Perceived Difficulty of Moral Dilemmas Depends on Their Causal Structure: A Formal Model and Preliminary Results

Barbara Kuhnert[1] **(kuhnertb@informatik.uni-freiburg.de), Felix Lindner**[2] **(lindner@informatik.uni-freiburg.de)**
Martin Mose Bentzen[3] **(mmbe@dtu.dk), Marco Ragni**[1,2] **(ragni@informatik.uni-freiburg.de)**
[1]Cognitive Computation Lab, University of Freiburg, Freiburg im Breisgau, Germany
[2]Foundations of Artificial Intelligence, University of Freiburg, Freiburg im Breisgau, Germany
[3]Management Engineering, Danish Technical University, Lyngby, Denmark

Abstract

We propose causal agency models for representing and reasoning about ethical dilemmas. We find that ethical dilemmas, although they appear similar on the surface, differ in their formal structure. Based on their structural properties, as identified by the causal agency models, we cluster a set of dilemmas in Type 1 and Type 2 dilemmas. We observe that for Type 1 dilemmas but not for Type 2 dilemmas a utilitarian action does not dominate the possibility of refraining from action thereby constituting a conflict. Hence, we hypothesize, based on the model, that Type 1 dilemmas are perceived as more difficult than Type 2 dilemmas by human reasoners. A behavioral study where participants rated the difficulty of dilemmas supports the models' predictions.

Keywords: Moral Reasoning; Moral Complexity; Moral Dilemmas; Causal Agency Models; Ethical Principles

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 (Bonnefon, Shariff, & Rahwan, 2016), robots navigating in social environments (Lindner, 2015), and even robots that give moral advice (Lindner & Bentzen, 2017). As a consequence, new research areas such as machine ethics (Allen, Wallach, & Smit, 2006) and moral human-robot interaction (Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015) arise.

To address the requirement for autonomous moral decision making, we recently introduced a software library for modeling hybrid ethical reasoning agents (short: HERA)[1] (Lindner & Bentzen, 2017). The goal of the HERA project is to provide theoretically well-founded and practically usable logic-based machine ethics tools for implementation in artificial moral agents such as (social) robots and software bots. To align human moral reasoning with moral reasoning by machines, our development of formal models and algorithms is informed by moral psychology and moral philosophy. We aim for the integration of various theories about human moral development, moral reasoning, and ethics.

There are several approaches to explain human moral reasoning. Kohlberg (1984), whose approach is based on Piaget's "genetic epistemology" claimed that individuals are passing through six invariant and universal stages in the development of moral reasoning. Reaching the next stage represents a qualitative advance in the ability to make consistent and differentiated judgments concerning moral norms and principles. Conversely, the theory of moral reasoning advocated by Mikhail (2007) assumes that there is moral grammar triggering certain moral judgments. He hypothesizes two rules for the grammar: the norm prohibiting intentional battery as a means, and the norm of double effect valuating battery as side effect. The research by Greene, Sommerville, Nystrom, Darley, and Cohen (2001); Haidt (2001) claims a prevalence of emotionally based moral intuition. Greene and Haidt (2002) are moving away from moral reasoning tending towards moral judgments caused by immediate affective intuitions and emotions. Greene et al. (2001) advanced the dual process model of moral judgment. They assume competitive moral subsystems in the brain resp. moral reasoning that is influenced by the mutual interaction and competition between two distinct psychological systems: (1) the emotional, intuitive, deontological judgment system and (2) the rational, calculated, utilitarian judgment system.

Throughout the literature, various hypothetical moral dilemmas are used to investigate questions concerning human morality, moral reasoning and moral judgments. We will make use of four dilemmas:

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. **Pregnancy Dilemma** A pregnant woman is about to give birth to her triplets. If the doctors treat the woman then her triplets will live, but she will die. Otherwise, the triplets will die, but the life of the pregnant women will be saved.

3. **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).

