A Model for Regulating of Ethical Preferences in Machine Ethics

Zohreh Baniasadi, Xavier Parent and Charles Max, Marcos Cramer

Interdisciplinary Center of Security and Trust , University of Luxembourg,
zohreh.baniasadi@uni.lu, xavier.parent@uni.lu, chalres.max@uni.lu, marcos.cramer@uni.lu

Abstract. Relying upon machine intelligence with reductions in the supervision of human beings, requires us to be able to count on a certain level of ethical behavior from it. Formalizing ethical theories is one of the plausible ways to add ethical dimensions to machines. Rule-based and consequence-based ethical theories are proper candidates for Machine Ethics. It is debatable that methodologies for each ethical theory separately might result in an action that is not always justifiable by human values. This inspires us to combine the reasoning procedure of two ethical theories, deontology and utilitarianism, in a utilitarian-based deontic logic which is an extension of STIT (Seeing To It That) logic. We keep the knowledge domain regarding the methodology in a knowledge base system called IDP. IDP supports inferences to examine and evaluate the process of ethical decision making in our formalization. To validate our proposed methodology we perform a Case Study for some real scenarios in the domain of robotics and automatous agents.

1 Introduction

Recently, autonomous agents such as robots, softbots, automated vehicles and drones have received a considerable amount of attention in the AI community, particularly in making ethical decisions [2,23,6]. There is a wide-spread skepticism when autonomous agents would eventually be in the charge of making ethical decision regarding the life of human beings. Theoretically, Machine Ethics is concerned with giving machines ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter [2].

The main contribution of AI in Machine Ethics for artificial moral agents (AMAs) is either through logical-based or machine learning approaches. Machine learning approaches have the potential to learn human ethics through the observation of human moral behavior. These methods are known as “bottom-up” approaches, which can include genetic algorithms, learning and deep learning algorithms [1,12]. Logical-based approaches have the potential to evaluate the ethical components of actions by formalizing ethical theories. These methods are called “top-down” approaches.

Among the ethical theories, deontology (rule-based) and utilitarianism (consequence-based), have the potential to be mechanized and formalized [6,28]. They support particular situations that come from the accepting or violating rules. In deontology the actions are deemed good or bad in themselves while in utilitarianism, results of the actions matter the most. It is open to discussion whether methodologies for each ethical theory separately might result in an action that is not always justifiable by human values. This inspires us to regulate preferences of an autonomous artificial agent by synthesizing the reasoning procedures of these two ethical theories.

The main research questions to be addressed in this context are:

1- *How to synthesize deontology and utilitarianism reasoning in a logical-based approach of Machine Ethics?*
2- *How to make a knowledge base specification for a logical formalization of Machine Ethics?*

It is worth to mention that, this present work deals with the problem of technically combining the procedure(s) of deontology and utilitarianism reasoning in a logic-based frame work. Being computer scientists and not philosophers, our object is not to introduce a new ethical theory. But, we would like to examine the out come of synthesizing two ethical reasoning procedures by one well developed approach. The ethicists and philosophers however provide reasoning and arguments to justify the autonomous agents’ ethical decisions.

Our methodological approach uses Utilitarian-based Deontic Logic (UDL) [16] as a logical agent-based framework for the ultimate need of formal proofs by combining deontological and utilitarian reasoning. And,

we use IDP as a knowledge representation (KR) language to simulate the ethical reasoning process in UDL as a knowledge domain in a knowledge base system.

STIT captures the concepts of agency by the fact that an agent acts when seeing to it that a certain state of affairs holds. Our intuition conducted by the two variants of STIT , utilitarian (UDL) model by Horty [16] and rational (DR-STIT) model by Lorini et. al. [21] . The main operators in these models i.e. deontic and rational operators are defined based on preference order between the states of affairs. they originate from social utility and legal obligatory actions, respectively. In this paper, we name the logics that is synthesizing these two methods of reasoning, E-STIT to regulate agent's ethical preferences.

