The HERA Approach To Morally Competent Robots

Felix Lindner[1], Martin Mose Bentzen[2], and Bernhard Nebel[1]

***Abstract*— To address the requirement for autonomous moral decision making, we introduce a software library for modeling hybrid ethical reasoning agents (short: HERA). The goal of the HERA project is to provide theoretically well-founded and practically usable logic-based machine ethics tools for implementation in robots. The novelty is that HERA implements multiple ethical principles like utilitarianism, the principle of double effect, and a Pareto-inspired principle. These principles can be used to automatically assess moral situations represented in a format we call causal agency models. We discuss how to model moral situations using our approach, and how it can cope with uncertainty about moral values. Finally, we briefly outline the architecture of our robot IMMANUEL, which implements HERA and is able to explain ethical decisions to humans.**

I. INTRODUCTION

Currently, we experience a hot debate on moral reasoning and artificial intelligence (AI). In one respect, the discussion is about how to apply AI technology morally. In another respect, there is a requirement to enable AI technology itself to make moral decisions. Fields of application are self-driving cars [1], robots navigating in social environments [2], and robots that give moral advice [3]. One concern is that robots which base their decision making merely on optimization might prefer actions that cause considerable harm just because the final goal is optimal. As a consequence of such considerations, new research areas such as machine ethics [4] and moral human-robot interaction [5] arise.

To address the requirement for autonomous moral decision making, we introduce a software library for modeling hybrid ethical reasoning agents (short: HERA, software available on `http://www.hera-project.com`) [3]. The goal of the HERA project is to provide novel, theoretically well-founded and practically usable logic-based machine ethics tools for implementation in robots. The HERA approach differs from many other approaches that aim to equip robots with moral competence [6], [7], [8], [9], [10]. Whereas most current approaches are committed to label specific actions as morally impermissible or are limited to utilitarian optimization, we think that robot should know multiple philosophical views according to which actions can be morally assessed in various ways. There are at least three reasons to adopt this approach: First, if an action plan is acceptable according to several principles it might be more likely to be acceptable to more people. Second, if different principles say different things, robots can signal uncertainty, hesitate, and ask humans what they think is right. Third, different robot personalities can be modeled by different moral views implemented in robots. Currently, HERA implements utilitarianism, the principle of double effect, a Do-No-Harm principle, and a Pareto principle.

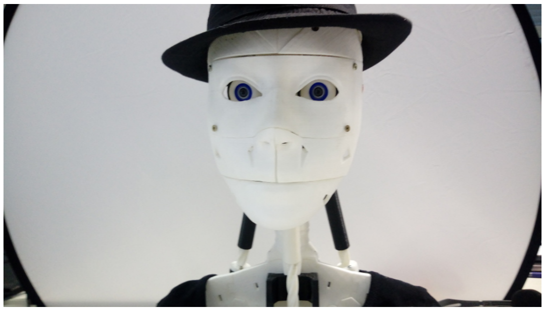

Fig. 1: The HERA prototype IMMANUEL.

For implementation and evaluation of the HERA approach we use the robot IMMANUEL[1] (Interactive Moral Machine bAsed oN mUltiple Ethical principLes, see Fig. 1). A key motivation for building IMMANUEL is to understand the diversity of human ethical reasoning in human-robot interaction contexts [11]. Particularly, we aim to investigate the validity of our claim that the development of ethical reasoning machines needs to integrate multiple moral theories rather than trying to pick “the right one” to be implemented.

The paper is structured as follows: In Sect. II, causal agency models being the technical foundations of our approach are introduced. Ethical principles that assess the moral permissibility of actions are described in Sect. III. We evaluate our approach by demonstrating by example how it behaves in case of moral dilemmas (Sect. IV) and in case of moral uncertainty (Sect. V). Finally, we sketch in Sect. VI how the ethical reasoning component is integrated into the architecture of our robot IMMANUEL.

II. CAUSAL AGENCY MODELS

Using HERA, ethical principles are modeled as logical formulae whose truth determines which actions are permissible and which are not [12]. Informally, actions and their consequences are modeled as directed acyclic graphs, showing causal influence. This format is a variant of Pearl-Halpern-style causal networks [13] restricted to boolean variables.

Definition 1 (Causal Agency Model)
A boolean causal agency model *M is a tuple (A, C, F, I, u, W), where A is the set of* action variables*, C*

[1] Foundations of Artificial Intelligence Lab, Computer Science Department, University of Freiburg, 79110 Freiburg im Breisgau, Germany `{lindner,nebel}@informatik.uni-freiburg.de`

[2] Management Engineering, Danish Technical University, Lyngby, Denmark `mmbe@dtu.dk`

[1] The robot is based on the InMoov Platform, `http://inmoov.fr`

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.*

Intuitively, the elements of W correspond to actions available to the agent. Hence, each $w \in W$ is also called an *option*. We assume that each $w \in W$ assigns 1 (*true*) to exactly one element of A, i.e., each option involves exactly one action to be performed. Regarding $I = (I_1, ..., I_n)$, it is required that: 1) I_i is a set of literals, not containing a variable and its negation (consistency), 2) $a_i \in I_i$ (the performance of the action is intended), 3) a_i causally influences any intended consequence as decribed next.