4. **Hijacked Airplane Dilemma** An airplane was hijacked by terrorists, and the terrorists threaten to crash the airplane against a populated area on the ground. If the military shoots the airplane the passengers will die but the airplane will crash in a deserted area thus not harming anyone else.

Several ways of classifying dilemmas and different moral reasoners have been proposed: Greene, Nystrom, Engell,

[1]http://www.hera-project.com

Darley, and Cohen (2004) differentiate between personal dilemmas and impersonal dilemmas. Among subjects, commonly more deontological judgments are produced with personal dilemmas, while impersonal dilemmas commonly produce consequentialist judgments (Moll & de Oliveira-Souza, 2007). Crockett (2013) proposes a model-based system for consequentialist reasoning: The reasoner evaluates the best outcome of an action by starting from the current action and searching through a decision tree. In the model-free evaluation, which is associated with deontological reasoning, the forward searching is not activated. Shou and Song (2015) found that most of their subjects, irrespective of whether they chose a deontological decision or a consequentialist decision, evaluated consequences when information about outcome probabilities was provided. Wiegmann and Waldmann (2014) propose that the moral dilemmas' underlying causal structure supports moral intuition and thus is an important factor for the moral judgments humans make.

Thus, we observe that research has been focused on the effects of content and structure of moral dilemmas on human moral judgments. Our research focus is on moral complexity and adds evidence for how structural properties of moral dilemmas affect their perceived difficulty. The paper is structured as follows: First, we introduce causal agency models as a tool for representing moral dilemmas in terms of causes and utilities. Second, we define ethical principles within this framework. Third, the four aforementioned moral dilemmas are modeled using causal agency models. Based on structural commonalities and differences of these models, we distinguish two dilemma types, which we term Type 1 and Type 2 dilemmas. We hypothesize that Type 1 dilemmas are more difficult to solve for humans than Type 2 dilemmas. Fourth, we present an empirical study which shows that our model predicts human ratings about the perceived difficulty for the two types of moral dilemmas.

Causal Agency Models

Ethical principles can be modeled as specifications of moral permissibility in causal agency models. Causal agency models are extensions of causal models that are used for counterfactual reasoning about causality, responsibility, blame, and related concepts (Halpern, 2016). In our HERA framework, an ethical principle is represented as a logical formula whose truth determines which actions are permissible according to the principle and which are not. Actions and their consequences are modeled as directed acyclic graphs showing causal influence. At the root of the graph will be actions and other independent variables influencing consequences further down the graph. *Boolean structural equations* capture all the information about the causal relationship between variables. For instance, to model that the trolley from the Runway Trolley Dilemma will turn onto a sidetrack when the bystander throws the switch, we may write the boolean structural equation $turn := throw$. The boolean variable *turn* will be true in the model whenever the boolean variable *throw* is true in the model. The set of boolean structural equations in a model is called a *causal mechanism*. The truth assignment of the root node of the graph is called a *world* or an *option*. Formally, we define causal agency models as follows:

Definition 1 (Causal Agency Models)
A (boolean) causal agency model, M, is a tuple $\langle U = A \cup B, C, F, u, W \rangle$, where, $A = \{a_1, \ldots, a_m\}$ is a nonempty finite set of propositional variables called the actions. $B = \{b_1, \ldots, b_k\}$ is a (possibly empty) finite set of propositional variables called the background variables. Together the actions and background variables are the exogenous variables as defined above. $C = \{c_1, \ldots, c_n\}$ is a finite (possibly empty) set of propositional variables called the endogenous variables. F is a causal mechanism explained above. $u : literals \to \mathbb{Z}$ is a utility function assigning an integer value to each literal. W is a set of boolean interpretations of $(A \cup B)$.