IDP[1] is an integrated knowledge base system of logic programming and classical logic introduced by Marc Denecker et. al. [9]. This KB system is multi-inferential. It allows to store information in a knowledge base and provide a wide range of inference tasks. In this paper IDP [22] is the KB system for knowledge representation which integrates declarative specification in FO(.) with imperative management of the specification via Lua language. This makes IDP a powerful tool for solving a large variety of problems. We close this paper with a *Case Study* to experimentally validate our methodology of an ethical decision making procedure by a machine via IDP inferences tasks. For us, a Case Study constitutes a research strategy, i.e., a practical examination of a problem from real life [29].

2 Background

Machine Ethics, Roboethics and Computer Ethics are three distinct topics in the field of ethics for artificial intelligent agents which should not be confounded with each other. Machine Ethics mostly targets the moral behavior of artificial entities by contrast to human beings or other artificial intelligent beings [2]. Roboethics deals with the moral behavior of the humans who design, construct, use and treat these entities [26]. Computer Ethics focuses on professional behavior towards computer and information [18].

Ethical theories such as, Consequentialism, Deontology, Virtue, Contractarianism, etcetera, propose different competing views to construct the values when evaluating the morality of an action. For instance, utilitarian ethics, as a form of Consequentialism and deontological ethics, puts its focus on the character of the action or the rule to justify the morality of an action while virtue ethics concentrates on the character of those who are acting. Therefore an action, such as lying, which is rejected by deontological ethical theories, might be supported by utilitarian ethical theories while, in virtue ethics it depends on the person who performs the action.

In Machine Ethics, one of the main challenges for computer scientist is to formalize the moral behavior of machines that behave autonomously with respect to ethical theories. In other words, to build ethics for machines.

The logic of STIT (Seeing To It That) is a logic of agency which has its origins in philosophy and was introduced by Belnap et al. [5]. One can find different semantics and extensions of STIT logic in the literature [5,4,7,17,21,20]. We are interested in Kripke semantics of STIT which has the benefit of being more close to standard deontic logic.

Originally, STIT by Belnap et al. [5] is defined in terms of Branching Time (BT) augmented with Agent Choices (AC), BT+AC structure. In BT+AC semantics, the evaluation is based on the pairs of moment-history (m/h or $index$) in a tree-like structure. Where moments (m) are partially ordered (transitive and anti-symmetric)[2] and, a history (h) is a maximal linear order set of the moments. When $m \in h$ we say that moment m is on history h. But, in Kripke-style semantics, the concept of world (w) is taken to be the primitive element of evaluation. Despite of the explicit difference between BT+AC and Kripke semantics, there is an implicit one-to-one corresponding relation between the two evaluation concepts i.e. world and index. Due to the space limitation we only review the syntax and semantics of TSTIT (Temporal STIT), DRSTIT and UDL in the Kripke structure [5,4,15]. We leave it to the reader to read about the relation between the world and index if interested [5,4] .

[1] A project in the University of Lueven

[2] for any $m_1, m_2, m_3 \in M$, if m_1<m_2 and m_2 <m_3, then either $m_1 = m_2$ or m_1<m_2 or m_2<m_1

Here we only explain the relation of world and index in form of a simple example illustrated in figure 1. In STIT, for every moment (m) in Kripke semantics, one can identify the set of histories ($Hist_m$)[3] passing through it. There exists a unique world w at the intersection between m and h, on the other way around, for every world w there exists a unique corresponding h passing through m which includes w. Clearly, each moment consist of a set of worlds that are induced by an equivalence relation $R_\Box$ on W [24]. In figure 1 for instance, moment m_1 includes a set of worlds $\{w_1, w_2, w_3, w_4\}$, where there is a one-to-one corresponding relation between this set of world and the set of all histories $\{h_1, h_2, h_3, h_4\}$ passing through m_1.