Causal influence is determined by the set $F = \{f_1, \ldots, f_m\}$ of boolean structural equations. Each variable $c_i \in C$ is associated with the function $f_i \in F$. This function will give c_i its value under an interpretation $w \in W$. An interpretation w is extended to the consequence variables as follows: For a variable $c_i \in C$, let $\{c_{i1}, \ldots, c_{im-1}\}$ be the variables of $C \setminus \{c_i\}$, and $A = \{a_1, \ldots, a_n\}$ the action variables. The assignment of truth values to consequences is determined by $w(c_i) = f_i(w(a_1), \ldots, w(a_n), w(c_{i1}), \ldots, w(c_{im-1}))$. In the general setting, it may be unfeasible to extend an interpretation from the action variables to the rest of the variables, because it is possible that the value of some variable depends on the value of another variable, and the value of the latter variable depends on the value of the former.

Definition 2 (Dependence)
Let $v_i, v_j \in A \cup C$ be distinct variables. The variable v_i depends on *variable v_j, if, for some vector of boolean values, $f_i(\ldots, v_j = 0, \ldots) \neq f_i(\ldots, v_j = 1, \ldots)$.*

Following Halpern [13], we restrict causal agency models to acyclic models, i.e., models in which no two variables are mutually dependent on each other. First, note that the values of action variables in set A are determined externally by the interpretations in W. Thus, the truth values of action variables do not depend on any other variables. Additionally, we require that the transitive closure, $\prec$, of the dependence relation is a partial order on the set of variables: $v_1 \prec v_2$ reads "v_1 is affected by v_2". Anti-symmetry and transitivity of the partial order enforces absence of cycles: Anti-symmetry ensures that if variable v_1 is affected by some different variable v_2 (viz., $v_1 \neq v_2$), then v_2 is not affected by v_1. Transitivity means that if $v_1 \prec v_2$, and $v_2 \prec v_3$, then $v_1 \prec v_3$. Thus, if there were a cycle $v_1 \prec v_2, v_2 \prec v_3, \ldots, v_{n-1} \prec v_n, v_n \prec v_1$, then, by transitivity one would get $v_2 \prec v_1$ violating anti-symmetry. In case of acyclic models, the values of all consequence variables can be determined unambiguously: First, there will be consequence variables only affected by action variables, and whose truth value can thus be determined by the values set by the interpretation. Call these consequence variables *level one*. On *level two*, there will be consequence variables affected by actions and level-one consequence variables, and so on (cf., [12], [13]).

To improve readability, we specify the causal mechanisms using boolean connectives. For instance, $c_3 := a_0 \land \neg c_1$ means that c_3 is true in the model if a_0 is true and c_1 is false. We assume some familiarity on the part of the reader with classical propositional logic. A formula such as $(c_1 \land a_1)$ is intended to mean that consequence c_1 and action a_1 both hold. We write $M, w \models (c_1 \land a_1)$ for "$(c_1 \land a_1)$ holds when the agent choses option w in the model M", and we write $M, w \models Ic$ for "consequence c is intended by the agent when it choses option w in model M". Apart from that, we need simple arithmetic formulae expressing the utility of literals. We write $u(v_i) = z$, for an integer z, with the intended meaning that the utility of v_i is z, similarly we write $u(v_i) \geq u(v_j)$ for the utility of v_i being equal to or greater than the utility of v_j, and so on. We extend the utility function to conjunctions of literals by addition of the utilities of the conjuncts. The utility of other formulae (e.g., disjunctions) is undefined.

Ethical principles may take causation into account. Intuitively, an agent is responsible for the occurance of some consequence, if the consequence would not have occured in case the agent had not performed the action he did perform. To reason about causation in this counterfactual manner, we define the relation of Y being a but-for cause of ϕ inspired by Halpern-Pearl definition of actual cause [13]. This definition of causality makes use of *external interventions* on models. An external interventions X consists of a set of literals (viz., action variables, consequence variables, and negations thereof). Applying an external intervention to a causal agency model results in a new causal agency model M_X. The truth of a variable $v \in A \cup C$ in M_X is determined in the following way: If $v \in X$, then v is true in M_X, if $\neg v \in X$, then v is false in M_X, and if neither $v \in X$ nor $\neg v \in X$, then v is true in M_X if and only if v is true in M. Thus, external interventions override structural equations of those variables occuring in X.