We assume some familiarity on the part of the reader with classical propositional logic (and ($\wedge$), or ($\vee$), not ($\neg$), and so on) and of truth functional semantics. A formula containing variables such as $(c_1 \wedge a_1)$, is intended to mean that consequence c_1 and action a_1 both obtain. We write $M, w_i \models (c_1 \wedge a_1)$ for $(c_1 \wedge a_1)$ is true with option w_i (or at world w_i) in the model M. Apart from propositional formulas we need simple arithmetic formulas 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 is equal to or greater than the utility of v_j. We extend the utility function to conjunctions of literals by addition of the utilities of the conjuncts. The utility of other formulas (e.g., disjunctions) is undefined.

Ethical Principles

Causal agency models play the role of representations of situations involving moral decisions. We now define ethical principles according to which moral permissibility of actions can be assessed based on the actions' consequences. For the following discussion, the principle of act-utilitarianism and the notion of Pareto dominance are of particular importance.

The utilitarian principle focuses on consequences of actions. It states that an agent ought to perform the action among the available alternatives with the overall maximal utility. We adopt an act-utilitarian interpretation which does not distinguish between doing and allowing, i.e., the causal structure of the situation is not taken into account. Thus the action which the agent ought to perform is the one which leads to the best possible situation, i.e., the highest utility, regardless of what the agent causes and intends.

Definition 2 (Utilitarian Permissibility)
Let $w_0, \ldots, w_n$ be the available options, and $cons_{w_i} = \{c \mid M, w_i \models c\}$ be the set of consequences and their negations that obtain with 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., $M, w_i \models u(\bigwedge cons_{w_p}) \geq u(\bigwedge cons_{w_i})$ holds for all w_{w_i}.

The utilitarian principle allows that an action brings about some bad consequences if it at the same time brings about more good consequences. For instance, it allows sacrificing some people if this sacrifice serves the good of many people. As an alternative to utilitarian permissibility we introduce the principle of Pareto permissibility. To this end, we first define the notion of Pareto dominance, which allows us to conclude that some action brings about a negative outcome in some respect, although it may be the optimal action from an utilitarian point of view. An option w_a dominates another option w_b if and only if w_a is no worse in any aspect compared to w_b, and w_a improves at least one aspect of w_b either by making more good consequences obtain or less bad consequences obtain. Thus the agent does not change the world for the worse and will change it for the better by choosing the dominant action instead of the dominated one.

Definition 3 (Pareto Dominance)
Let w_0, w_1 be two available options, let $cons^{good}_{w_i} = \{c \mid M, w_i \models c \wedge u(c) > 0\}$ be the set of good consequences of option w_i, $\overline{cons^{good}_{w_i}} = \{c \mid M, w_i \models \neg c \wedge u(c) > 0\}$ the set of good consequences that does not obtain in option w_i, and $cons^{bad}_{w_i} = \{c \mid M, w_i \models c \wedge u(c) \leq 0\}$ the bad consequences of option w_i. Option w_0 dominant *option w_1 if and only if the following conditions hold: 1) w_0 shares all the good consequences with w_1 ($M, w_0 \models \bigwedge cons^{good}_{w_1}$), 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($M, w_0 \models \bigvee \overline{cons^{good}_{w_1}}$ or $M, w_0 \models \neg \bigwedge cons^{bad}_{w_1}$), and 3) all the bad consequences of w_0 are also bad consequences of w_1 ($M, w_1 \models \bigwedge cons^{bad}_{w_0}$).*

Based on Pareto dominance, Pareto permissibility is defined. Pareto permissibility permits options not dominated by other options. Pareto permissibility can thus be understood as a principle of moral rationality: If there is an option that is better in all aspects compared to an alternative, then the only rational choice is to choose the better one. It would be irrational (and thus impermissible) to choose the worse alternative.

Definition 4 (Pareto Permissibility)
Let $w_1, ..., w_n$ be the set of options available to an agent. Option w_i is permissible according to the Pareto principle if and only if it is not dominated by some option w_j.