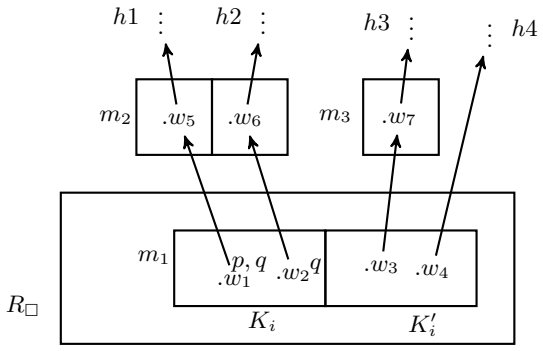

Fig. 1. Temporal Kripke semantics of STIT logic for a single agent i

T The available choices for agent i is characterized by the equivalent class over a set of worlds in each moment. For example in figure 1, the choices available for agent i (namely $Choice(i)$), are K_i and K_i' where $K_i = \{w_1, w_2\}$ and $K_i'\{w_3, w_4\}$. For the purpose of notational convenience, we define the arbitrary binary relation R as follows: Given a set of worlds W with $w \in W$ and an arbitrary binary relation R on W, $R(w) = \{v \in W \mid vRw\}$ is an equivalence relation. Inductively, for all $i \in Agent$, we define $R_i(w) = \{v \in W \mid vR_iw\}$ as an equivalence relation on W with respect to agent i and $R_\Box(w)$ a set of the worlds that are alternative to w ($R_\Box(w) = \{v \in W \mid vR_\Box w\}$). In the given example in figure 1, $R_i(w_1) = \{w_2\}$ (or $w_2R_iw_1$) and $R_\Box(w_1) = \{w_2, w_3, w_4\}$. We define $Choice(i)$ as a set of all choices available for agent i, a partition of the set of all worlds with respect to R_i. In figure 1, $Choice_i = \{K_i, K_i'\}$. p, q and q are atomic propositions that set to be true at w_1 and w_2, respectively.

If preference relations exist between the worlds, then a choice is preferred to the other choice for agent i if the set of worlds representing the first choice are better (preferable) to the set of world representing the other choice. This might restrict the set of available choices so that agent's preferable choices are subset of agents current choices. In order to capture the concept of two ethical theories, i.e. utilitarianism and deontology, we follow two variants of STIT: Utilitarian-base Deontic Logic (UDL) introduced by Horty [16] and STIT-with Deterministic time and Rational choices (DR-STIT) introduced by Lorini and Satror [21]. The ought($\bigcirc$) operator in Horty's logic represents the agent's best choice regarding to utilitarianism [16] while the rational $([ir])$ operator in the work of Lorini and Satror [21] indicates the agent's best choice regarding to deontology. Both operators are based on the preference relations between the worlds.

In the next section and as told already above, we review the syntax and semantics of TSTIT, DRSTIT and UDL in the Kripke structure [4,15].

2.1 Syntax of TSTIT

The language of the TSTIT is built from a finite set $Agent = \{1, \ldots, n\}$ of agents and a countable set $Atm = \{p, q, r, \ldots\}$ of propositional letters (atomic propositions). Let $p \in Atm, i \subseteq Agent$. The language of T-STIT, $\mathfrak{L}_{T-STIT}$ is given by the following Backus-Naur Form:

[3] $Hist_m = \{h \mid h \in Hist, m \in h\}$ where $Hist$ is the set of all histories [5].

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [i]\varphi \mid \Box\varphi \mid G\varphi \mid H\varphi$$

$[i]\varphi$, is read as "agent i sees to it that φ", and $\Box\varphi$, is read as "φ is settled to be true". Intuitively, the latter captures the concept of ability and the former captures the concept of historical necessity for agents. Finally, G and H are standard temporal operators where $G\varphi$ is read as "φ will always be true in the future" and $H\varphi$ is read as "φ has always been true in the past" . The dual operators are defined as usual; $\Diamond\varphi \equiv \neg\Box\neg\varphi$ and $\langle i\rangle \equiv \neg[i]\neg\varphi$ [20] .

Remark 1. The STIT semantics supports different concepts of agency. It captures by the fact that an agent i acts only if i sees to it that a state of affairs holds. Various operators of agency exist of which the most famous STIT operators are: Chellas and deliberative STIT operators; $[i\ cstit : \varphi]$ and $[i\ dstit : \varphi]$ respectively [17]. In this paper, $[i]$ is an abbreviation for the Chellas STIT operator and $[i\ dstit : \varphi]$ is inter-definable by $[i]$ ($[i\ dstit : \varphi] \equiv [i]\varphi \wedge \neg\Box\varphi$).[4] $[i]\varphi$ holds if and only if φ holds in all the set of worlds represented by the agent's choice.