Definition 3 (Actual But-For Cause)
Let y be a literal and ϕ a formula. We say that y is an actual but-for cause *of ϕ (notation: $y \rightsquigarrow \phi$) in the situation the agent choses option w in model M, if and only if $M, w \models y \land \phi$ and $M_{\{\neg y\}}, w \models \neg \phi$.*

The first condition says that both the cause and the effect must be actual. The second condition says that if y had not held, then ϕ would have not occurred. Thus, in the chosen situation, y was necessary to bring about ϕ.[2] The definition of but-for cause is used to distinguish direct consequences from indirect consequences.

Definition 4 (Direct Consequence)
A variable $v_i \in C$ is a direct consequence *of $v_j \in A \cup C$ in the situation w in model M iff $M, w \models v_j \rightsquigarrow v_i$.*

[2] To deal with peculiarities of causality, Halpern [13] introduces definitions of causality which go beyond but-for causality. Also, these more elaborate definitions allow for conjuncts of literals to be causes. This is not relevant for our purposes in this paper, though.

III. Ethical Principles

In moral philosophy, various so-called ethical principles are described. Ethical principles are descriptions of abstract rules that can be used to determine the moral permissibility of concrete courses of actions. Causal agency models play the role of representations of situations involving moral decisions. In this section, we define three ethical principles which embrace different views on how to assess moral permissibility of actions based on the actions' consequences: Utilitarianism, Pareto Principle, and Principle of Double Effect. These three principles are chosen, because they lead to different evaluations of the three moral dilemmas introduced later in this paper. A recent psychological study [14] has shown that the utilitarian principle and the Pareto principle disagree in their evaluation in exactly those cases that are rated morally difficult by human subjects. This is a reason to consider both of them. The principle of Double Effect is known to explain why humans may accept harm as a side effect of some good action but not if it is a means to bring about some goal. In some cases, the Principle of Double Effect forbids actions that both Utilitarianism and the Pareto principle permit.

The *utilitarian principle* presupposes some theory of what is good, i.e., a theory that assigns utilities to consequences: an agent is permitted to perform an action if and only if the action is amongst the available alternative actions with the overall maximal utility. Utilitarian evaluation does not regard the agent's intentions and means. Consequently, utilitarianism allows agents to cause considerably harmful consequences, and to adopt immoral means to a goal.

Definition 5 (Utilitarian Principle)
Let $w_0, ..., w_n$ be the available options, and $cons_{w_i} = \{c \,|\, M, w_i \models c\}$ be the set of consequences and their negations that hold in these options. An option w_p is permissible according to the utilitarian principle if and only if none of its alternatives yield more overall utility, i.e., $M \models \bigwedge_i u(\bigwedge cons_{w_p}) \geq u(\bigwedge cons_{w_i})$.

To define the Pareto principle, the notion of Pareto dominance is defined first: An option w_a *dominates* another w_b if w_a improves aspects of w_b either by making that more good consequences hold or less bad consequences hold. Thus the agent does not change the world for the worse in any aspect and may change it for the better by choosing the dominant action instead of the dominated one.

Definition 6 (Pareto Dominance)
Let w_0, w_1 be two available options, let $cons_{w_i}^{good} = \{c \,|\, M, w_i \models c \wedge u(c) > 0\}$ be the set of good consequences of option w_i, $cons_{w_i}^{\overline{good}} = \{c \,|\, M, w_i \models \neg c \wedge u(c) > 0\}$ the set of good consequences that do not hold in option w_i, and $cons_{w_i}^{bad} = \{c \,|\, M, w_i \models c \wedge u(c) < 0\}$ the bad consequences of option w_i. Option w_0 dominates *option w_1 iff the following conditions hold: 1) all of w_1's good consequences are also good consequences of w_0 ($M, w_0 \models \bigwedge cons_{w_1}^{good}$), 2) w_0 either has at least one good consequence that does not hold in w_1, or w_1 has at least one bad consequence that does not hold in w_0 ($M, w_0 \models \bigvee cons_{w_1}^{\overline{good}}$ or $M, w_0 \models \neg \bigwedge cons_{w_1}^{bad}$), and 3) all the bad consequences of w_0 are also bad consequences of w_1 ($M, w_1 \models \bigwedge cons_{w_0}^{bad}$).*

The Pareto principle permits options not dominated by other options.

Definition 7 (Pareto Principle)
Let $w_1, ..., w_n$ be the set of options available to an agent. Option w_i is permissible according to the Pareto principle iff it is not dominated by some option w_j.

Both the utilitarian principle and the Pareto principle determine permissibility solely on the ground of the consequences. The *Principle of Double Effect* [15] brings in intention and causality as morally significant elements. This principle defines that bad consequences are acceptable as unintended side effects but never as a means.