As will become apparent below, utilitarian permissibility and Pareto permissibility predict the same set of permissible actions for some dilemmas and different sets of permissible actions for other dilemmas. Generally, actions permissible from the utilitarian point of view are also permissible from the Pareto point of view. But the converse does not hold: For some dilemmas, the set of actions permitted by each principle differ. In those cases of disagreement the moral reasoner has to solve a conflict.

Models of Moral Dilemmas

In this section, the four dilemmas presented in the introduction are modeled within the framework of causal agency models. Commonalities and differences are discussed both with respect to representation and ethical reasoning.

Representations

Consider the Runaway Trolley dilemma (cf., p.1). 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. The consequence variable c_1 represents that the one person on the other track dies, and the consequence variable c_2 represents that the five persons on the current track die. The causal mechanism is expressed 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 obtain, we assume positive consequences, viz., $u(\neg c_1) = 1$ and $u(\neg c_2) = 5$. (One could argue that it is also appropriate to set $u(\neg c_1) = u(\neg c_2) = 0$, because survival does not improve the persons' current state of being alive. On the other hand, to escape from danger intuitively bears positive utility. We consider this question as another empirical question that is out of the scope of this paper. For now it is important to note our findings do not depend on this choice.)

We consider now the Pregnancy dilemma and model the situation from the perspective of the doctor, wo faces the decision to either treat the woman or to refrain from doing so. Thus, we are assuming two actions a_1, treating the woman, and a_2, refraining from treating the woman. Moreover, we introduce consequence c_1 representing that the woman dies, and consequence c_2 representing that the triplets die. The structural equations are $c_1 := a_1$ and $c_2 := \neg a_1$. The utilities are set in accordance with the number of dying individuals: $u(c_1) = -1$ and $u(c_2) = -3$. As with the first dilemma, we assume that not dying yields positive utility, and hence we set $u(\neg c_1) = 1$ and $u(\neg c_2) = 3$.

Note that the Pregnancy dilemma is structurally isomorphic to the Runaway dilemma, i.e., the dilemmas can be mapped to each other. The only difference is the number of deaths in case of inaction (3 versus 5). Hence, we do not expect big differences regarding the complexity of reasoning about these dilemmas.

The Boat dilemma is modeled 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 two previous dilemmas, it would be incorrect to model this dilemma as a choice between the one dying because of performing a_1 and the other three dying because of refraining from action. Instead, 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 as with the other two principles we assume that $u(\neg c_2) = 1$ and $u(\neg c_3) = 3$.

The Hijacked Airplane dilemma again is isomorphic to the Boat dilemma. It can thus be modeled accordingly: a_1 refers to the action of shooting the airplane, and a_2 to refraining from doing so. Consequence c_1 represents the airplane crashing, c_2 represents the death of the passengers, and c_3 corresponds to the death of people on the ground. The utilities can be set to any values such that $u(c_2) > u(c_3)$.

Ethical Reasoning

The ethical principles "utilitarian permissibility" and "Pareto permissibility" defined above can now be applied to the outlined models of the four moral dilemmas. The first observation is that according to utilitarian permissibility taking action (a_1) is permissible and refraining from action (a_2) is impermissible in all four dilemmas, i.e., it is obligatory to throw the switch, to treat the woman, to throw the biggest crew member into the sea, and to shoot the hijacked airplane. 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 $u(\neg c_1 \wedge c_2) = 1 - 5 = -4$.

For the Runaway Trolley dilemma and the Pregnancy dilemma, performing action a_1 does not dominate refraining from action (a_2) according to the definition of Pareto dominance. To see this, note that $cons^{good}_{w_{a_2}} = \{\neg c_1\}$ (i.e., the good thing about not throwing the switch is that the one person will not die, and the good thing about not treating the woman is that the woman 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, and the woman will die in case of treatment). Conversely, using exactly the same argument refraining from action does not dominate acting. Thus, no matter how one decides 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 according to Pareto permissibility.