2.2 Semantics of TSTIT

Definition 1 (Kripke model [20]). *A Kripke model for the TSTIT logic is a tuple $M = (W, R_i, R_\Box, R_G, V)$ where:*

- *W is a non-empty set of possible worlds.*
- *R_i and $R_\Box$ are equivalence relations on W such that:*
  - *C1.* $\forall i \in Agent: \quad R_i \subseteq R_\Box$.
  - *C2.* $\forall w_1, \ldots, w_n \in W$, *if* $w_i R_\Box w_j$ *then,* $\forall\{1, \ldots, n\} \in Agent$ *it holds that,* $\cap_{1\leq i\leq n} R_i(w_i) \neq \emptyset$.
- *R_G is a serial and transitive binary relation on W such that:*
  - *C3.* $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_G(w_1)$ *and* $w_3 \in R_G(w_1)$ *then* $w_2 \in R_G(w_3)$ *or* $w_3 \in R_G(w_2)$ *or* $w_2 = w_3$.
  - *C4.* *If* $R_H = R_G^{-1} = \{(x, y) \mid (y, x) \in R_G)\}$ *then* $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_H(w_1)$ *and* $w_3 \in R_H(w_1)$ *then* $w_2 \in R_H(w_3)$ *or* $w_3 \in R_H(w_2)$ *or* $w_2 = w_3$.
  - *C5.* $\forall w_1, \ldots, w_i \in W$, *if* $w_i \in R_\Box(w_j)$ *then,* $w_i \notin R_G(w_j)$.
- *V is an evaluation function for atomic formulas, $V : Atm \mapsto 2^W$.*

As it is explained above, $R_\Box(w)$ is the set of worlds that are alternative to w and $R_i(w)$ is the set of all alternatives that are "enforced" by agent i's actual choice at W. They are restricted by two constraints $C1$ and $C2$. $C1$ says that an agent can only choose among possible alternatives. $C2$ expresses the assumption of the independence of choices among agents. Following [24], a moment is defined as the equivalence classes induced by the equivalence relation $R_\Box$ on W. $R_G(w)$ is the set of worlds in the future of w. Constraints $C3, C4$ and $C5$ imply that R_G is serial, transitive and irreflexive.

Definition 2 (Satisfaction conditions, [20]). *Let $M = (W, R_i, R_\Box, R_G, V)$ be a Kripke TSTIT model, we have:*

$M, w \models p$ *iff $w \in$V(p);*
$M, w \models \neg\varphi$ *iff it is not the case that* $M, w \models \varphi$;
$M, w \models \varphi \wedge \psi$ *iff* $M, w \models \varphi$ *and* $M, w \models \psi$;
$M, w \models [i]\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_i w$;
$M, w \models \Box\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_\Box w$;
$M, w \models G\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_G w$;
$M, w \models H\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_G^{-1} w$;

The notations of *validity* and *satisfiability* are defined as usual.

[4] The deliberative STIT logic is more appropriate to capture the original concept of action but since they are inter-definable, one can choose any as primitive operator [17].

2.3 Syntax and semantics of DR-STIT

DR-STIT is the variant of STIT with discrete time and rational choices interpreted in a Kripke semantics. The DR-STIT language, $\mathfrak{L}_{DR-STIT}$ is the set of formulas given by the BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid X\varphi \mid Y\varphi \mid [i]\varphi \mid [ir]\varphi$$

The other Boolean connectives $\top, \bot, \vee, \rightarrow$ and $\leftrightarrow$ are defined in the standard way.

Intuitively, $\Box\varphi$ and $[i]\varphi$ are read as before. The formulas $X\varphi$ and $Y\varphi$ are respectively the tense operators similar to $G\varphi$ and $H\varphi$ in TSTIT where X is an operator for the facts in the next moment and $Y$is an operator for the facts in the previous moment. $[ir]\varphi$ is read as "agent i rationally sees to it that φ" and captures the fact that φ is the rational choice of the agent i .