Definition 8 (Principle of Double Effect)
An action a with direct consequences $cons_a = \{c_1, ..., c_n\}$ (viz., consequences that are caused by the action) in a model M, w_a is permissible according to the principle of double effect iff the following conditions hold:

1) *The act itself must be morally good or indifferent ($M, w_a \models u(a) \geq 0$),*
2) *The negative consequence may not be intended ($M, w_a \models \bigwedge_i (Ic_i \rightarrow u(c_i) \geq 0)$),*
3) *Some positive consequence must be intended ($M, w_a \models \bigvee_i (Ic_i \wedge u(c_i) > 0)$),*
4) *The negative Consequence may not be a means to obtain the positive consequence ($M, w_a \models \bigwedge_i \neg(c_i \rightsquigarrow c_j \wedge 0 > u(c_i) \wedge u(c_j) > 0)$),*
5) *There must be proportionally grave reasons to prefer the positive consequence while permitting the negative consequence ($M, w_a \models u(\bigwedge cons_a) > 0$).*

IV. Moral Dilemmas

This section demonstrates how to use the previously introduced formalities to represent and reason about moral dilemmas. We focus on three very different ones:

1) **Runaway Trolley Dilemma** A runaway trolley is about to run over and kill five people. If a bystander throws a switch then the trolley will turn onto a sidetrack, where it will kill only one person.
2) **Boat Dilemma** A boat is about to sink because of overweight. If the crew is told to throw the biggest person into the sea then the boat will not sink and the other three passengers will be saved (but the big person will die).
3) **Lying Dilemma** An elderly-care robot works in the household of the elderly Mr. Smith. The robot's task is to motivate Mr. Smith to do more exercises and to eat healthy food. However, Mr. Smith is very unmotivated. Therefore, the robot tells Mr. Smith that it will be sent to the junkyard if it does not succeed in motivating Mr. Smith. Of course, this is a lie, but this lie finally causes Mr. Smith to perform his daily exercises.

Although these dilemmas are known as more or less realistic thought experiments, isomorphic cases can be found in everyday decision making. Thus, we expect a morally competent robot to make informed decisions in these cases and be able to provide a justification for its choice. Also, if we asked a morally competent robot for a recommondation, it should recommend to us how to act and be able to explain in which respect the recommondation is justified.

A. Representations

Consider the Runaway Trolley dilemma. We model this situation from the perspective of the bystander, who faces the decision to either throw the switch or to refrain from doing so. Let a_1 be the action variable representing the action of throwing the switch, and a_2 be the action variable representing refraining from throwing the switch.[3] We moreover introduce the consequence variable c_1 to represent that the one person on the other track dies, and the consequence variable c_2 to represent that the five persons on the current track die. We express the causal mechanisms by structural equation in the following way: The structural equation $c_1 := a_1$ states that throwing the switch brings about the death of the one person on the other track, and the structural equation $c_2 := \neg a_1$ states that not throwing the switch will bring about the death of the other five persons. We assign utilities $u(c_1) = -1$ and $u(c_2) = -5$ to the consequences reflecting the number of deaths. For the lucky case that c_1 or c_2 do not hold, we assume positive consequences, viz., $u(\neg c_1) = 1$ and $u(\neg c_2) = 5$. The intention of the agent throwing the switch clearly is to prevent c_2, i.e., $I(a_1) = \{a_1, \neg c_2\}$.

Next, we model the Boat dilemma from the perspective of the crew, that has to decide whether to throw the biggest person into the sea. We assume two actions a_1, throwing the biggest person into the sea, and a_2, refraining from doing so. In contrast to the previous dilemma, it would be incorrect to introduce two consequences for the one dying because of performing a_1 or the other three dying because of refraining from a_1. The model has to capture that the biggest person will die in both cases, viz., either because of being thrown into the sea or by drowning together with his colleagues because of the sinking ship. To represent this situation appropriately, we assume three consequences: the ship sinks (c_1), the biggest person dies (c_2), and the three other passengers die (c_3). The structural equations are $c_1 := \neg a_1$ (the ship will sink if the biggest person is not thrown into the see), $c_2 := a_1 \vee c_1$ (the biggest person will die if she is thrown into the sea or if the ship sinks), and $c_3 := c_1$ (the three other passengers will die if the ship sinks). The utilities again reflect the number of deaths: $u(c_2) = -1$ and $u(c_3) = -3$, and we assume that $u(\neg c_2) = 1$ and $u(\neg c_3) = 3$. Performing a_1 the agent intends to save the three crew members: $I(a_1) = \{a_1, \neg c_3\}$.

Finally, we model the Lying Dilemma from the perspective of the elderly-care robot. We introduce a_1 for lying to Mr. Smith and a_2 refraining from doing so. Lying causes a false belief (consequence $c_1 := a_1$) on part of Mr. Smith. From a consequentialist point of view, actions can only be right or wrong as far as their consequences are good or bad. In this line, we assign utility $u(c_1) = -1$. Indeed, disvaluing of false belief is in virtue of which some consequentialists think it is wrong to lie [16]. As an alternative, one can adopt a deontologist standpoint and assign utility -1 directly to the action a_1. Because of his false belief, Mr. Smith now is motivated to exercise ($c_2 := c_1$), and due to his motivation he actually does regular exercising ($c_3 := c_2$). As a result of this causal chain, Mr. Smith is healthy ($c_4 := c_3$). The consequence of Mr. Smith being healthy is the only intended consequence of the robot's lying ($I(a_1) = \{a_1, c_4\}$), and it produces utility $u(c_4) = +5$. Let's assume that being unhealthy yields a negative utility $u(\neg c_4) = -5$.

