reasons for the sake of generating explanations for judgments about the moral permissibility or impermissibility of actions. The approach operates on the formulae that represent the ethical principle used for the judgment—no additional knowledge engineering is necessary. The procedure is based on computing prime implicants and prime implicates of the ethical-principle formulae. We take prime implicants and prime implicates to correspond to sufficient reasons and necessary reasons for the (im-)permissibility judgments. Further, a (weakened) version of the INUS condition was proposed to serve as a definition of a third type of reasons. It identifies necessary reasons that are part of sufficient reasons as constituents of explanations. However, we have seen that such reasons do not always exist like in the case of overdetermination. We are currently working towards linguistically framing explanations. This work will investigate which type of reasons are best suited for communicating different aspects of the situation. INUS reasons seem to be quite concise and straight to the point. Sufficient reasons seem to give a good idea about the regularities that underlie the judgment. And necessary reasons give an idea about what should have been different in order to enforce a different judgment, thus, they provide the basis for contrastive explanations. A user study to investigate these questions is underway.

Acknowledgments

We would like to thank the three anonymous reviewers for their constructive comments.

References

1. Lindner, F., Bentzen, M. M.: The hybrid ethical reasoning agent IMMANUEL. In HRI 2017, pp. 187–188 (2017)

2. Lindner, F., Bentzen, M. M., Nebel, B.: The HERA approach to morally competent robots. In IROS 2017, pp. 6991–6997 (2017)
3. Halpern, Y.: Causality. MIT Press (2016)
4. Kuhnert, B., Lindner, F., Bentzen, M. M., Ragni, M.: Perceived difficulty of moral dilemmas depends on their causal structure: A formal model and preliminary results. In CogSci 2017, pp. 2494–2499 (2017)
5. Anderson, M., Anderson, S. L.: Machine Ethics. Cambridge University Press (2011)
6. Wachter, S., Mittelstadt, B., Floridi, L.: Transparent, explainable, and accountable AI for robotics. Science of Robotics 2(6) (2017)
7. Mittelstadt, B., Russel, C., Wachter, S.: Explaining explanations in AI. In: FAT*'19, pp. 279–288 (2019)
8. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38 (2019)
9. Mackie, J. L.: Causes and Conditions. American Philosophical Quarterly, vol. 12, pp. 245–65 (1965)
10. Lewis, D.: Causation. Journal of Philosophy 70, 556–567 (1973)
11. Dannenhauer, D., Floyd, M. W., Magazzeni, D., Aha, D. W.: Explaining rebel behavior in goal reasoning agents. In ICAPS-18 Workshop on Explainable Planning, pp. 12–18 (2018)
12. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In Twenty-Ninth Annual Conference on Innovative Applications of Artificial Intelligence, pp. 4762–4763 (2017)
13. Russell, C.: Efficient search for diverse coherent explanations. In FAT* 19, pp. 20–28 (2019)
14. Shih, A.,Choi, A., Darwiche, A.: A symbolic approach to explaining Bayesian network classifiers. In IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI), pp. 144–150 (2018)
15. Ignatiev, A., Margado, A., Marques-Silva, J.: PySAT: A Python toolkit for prototyping with SAT oracles. In: SAT, pp. 428–437 (2018)
16. Said, J., Marques-Silva, J., Sais, L., Salhi, Y.: Enumerating prime implicants of propositional formulae in conjunctive normal form. In: JELIA 2014, pp. 152–165 (2014)
17. Rosenthal, S., Selvaraj, S. P., Veloso, M.: Verbalization: Narration of autonomous robot experience. In: IJCAI-16, pp. 862–868 (2016)
18. Baum, K., Hermanns, H., Speith, T.: From machine ethics to explainability and back. In: International Symposium on Artificial Intelligence and Mathematics (ISAIM 2018) (2018).
19. Hölldobler, S.: Ethical decision making under the weak completion semantics. In: Proceedings of the Workshop on Bridging the Gap between Human and Automated Reasoning, pp. 1–5 (2018)
20. Pereira, L. M., Saptawijaya, A.: Programming Machine Ethics. Springer (2016)
21. Shanahan, M.: Prediction is deduction but explanation is abduction. In IJCAI'89, pp. 1055–1060 (1989)
22. Borgo, R., Cashmore, M., Magazzeni, D.: Towards providing justifications for planner decisions. In: Proceedings of IJCAI-18 Workshop on Explainable AI (2018)
23. Previti, A., Ignatiev, A., Morgado, A., Marques-Silva, J.: Prime compilation of non-clausal formulae. In: IJCAI 2015, pp. 1980–1987 (2015)
24. Stocker, M.: The schizophrenia of modern ethical theories. The Journal of Philosophy 73(14), 453–466 (1976)