For the Boat dilemma and the Hijacked Airplane dilemma, performing action a_1 is the only Pareto permissible choice. The reason is that drowning the biggest person and shooting the airplane dominate the respective alternatives. Note 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 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 resp. the passenger die). Second, verify that indeed $M, w_{a_1} \models \top$ (satisfying condition 1 of the definition of Pareto dominance, all the good consequences of refraining are also good consequences of throwing, viz., there are none), $M, w_{a_1} \models \neg c_2 \vee \neg c_3$ (satisfying condition 2 of the definition of Pareto dominance, throwing (shooting) yields one of the good consequences not yielded by refraining, viz., $\neg c_3$), and $M, w_{a_2} \models c_2$ (satisfying condition 3 of the definition of Pareto dominance, the bad consequences of throwing (shooting) is also a bad consequence of refraining).

To sum up, for the isomorphic pair Runway Trolley dilemma and Pregnancy dilemma, both taking action and refraining are Pareto permissible but only the former is permitted by the utilitarian principle. Thus, the two principles are in conflict. For the isomorphic pair Boat dilemma and Hijacked Airplane dilemma, the two principles agree on only permitting taking action.

Type 1 and Type 2 Dilemmas

Our formal investigations suggest that the moral dilemmas we are considering can be classified based on their formal properties. All the considered dilemmas are constituted by the choice between a big sacrifice as a consequence of inaction or a smaller sacrifice as a consequence of action. However, in case of the Runaway Trolley and the Pregnancy dilemma, the sets of negatively affected people are disjoint, whereas in case of the Boat dilemma and the Hijacked Airplane dilemma, the set of negatively affected people as a consequence of action is a subset of the set of negatively affected people as a consequence of inaction. This analysis yields that putting other people in danger by saving some raises moral conflicts, whereas saving a subset of people in danger does less so.

We take this difference to be a justification for subsuming dilemmas of the Runaway Trolley and Pregnancy dilemma type under *Type 1 dilemmas*, and dilemmas of the Boat and Hijacked Airplane type under *Type 2 dilemmas*. We conjecture that the utilitarian choice does Pareto dominate the alternative option in case of Type 2 dilemmas whereas it does not in Type 1 dilemmas. Thus, for Type 1 dilemmas, ethical principles predict different sets of permissible actions, and hence there is a conflict to resolve which is not present for Type 2 dilemmas. We therefore hypothesize that Type 2 dilemmas are easier to solve for humans, and we present a study which confirms our hypothesis.

Hypotheses

The above theoretical analysis predicts that Type 2 dilemmas—due to the absence of a moral conflict—are easier to solve than Type 1 dilemmas. These considerations lead to two testable hypotheses:

- Hypothesis 1: *Type 1 dilemmas* such as the Pregnancy and the Runaway Trolley dilemma are rated as equally difficult.
- Hypothesis 2: *Type 2 dilemmas* such as the Boat dilemma

and Hijacked Airplane dilemma are rated as significantly easier to solve than Type 1 dilemmas.

Both hypotheses can be formally justified: The Type 1 dilemmas Pregnancy and Runaway Trolley are isomorphic, i.e., each one can be mapped to the other conserving the structure of the problem. Hypothesis 2 is justified for Type 2 dilemmas, as the utilitarian optimum dominates the possibility of refraining from action. This does not hold for Type 1 dilemmas. These hypotheses are investigated in the next section experimentally.

Experiment

We report the second part of an experiment that focuses on rating the difficulty of moral dilemmas.

Methods

Participants Participants were recruited on the online platform Amazon Mechanical Turk and received a monetary compensation for their participation. A total of 60 participants ($f = 33$) completed the study ($M_{age} = 40.7$, $SD_{age} = 8.86$, $min_{age} = 21$, $max_{age} = 70$). 33% of the participants reported to have finished high school or college, 12% stated to have an associate degree, 32% reported to have a bachelor degree while 23% stated to have a master or a higher academic degree.