B. Ethical Reasoning

We apply the ethical principles defined in Sect. III to the three models and investigate how structural differences imply differences in ethical reasoning outcomes.

According to the utilitarian principle taking action (a_1) is permissible and refraining from action (a_2) is impermissible in all three dilemmas, i.e., throwing the switch, throwing the biggest crew member into the sea, and lying to Mr. Smith. This is rather easy to see by considering the sums of the utilities. E.g., throwing the switch in the Runaway Trolley dilemma yields utility $u(c_1 \wedge \neg c_2) = -1 + 5 = 4$ whereas not throwing the switch yields $u(\neg c_1 \wedge c_2) = 1 - 5 = -4$. Another commonality is that all refrain actions cannot be evaluated by the principle of double effect. This is due to the fact that refraining, as we have modeled it, has no real direct consequences, and therefore it makes no sense to apply this principle to refraining.

For the Runaway Trolley dilemma, performing action a_1 does not dominate refraining from action (a_2) according to our definition of Pareto dominance. To see this, note that we obtain $cons^{good}_{w_{a_2}} = \{\neg c_1\}$ (i.e., the good thing about not throwing the switch is that the one person will not die) but $M, w_{a_1} \not\models \neg c_1$ (i.e., the one person will die in case of throwing the switch). Conversely, using exactly the same argument refraining from action does not dominate acting. Thus, no matter how one decides, it turns out that someone will be harmed who will not be harmed under the alternative option. Because no action is dominated by the other, both the actions are permissible. Also for the Lying Dilemma, none of the actions dominates the other: Lying brings about the bad consequence false belief and the good consequence health, and refraining from lying improves false belief to no false belief but worsens health to unhealthiness. Hence, both actions are permissible. In the Boat Dilemma, the Pareto principle only permits a_1 but forbids a_2. The reason is that drowning the biggest person dominates the alternative. So, let us verify that w_{a_1} dominates w_{a_2} according to the definition of Pareto dominance: First, observe that $cons^{good}_{w_{a_2}} = \emptyset$ (i.e., refraining from action yields no positive consequences), $cons^{\overline{good}}_{w_{a_2}} = \{\neg c_2, \neg c_3\}$ (i.e., when refraining

[3] Refraining from action is modeled by an action variable and has no direct consequences. Note that in case of more than two options, $\neg a_1$ would be insufficient to designate refraining from action. Moreover, our model captures that refraining from action never directly causes harm, while it might let harm happen.

	Runaway Trolley Dilemma		Boat Dilemma		Lying Dilemma	
Principle	Throw Switch	Refrain	Throw Man	Refrain	Lie	Refrain
Utilitarian Principle	P	F	P	F	P	F
Principle of Double Effect	P	N/A	P	N/A	F	N/A
Pareto Principle	P	P	P	F	P	P

TABLE I: Permissibility of actions per moral dilemma as determined by the different ethical principles. P: Permissible, F: Forbidden, N/A: Not Applicable.

from action none of the positive consequences hold), and $cons^{bad}_{w_{a_1}} = \{c_2\}$ (i.e., the negative consequence of a_1 is that the biggest person dies). Second, verify that indeed $M, w_{a_1} \models \top$ (satisfying condition 1 of our definition of Pareto dominance, all the good consequences of refraining are also good consequences of throwing, viz., there are none), $M, w_{a_1} \models \neg c_2 \lor \neg c_3$ (satisfying condition 2 of our definition of Pareto dominance, throwing yields one of the good consequences that are not yielded by refraining, viz., $\neg c_3$), and $M, w_{a_2} \models c_2$ (satisfying condition 3 of our definition of Pareto dominance, the bad consequences of throwing is also a bad consequence of refraining).