Definition 3 (Semantics for DR-STIT [21]). *A Kripke model for DR-STIT is a tuple* $M = (W, C_i, RC_i, \equiv, \rightarrowtriangle, \leftarrowtriangle, V)$ *where:*

- W *is a non-empty set of possible worlds.*
- C_i *and* $\equiv$ *are equivalence relations on* W *such that:*
  C1. $\forall i \in Agent: \quad C_i \subseteq \equiv$.
  C2. $\forall w_1, \ldots, w_i \in W$, *if* $w_i \equiv w_j$ *then,* $\forall i, j \in \{1, \ldots, n\} \in Agent$ *it holds that,* $\cap_{1 \leq i \leq n} C_i(w_i) \neq \emptyset$.
- $\rightarrowtriangle$ *is a serial and deterministic relation on* W *such that:*
  C3. $\forall w_1, w_2 \in W$ *if* $w_1 F w_2$ *then* $w_1 \not\equiv w_2$ *where* F *denoting the transitive closure of the binary relation* $\rightarrowtriangle$ *and supposed to be deterministic time.*
- $\leftarrowtriangle$ *is the inverse relation of* $\rightarrowtriangle$*, and is supposed to be deterministic.*
- *Every* RC_i *is a subset of the partition of* W *induced by the equivalence relation* C_i.
  C4. $\forall w_1 \in W$ *and* $\forall i \in Agent: \exists w_2 \in W$ *such that* $w_1 \equiv w_2$ *and* $C_i(w_2) \in RC_i$.
- V *is an evaluation function for atomic formulas,* $V : Atm \mapsto 2^W$.

The satisfaction conditions in a DR-STIT model for atomic formulas, negation and conjunction is similar to satisfaction conditions in a Kripke model for TSTIT. Also C_i , $\equiv$, $\rightarrowtriangle$ and $\leftarrowtriangle$ in DR-STIT model are respectively equivalent to R_i, $R_\Box$, R_G and R_G^{-1} in a TSTIT model.

Definition 4 (Satisfaction conditions for the rational operator in DR-STIT [21]). *Let* $M = (W, C_i, RC_i, \equiv, \rightarrowtriangle, \leftarrowtriangle, V)$ *be a Kripke model for DR-STIT, The new operator (*$[ir]\varphi$*) is evaluated as follows:*
$M, w \models [ir]\varphi$ *iff* *IF* $C_i(w) \in RC_i$ *THEN* $\forall w' C_i w : M, w' \models \varphi$

In figure 1, we consider $w_1 C_i w_2$. In this case, $w_2 \models [ir]q$.

2.4 Syntax and semantics of UDL

We now describe the Utilitarian Deontic Logic (UDL) of Horty [16]. The language of UDL, $\mathfrak{L}_{udl}$ is defined by the following Backus-Naur Form:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [i]\varphi \mid \bigcirc_i \varphi$$

Intuitively, $\Box\varphi$ and $[i]\varphi$ are read as before. $\bigcirc_i\varphi$ is read as "agent i ought to see to it that φ".

The semantics of UDL is based on utilitarian models, where the utility for each world is a relation induced by the real numbers assigned to the corresponding worlds as social utilities (equivalent to the definition of payoff in game theoretical settings).

Definition 5 (Utilitarian Kripke model). *An utilitarian Kripke model is a tuple* $M = (W, R_\Box, R_i, \leq_u, V)$, *where* $W, R_\Box, R_i$ *and* V*are defined as in a Kripke TSTIT model, and* $\leq_u$*, representing the utility order which is a reflexive and transitive relation over* W.

$w \leq_u v$ *is read as* v *is at least as good as* w. $w \approx_u v$ *is short for* $w \leq_u v$ *and* $v \leq_u w$. $w <_u v$ *if and only if* $w \leq_u v$ *and* $v \not\leq_u w$. $w <_u v$ *is read as* v *is strictly better than* w.

For convenience, we can also use numbers to express the utility of each world. In figure 1, let $w_1 = 0$, $w_2 = 2$, $w_3 = 0$ and $w_4 = -1$ then, the one can define the utility order ($\leq_u$) between corresponding worlds as follows: $w_4 <_u w_1 \approx_u w_3 <_u w_2$.