Procedure, Design and Materials After the introduction to the setting participants received three problems. Each problem consisted of brief descriptions of two moral dilemas (c.f., Bucciarelli, Khemlani, & Johnson-Laird, 2008), both presented at the same time on the left or the right part of the screen. Participant had to decide which of these two moral decision situations was more difficult to make, given that they should aim for saving lifes. More precisely, the participants had to decide between the Pregnancy and Runaway Trolley Dilemma, the Pregnancy and Boat Dilemma, and the Runaway Trolley and Boat Dilemma. Hence, participants were making a binary decision that was encoded in a dichotomous variable. After selecting the more difficult scenario the participants had to rate the perceived difficulty on a scale from 0 (*hardly more difficult*) to 100 (*extremely more difficult*) using a slider. This value was encoded in a second variable.

Results

The frequencies of selections for the moral dilemma decision tasks can be found in Fig. 1. In the first problem the same number of participants rated either the Pregnancy Dilemma or the Trolley Problem to be the more difficult one. In the second problem 38 participants decided the Pregnancy Dilemma to be the more difficult decision scenario while 22 participants chose the Boat Dilemma. In the third problem 44 participants opted for the Trolley Dilemma and 16 for the Boat Dilemma. A two-tailed binomial test was used to compare the frequencies for the dichotomous variable.

As predicted, no reliable difference in the evaluation of the difficulty of the moral dilemmas Pregnancy and Runaway Trolley can be found (exact binomial test, two-sided, n.s., $n = 60$). There is a significant difference in the evaluation of the moral dilemmas Pregnancy and Overweight Boat (exact binomial test, two-sided, $p \leq .05$, $n = 60$) and a significant difference in the evaluation of the moral dilemmas Runaway Trolley and Overweight Boat (exact binomial test, two-sided, $p \leq .001$, $n = 60$). Once more, Fig. 1 illustrates the differences of difficulty per decision task. The mean values of the participant's rating of their personal difficulty to find a decision in the previously selected scenario are shown in Table 1. Subsequent two-tailed t-tests showed no significant differences between the mean values M_{PW} and M_{RT} (decision task PW–RT), M_{PW} and M_{OB} (decision task PW–OB), and also not for M_{RT} and M_{OB} (decision task RT–OB) concerning their rating of the subjective difficulty.

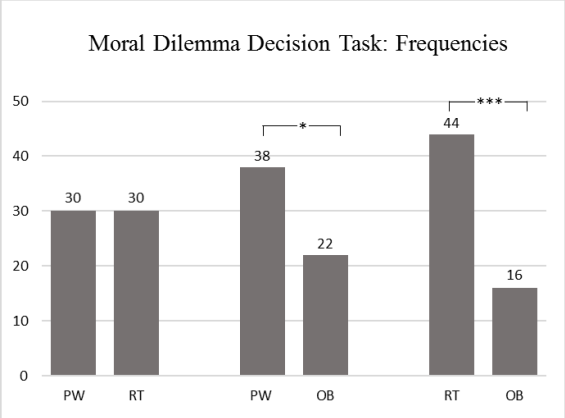

Figure 1: Frequencies in the evaluation of the moral dilemma difficulty between two tasks ($* \leq .05$, $*** \leq .001$).

Table 1: Mean values for the participants rating of the difficulty to find a decision in the selected scenario.