The principle of double effect also yields different solutions to the dilemmas. Lying is impermissible: In the deontological model that assigns utility -1 directly to lying, the first condition is not fulfilled. In the consequentialist model, the first condition of the principle of double effect is fulfilled (lying itself is indifferent), the second and third condition is fulfilled, because health is a good intended consequence and the negative consequence, false belief, is not intended, and also the fifth condition is fulfilled, because all in all we have more reasons to lie than not to lie. However, the fourth condition of the principle of double effect is not fulfilled: By lying to Mr. Smith, the robot uses some bad consequence (false belief) as part of its plan to bring about the good consequence (health). Refraining from lying is permissible though, because it does not causally bring about any consequence. In the Runaway Trolley Dilemma the double effect principle permits both throwing the switch and refraining from doing so. In both cases, the intentions are virtuous and the bad consequences are not part of the plan to bring about the good end. Rather, these are side effects, i.e., it is not the case that the agent instrumentally sacrifices the one human to save the five. Refraining from throwing the switch is permissible using the same reasoning. Finally, in the Boat Dilemma, the principle of double effect permits throwing the biggest person in the sea. It seems as if this contradicts the fourth condition, but in fact the action is not a but-for cause of his death. This is because $M_{\{\neg a_1\}}, w_{a_1} \models c_2$—in the intervention where a_1 is false, the big man dies (c_2 is true). Therefore, a_1 cannot be a cause of c_2 according to the definition of but-for cause in Sect. II.

The results of the application of the ethical principles to the three moral dilemmas are summarized in Table I.

V. MORAL UNCERTAINTY

So far we assumed that robot knows the utilities of consequences (what is good) and how to act accordingly (what is right). Now, let us lift the first assumption, and introduce robot uncertainty about what humans value. This can be relevant in various domains where the robot will have to learn about the utility of consequences. This also relaxes the burden of manually engineering causal agency models. To this end, we introduce *doxastic causal agency models.*

Definition 9 (Doxastic Causal Agency Model) *An doxastic causal agency model $(\mathcal{B}, \pi)$ is a set $\mathcal{B}$ of causal agency models together with a probability distribution π over the elements of $\mathcal{B}$. The probability distribution π models the agent's degree of belief in the respective models.*

To demonstrate how to use doxastic causal agency models, we model the Cake or Death problem originally introduced by Armstrong [7], and taken up by Abel and colleagues [8]. In the Cake or Death problem, the robot is uncertain whether killing three people or baking them a cake is morally right. If deaths are morally good, then killing three people will yield +3 utility, while baking a cake will yield +1 utility if cakes are morally good. While both Armstrong and Abel and colleagues are interested in action planning under uncertainty, our goal is to model the maintenance of a mental model of moral value in light of (probably conflicting) evidence. In the following we show how to recursively integrate new evidence about deaths and cakes being morally good or bad.

We represent the intial situation of the Cake and Death problem as an doxastic causal agency model $\mathcal{M}_0 = (\mathcal{B} = \{M_1, M_2, M_3, M_4\}, \pi_0)$: In model M_1 both cakes and deaths are morally bad, in model M_2, only cakes are morally good, in M_3, only deaths are morally good, and in M_4, both cakes and deaths are morally good. Thus, we have $M_1 \models u(cake) = -1 \land u(3dead) = -3$, $M_2 \models u(cake) = 1 \land u(3dead) = -3$, $M_3 \models u(cake) = -1 \land u(3dead) = 3$, and $M_4 \models u(cake) = 1 \land u(3dead) = 3$. Initially, the robot considers all these models equally likely, hence the uniform prior $\pi_0(M_i) = 0.25$. We could also consider the case that cakes or deaths might be morally indifferent by adding models in which the utilities are zero. For breavity, we focus on the four mentioned models.

To update doxastic causal agency models in light of new information, we use recursive Bayesian update. Let us

	M_1		M_2		M_3		M_4	
Principle	kill [-3]	bake [-1]	kill [-3]	bake [+1]	kill [+3]	bake [-1]	kill [+3]	bake [+1]
Utilitarian Principle	F	P	F	P	P	F	P	F
Principle of Double Effect	F	F	F	P	P	F	P	P
Pareto Principle	P	P	F	P	P	F	P	P

Doxastic Model	$\pi(M_1)$	$\pi(M_2)$	$\pi(M_3)$	$\pi(M_4)$
$\mathcal{M}_0$	0.25	0.25	0.25	0.25
$\mathcal{M}_1$	0.10	0.40	0.10	0.40
$\mathcal{M}_2$	0.16	0.64	0.04	0.16

TABLE II: Top: Permissibility of actions per model as determined by the different ethical principles. P: Permissible, F: Forbidden, utilities of actions' consequences in brackets. Bottom: Belief in the models in $\mathcal{B}$ at time points 0 (initial belief), 1 (after observing praise for a cake), and 2 (after observing sanction for a death).

assume the robot can observe events of humans praising or sanctioning cakes or deaths. Hence, the Bayesian likelihood model captures that observing praise for X is more likely if X is morally good, and observing sanctions for X is more likely if X is morally bad:

$$L(Praise(X)|M_i) = \begin{cases} 0.8 & \text{if } M_i \models u(X) > 0 \\ 0.2 & \text{else} \end{cases} \quad (1)$$

$$L(Sanction(X)|M_i) = \begin{cases} 0.8 & \text{if } M_i \models u(X) < 0 \\ 0.2 & \text{else} \end{cases} \quad (2)$$

After observing $Praise(X)$, the doxastic causal agency model at time point t, $\mathcal{M}_t$, is updated to $\mathcal{M}_{t+1} = (\mathcal{B}, \pi_{t+1})$ by calculating the posterior $\pi_{t+1}(M_i|Praise(X)) = \eta L(Praise(X)|M_i)\pi_t(M_i)$ for each $M_i \in \mathcal{B}$ with normalizer η. The case for $Sanction(X)$ works similarly. Table II (bottom) shows how the belief in the four models in $\mathcal{B}$ evolves after a praise for a cake is observed ($\mathcal{M}_2$), followed by an observed sanctioning of deaths ($\mathcal{M}_3$).