Definition 6 (Individual agent choice [16]). *We define $Choice(i)$: $i \mapsto \{[w]_{R_i} : w \in W\}$ as the set of individual agent's choices for each $i \in Agent$. Let $[w]_{R_i}$ be the equivalence classes of W with respect to R_i.*

Preference over a set of worlds is defined by lifting preferences over worlds. There is no standard way of lifting preferences. Lang and van der Torre [19] summarize three ways of lifting; Strong Lifting, Optimistic Lifting and Pessimistic Lifting. Following Horty we define Strong Lifting according to Weak lifting.

Definition 7 (Preferences over sets of worlds via Weak Lifting [16]). *Let $X, Y \subseteq W$ be two sets of worlds. $X \preceq Y$ (Y is weakly preferred to X) iff $\forall x \in X$ and $\forall y \in Y \quad x \leq y$.*

Definition 8 (Preferences over sets of worlds via Strong Lifting [16]). *$X \prec Y$ (Y is strongly preferred to X) if and only if $X \preceq Y$ and $Y \npreceq X$ iff:*
(1) $\forall x \in X, \forall y \in Y$, $x \leq y$ *and*
(2) $\exists x' \in X, \exists y' \in Y$, $x' < y'$.

Definition 9 (Optimal [16]). *Let i be an agent,*

- $Optimal_i$ = *{$K \in Choice(i)$: there is no $K' \in Choice_i$ such that $K \prec K'$}.*

In the semantics of UDL, the optimal choices is used to interpret the deontic operators.

Definition 10 (Satisfaction condition of UDL for ought operator [16]).
Let $M = (W, R_\Box, R_i, \leq_u, V)$ be a utilitarian Kripke model and $w \in W$. The truth conditions of atomic formulas, negation, conjunction and $[i]$ operators in M is similar to a Kripke TSTIT model. The satisfaction condition for obligation is as follows:
$M, w \models \bigcirc_i \varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in K$ *such that* $K \in Optimal_i$*;*

Considering figure 1 as a Kripke model M we have, $M, w_1 \models \bigcirc_i q$ because, one of the choice of the agent i is K_i such that $K_i \in optimal_i$ and $K_i = \{w_1, w_2\}$.
The notations of Validity and Satisfiability are defined as usual.

2.5 FO(·) knowledge base language and IDP system

A knowledge base system aims to express the domain knowledge in an expressive way to solve various problems in domain using inference tasks and rules. The logic for the knowledge base system that is used in this paper is FO(·) [8] and the system build upon this logic is IDP [27].

The language FO(·) refers to the class of extensions of first order logic (FO) based on a logical framework. Currently, the language of the IDP system is FO(T, ID, Agg, arit, PF) [8], i.e. FO which is extended with types, definitions, aggregates, arithmetic and partial functions. In this paper we use the subset language FO(T, ID, Agg ,PF). We use FO(·) as an abbreviation for this language. Below, we introduce the aspects of the logic and its syntax on which our formalization of E-STIT relies.

In IDP variables x, y (represented by lower case English letters), atoms A (upper case English letters), FO-formulas φ (Greek letters) are defined as usual. A vocabulary Σ consists of a set of symbols, predicate symbols and function symbols. Symbols are types (sorts). Each predicate (function) symbol has an associated arity, i.e. the number of arguments.[5] A function can be declared as partial, indicating that for some inputs, the output can be undefined; otherwise it is total. ID stands for the following definition structure for inductive definitions: An inductive definition Δ is a set of rules of the form $\{\forall \overline{x} : P(\overline{x}) \leftarrow \varphi(\overline{y})\}$, where $\overline{x}$ is a tuple of variables and $\varphi(\overline{y})$ is a first-order logic (FO) formula. Agg are aggregate terms of the form $Agg(E)$ with aggregate function for cardinality, sum, product, maximum and minimum which accept E as an expression of the form $\{(\overline{x}, F(\overline{x}) \mid \varphi(\overline{x})\}$. For example $sum\{(x, F(x) \mid \varphi(x)\}$ is $\Sigma_{x \in \varphi(x)} F(x)$. A FO(·) theory is a set of symbols, inductive definitions, aggregate functions (a set of FO(·) formulas) and FO formulas.