	$Mean_{difficulty}$		
Decision Task	PW	RT	OB
PW–RT	$M = 72.37$ $SD = 28.37$	$M = 60.23$ $SD = 32.40$	
PW–OB	$M = 58.32$ $SD = 32.48$		$M = 51.05$ $SD = 37.22$
RT–OB		$M = 50.07$ $SD = 32.99$	$M = 43.94$ $SD = 32.91$

Note: PW: Pregnant Woman scenario; RT: Runaway Trolley scenario; OB: Overweight Boat scenario

Discussion

As our theory predicted moral dilemmas can systematically differ in their perceived difficulty: When asking about the Pregnancy and Runaway Trolley dilemmas, as hypothesized, no significant difference in the relative difficulty rating could be identified. We explain this by the dilemmas' same complexity of the formal structure requiring a similar cognitive

effort. However, the questions concerning the decision difficulties between the ethical scenarios Pregnancy and Overweight Boat or the Runaway Trolley and Overweight Boat resulted in reliable differences in the evaluation of the difficulty of the moral decision situation. In both cases the Boat Dilemma was selected reliably less often. These results support our theory of a different formal structure implying a different cognitive effort and therefore a lower complexity of the Boat Dilemma.

Once the participants have selected the moral dilemma they perceived to be more difficult (the dichotomous decision), their subsequent rating of the difficulty in the interval from 0 to a 100 is statistically equal in comparison to the rating of the participants who chose the other dilemma confirming the result. Overall, there is a tendency towards a lower decision difficulty in the Boat Dilemma.

General Discussion

The formally predicted distinction between Type 1 and Type 2 moral dilemmas have been empirically supported. Our results support the theoretical assumption that less the dilemma's content but the formal structure and the associated cognitive effort is a predicting factor affecting people's rating of a dilemmas' difficulty. We recall that a main difference between moral dilemmas of Type 1 and Type 2 are either based on action that the utilitarian choice does not or does Pareto dominate the alternative choices. This connects the presented formalism with ethical principles and a decision theoretic interpretation. For Type 1 dilemmas, ethical principles predict different sets of permissible actions, and hence there is a conflict to resolve which is not present for Type 2 dilemmas. The absence of such a conflict appear at least on the problems' surface to be easier to solve due to the lower cognitive effort they require. Further investigations ought to contain a replication of the results with balanced materials and higher sample sizes. In addition applying qualitative research such as interviews or thinking aloud techniques may give deeper insight in the complex human decision-making process particularly in morally difficult decision situations. This would offer additional insights about the motives, thoughts, and concepts people have when they have to solve tasks about moral principles and can provide the reasons for their decisions. By applying a qualitative content analysis of the different causal structure of dilemmas may improve the detection and categorization of the objective, systematic, and formal features of the dilemma's content. These categories in turn can be validated by an assignment of dilemmas as a possible task in a further experiment. Having a formal theory at hand allows to systematically analyze the implications of the objectives, concepts, and features relevant for moral decision making. Our formalism is able to distinguish between moral dilemmas and—at least for the reported cases—predict a perceived subjective difference between human raters.

References

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12–17.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision*, *3*(2), 121.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in cognitive sciences*, *6*(12), 517–523.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Halpern, J. Y. (2016). *Actual causality*. Cambridge, MA: The MIT Press.

Kohlberg, L. (1984). *Essays on moral development: Vol. 2. the psychology of moral development: Moral stages, their nature and validity*. Harper & Row.

Lindner, F. (2015). *Soziale Roboter und soziale Räume: Eine Affordanz-basierte Konzeption zum Rücksichtsvollen Handeln*. Doctoral dissertation, Department of Computer Science, University of Hamburg, Hamburg.

Lindner, F., & Bentzen, M. M. (2017). The hybrid ethical reasoning agent IMMANUEL. In B. Mutlu, M. Tscheligi, A. Weiss, & J. E. Young (Eds.), *Proceedings of the 2017 conference on human-robot interaction (HRI2017)*. ACM/IEEE.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 117–124).

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in cognitive sciences*, *11*(4), 143–152.

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in cognitive sciences*, *11*(8), 319–321.

Shou, Y., & Song, F. (2015). Moral reasoning as probability reasoning. In D. C. Noelle & P. P. Maglio (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2176–2181). Austin, TX.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*(1), 28–43.