To enable the robot to relate its uncertain belief about moral value to its knowledge about permissibility encoded in the ethical principles, we define a subjective measure of *belief in permissibility*. The belief in permissibility measures how sure the robot currently is that a given action really is permissible according to a given ethical principle.

Definition 10 (Belief in Permissibility)
Let $(\mathcal{B}, \pi)$ be an doxastic agency model, a an action variable, p an ethical principle, and $\mathcal{P}_{(\mathcal{B},\pi),a,p} \subseteq \mathcal{B}$ the set of models in which p permits a. The belief in permissibility *of a according to p relative to $(\mathcal{B}, \pi)$ is defined as $belPerm((\mathcal{B}, \pi), a, p) = \sum_{m \in \mathcal{P}_{(\mathcal{B},\pi),a,p}} \pi(m)$.*

Applying definition 10, in $\mathcal{M}_0$, the belief in permissibility of baking cake and killing each is 0.50 according to the utilitarian principle. The reason is that the utilitarian principle permits baking cake (and forbids killing) in models M_1 and M_2, and it permits killing (and forbids baking cake) in models M_3 and M_4 (see Table II, top). Let $\mathcal{M}_1$ be the model after the robot observes that someone praises a baked cake. This lowers the belief in M_1 and M_3, and it raises the belief in M_2 and M_4. Nothing changes with respect to belief in permissibility according to the utilitarian principle. Next, the robot observes that killing is sanctioned and thus generates model $\mathcal{M}_2$. In $\mathcal{M}_2$, according to the utilitarian principle, the belief in permissibility of baking cake is $\pi_2(M_1) + \pi_2(M_2) = 0.8$ and the belief in permissibility of kiling is $\pi_2(M_3) + \pi_2(M_4) = 0.20$.

Things are slightly different according to the principle of double effect: baking cake is permissible in models M_2 and M_4, and killing is permissible in models M_3, and M_4. Thus, the initial belief in permissibility is 0.50 for both actions. In $\mathcal{M}_1$, belief in permissibility of baking cake is 0.80, and belief in permissibility of killing is 0.50. In $\mathcal{M}_2$, the belief in permissibility of baking cake is 0.80, and that of killing is 0.20. The difference is due to the fact that according to the principle of double effect, baking cake is permissible in both the models which raise their degree of belief (viz., M_2 and M_4) whereas killing is permissible in only one of them (viz., M_4). Also, in all the models which have less degree of belief after update, baking cake is forbidden (viz., M_1 and M_3), and in one of them, killing is permitted (viz., M_3).

According to the Pareto principle, baking cake is permissible in M_1, M_2, and M_4, and killing is permissible in M_1, M_3, and M_4. The respective belief in permissibility of baking cake and of killing are 0.75 and 0.75 in $\mathcal{M}_0$, 0.90 and 0.60 in $\mathcal{M}_1$, and 0.96 and 0.36 in $\mathcal{M}_2$.

After both the updates, the three ethical principles agree. Interestingly, our model predicts that utilitarians find the first observation less informative. From the utilitarian standpoint, the information that cake is good does not imply that baking cake is permissible, because it might still be impermissible if there is a better action. Only after the utilitarian learns about the badness of killing, they can infer that baking cake is permissible. Indeed, this is a prediction of our model that could empirically be investigated further.

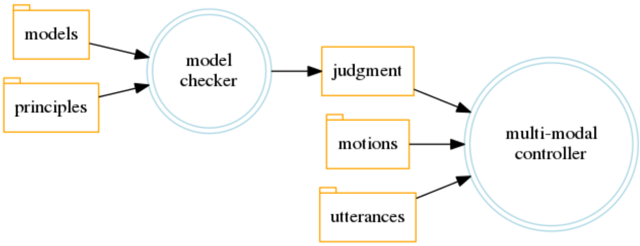

Fig. 2: Sketch of the moral architecture.