[5] We often use P/n (f/n) to denote the predicate symbol P (respectively function symbol f) with arity n.

Definition 11. *A* partial set *on the domain D is a function from D to $\{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$, where* **t***,* **f** *and* **u** *stand for the three truth-values* true*,* false *and* undefined*. A partial set is two-valued (or total) when* **u** *does not belong to its range.*

A (partial) structure S is a tuple that consists of a domain D_τ for all types τ in the vocabulary Σ and an assignment of a partial set $\mathcal{I}$ to each symbol in Σ, called the interpretation of type symboles in S. For a predicate symbol P of arity n, the interpretation $P^{\mathcal{I}}$ is a partial set on the domain D^n; for a function symbol f of arity n, $f^{\mathcal{I}}$ is a function from D^n to D. Where the interpretation of symbols in S is a two-valued set, we say the structure is total. In fact, we call a partial structure total if and only if $P^{\mathcal{I}}(\bar{d})$ is total for all $\tau \in \Sigma$, there is a $\bar{d} \in D^n$. We call a partial structure S finite if and only if its domain D is finite. The interpretation of terms $t^{\mathcal{I}}$ and the satisfaction relation $\models$ for total structures $S \models \varphi$ are defined as usual.

The precision-order on the truth values is given by $u <_p f$ and $u <_p t$. It can be extended pointwise to partial sets and partial structures, denoted $S \leq_p S_0$. Notice that total structures have the maximal precision . We say that S_0 extends S if $S \leq_p S_0$. A total structure S is called functionally consistent if for each function f of arity n, the interpretation $F^{\mathcal{I}}$ is the graph of a function $D^n \mapsto D$. A partial structure S is functionally consistent if it has a functionally consistent two-valued extension. Unless stated otherwise, we will assume for the rest of this paper that all (partial) structures are functionally consistent.

Inferences tasks In the KB system, a specification is a bag of information. This information can be used for solving various problems by applying a suitable form of inference on it. FO is standardly associated with deduction inference. It takes as input a pair of theory T and sentence φ as input and returns t if $T \models \varphi$, and f otherwise. This is well-known to be undecidable for FO, and by extension for FO($\cdot$). However, to achieve desired results in our meta modelling of a modal logic we can use simpler forms of inference. Indeed, in many domains alike a fixed finite domain is specified by translating the real world scenarios into the structural form that is required.

In logic a natural format to describe these finite domains is either by definition a total fixed structure or a partial structure with a finite domain. Also other data that are often available in such problems can be represented in that structure. As such various inference tasks are solvable by finite domain reasoning and become decidable. In our case, we define total structures for the time being and we leave partial structures and auto completion as a future work. However, reasoning on finite domain is decidable. Below, we introduce base forms of inference and recall their complexity when using finite domain reasoning. We assume a fixed vocabulary V and theory T.

Modelcheck(T, S)**:** A total structure S and theory T over the vocabulary interpreted by S; output is the boolean value $S \models T$. Complexity is in **P**.

Modelexpand(T, S)**:** Theory T and partial structure S; output: a model I of T such that $S \leq_p I$ or $UNSAT$ if there is no such I. Complexity of deciding the existence of a modelexpansion is in **NP**.

Query(S, E)**:** A (partial) structure S and a set expression $E = \{x \mid \varphi(x)\}$; output: the set $AQ = \{x \mid \varphi(x)^S = t\}$. Complexity of deciding that a set A is AQ is in **P**.

3 Machine Ethics in STIT Logic

In this paper we call the STIT logic to formalize Machine Ethics E-STIT.

3.1 E-STIT logic

The logic of Ethics in STIT (E-STIT) is a language for ethical autonomous agents. We use E-STIT to capture the fact about the main question of ethical theories. The main question of ethical theories is more prominently about the agents choices when they know the ethical status of the worlds and not only the moral status of the worlds. In fact, E-STIT provide ethical metrics for agents to evaluate the ethical status of their Choices. It is

simply a variant of an extension of TSTIT logic. Generally, in any variant of STIT logic, agents make choices and each choice is represented by a set of possible worlds, a partition of equivalent classes in each moment over all sets of available worlds (W). Therefore the interpretation of deontic modality for ethical choices is based on best choices, which can only be defined on top of preference over sets of worlds. Preference over sets of worlds is defined by lifting preferences over worlds. As we discussed before in Section 2.4 following Horty [16] we adopt Strong Lifting defined in Definition 8. Therefore, between two choices K and K', K is better than K' from the ethical point of view if and only if all the worlds in K' are at least as good as all the worlds in K and there exists at least a world in K' that is strictly better than all the worlds in K.