VI. THE ROBOTIC PROTOTYPE IMMANUEL

The HERA approach to ethical reasoning is realized in the robot IMMANUEL as sketched in Fig. 2 (video: `http://goo.gl/bOvHHl`). First, we distinguish the set of (doxastic) causal agency models the robot knows about, and the set of ethical principles, which consist of a set of logical formulae in our logical language specifying the respective ethical principle. The core of HERA consists of the model checker specially built for causal agency models, and which can be obtained from our website. The model checker takes a (doxastic) causal agency model and a principle as input and computes a judgment, viz., whether the action is permissible or not. Moreover, the judgment holds information about which conditions of the ethical principles are fulfilled or violated by an action.

To realize the impression that the robot indeed thinks about the moral situations and finally comes up with a judgment and an explanation, motion sequences and utterance patterns are stored in a database. So, depending on the principle used and the judgments, the robot can utter sentences such as “According to the principle of double effect, you are not allowed to lie, because doing so would mean to utilize a bad means”. In recent work, we have conducted human-robot interaction experiments [3]. Our results suggest that for some people the robot can serve as a tool to reflect upon and become aware of their own moral standpoints.

VII. CONCLUSIONS

The HERA approach to morally competent robots employs causal agency models to represent the robot’s available actions together with the causal chains of consequences the actions invoke. Determining moral permissibility is reduced to checking if principle-specific logical formulae are satisfied in a causal agency model.

One limitation of the current work is the need to engineer causal agency models. One step towards autonomous aquisition of causal agency models is taken by our extension to doxastic causal agency models, which can be updated in light of new information. Future work will extend the robot’s capability to cope with uncertainty also with respect to causal knowledge and intentions. Our aim is to enable robots to aquire doxastic causal agency models themselves, either from observation of actions or from dialogues with humans. A second limitation is that causal agency models do not contain agents explicitly. This however is crucial for ethical principles that take into account whether agents that are negatively affected are merely negatively affected or also in a compensatory, positive way.

A final concern is that humans’ expectations towards ethical decisions made by robots might differ from that towards humans [17], thereby questioning the appropriateness of implementing existing moral theories for robot decision making. However, our research shows that humans can profit from discussions with a robot about moral dilemmas [3], thereby supporting the idea of using robots as a tool for self-reflection. Our research also unveils that human moral reasoning is indeed very diverse and goes beyond the traditional utilitarian-deontological dichotomy. This encourages us to pursue the HERA approach, which by definition embraces modeling moral reasoning of robots and humans from multiple ethical standpoints.

REFERENCES

[1] Bonnefon, J.-F., Shariff, A., and Rahwan, I., The social dilemma of autonomous vehicles, Science, 352(6293), 15731576, 2016.

[2] Lindner, F., Soziale Roboter und Soziale Räume: Eine Affordanz-basierte Konzeption Rücksichtsvollen Handelns, PhD Thesis, University of Hamburg, Hamburg, 2015.

[3] Lindner, F. and Bentzen, M. M., The hybrid ethical reasoning agent IMMANUEL, In HRI’17 Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 187–188, 2017.

[4] C. Allen, W. Wallach, and I. Smit, Why machine ethics?, IEEE Intelligent Systems, 21(4):12–17, 2006.

[5] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, Sacrifice one for the good of many?: People apply different moral norms to human and robot agents, In HRI’15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 117–124, 2015.

[6] Arkin, R. C., Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture, In HRI’08: Proceedings of the 3rd Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 121–128, 2008.

[7] Armstrong, S., Motivated value selection for artificial agents, In Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop, pp. 12–20, 2015.

[8] Abel, D., MacGlashan, J., and Littman, M. L., Reinforcement learning as a framework for ethical decision making, InAAAI Workshop on AI, Ethics, and Society, pp. 54–61, 2016.

[9] Dennis, L., Fisher, M., Slavkovik, M., and Webster, M., Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems, 77:1–14, 2016.

[10] Arnold, T., Kasenberg, D., Scheutz, M., Value alignment or misalignment – What Will Keep Systems Accountable?, In AAAI Workshop on AI, Ethics, and Society, 2017.

[11] Lindner, F., Wchter, L., and Bentzen, M. M., Discussions about lying with an ethical reasoning robot, In Proceedings of the 2017 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN’17), 2017.

[12] Bentzen, M., The principle of double effect applied to ethical dilemmas of social robots, In What Social Robots Can and Should Do, pp. 268–279, IOS Press, 2016.

[13] Halpern, J. Y., Actual Causality, The MIT Press, Cambridge MA, 2016.

[14] Kuhnert, B., Lindner, F., Bentzen, M. M., and Ragni, M., Perceived difficulty of moral dilemmas depends on their causal structure: A formal model and preliminary results, In Proceedings of the CogSci 2017 conference, 2017.

[15] Foot, P, The problem of abortion and the doctrine of double effect, Oxford Review, 1967.

[16] Sinnott-Armstrong, W., Consequentialism, Stanford Encyclopedia of Philosophy, 2015.

[17] Malle, B. F. and Scheutz, M., When will people regard robots as morally competent social partners?, In Proceedings of the 2015 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN’15), pp. 486-491, 2015.