3.2 Syntax of E-STIT

The language of E-STIT, $\mathfrak{L}_{E-STIT}$ is defined by the BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [i]\varphi \mid \bigcirc_i\varphi \mid X\varphi \mid Y\varphi$$

Where $\bigcirc_i\varphi$ is read as "agent i ought to see to it that φ" enforces the concept of obligation from the ethical point of view. The dual operator of $\bigcirc_i$ is $P_i\varphi \equiv \neg \bigcirc_i \neg\varphi$.

3.3 Semantics of E-STIT

Definition 12 (The Kripke model of E-STIT). *A Kripke model for the E-STIT logic is a tuple* $M = (W, R_i, R_\Box, R_X, \leq_r, \leq_u, V)$ *where;*

- W *is a nonempty set of possible worlds.*
- R_i *and* $R_\Box$ *are equivalent relations on* W *such that:*
  - *C1.* $\forall i \in Agent: \quad R_i \subseteq R_\Box$.
  - *C2.* $\forall w_1, \ldots, w_i \in W$, *if* $w_i \in R_\Box(w_j)$ *then,* $\forall i \in Agent$ *such that* $i \in \{1, \ldots, n\}$ *it is hold that,* $\cap_{1\leq i\leq n} R_i(w_i) \neq \emptyset$.
- R_X *is a serial and transitive binary relation on* W *such that:*
  - *C3.* $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_X(w_1)$ *and* $w_3 \in R_X(w_1)$*then* $w_2 \in R_X(w_3)$ *or* $w_3 \in R_X(w_2)$ *or* $w_2 = w_3$.
  - *C4. If* $R_Y = R_X^{-1} = \{(x, y) \mid (y, x) \in R_X)\}$ *then* $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_Y(w_1)$ *and* $w_3 \in R_Y(w_1)$ *then* $w_2 \in R_y(w_3)$ *or* $w_3 \in R_Y(w_2)$ *or* $w_2 = w_3$.
  - *C5.* $\forall w_1, \ldots, w_i \in W$, *if* $w_i \in R_\Box(w_j)$ *then,* $w_i \notin R_X(w_j)$.
- $\leq_u$ *and* $\leq_r$ *are transitive and reflexive relations over* $R_\Box(w)$.
- V *is an evaluation function for atomic formulas,* $V : Atm \mapsto 2^W$.

We define the ethical relation between the worlds in terms of combining the utility and rationality preference relations. To this end we follow multi-preference-based semantics by Goble [11] for Standard Deontic Logic (SDL) and Dyadic Standard Deontic Logic (DSDL) [10]. It allows SDL and DSDL to accommodate deontic dilemmas without inconsistency by combining multi-preference relations, and to do so in a way that follows conventional methods for interpreting deontic statements. He simply consider the union of all preferences to interpret a deontic operator [11].

Definition 13 (Ethical preferences in E-STIT model). *Given a Kripke structure of E-STIT* $M = (W, R_i, R_\Box, R_X, \leq_r, \leq_u)$ *where* $\leq_r$ *and* $\leq_u$ *are individual agent's utility and rational preferences over worlds in* $R_\Box$ *then, the ethical preference of* i *over* $w, w' \in R_\Box$ *is defined with "*$\leq_e$*", such that* $w \leq_e w'$ *read as "*w' *is ethically as good as* w*" if and only if :*

$$w \leq_r w' \text{ or } w \leq_u w'.$$

$w \approx_e w'$ *is short for* $w \leq_e w'$ *and* $w' \leq_e w$.$w <_e w'$ *if and only if* $w \leq_e w'$ *and* $w' \not\leq_e w$. $w <_e w'$ *is read as "*w' *is strictly better than* w *from the ethical point of view".*

Comparing two rational and utility orderings, the two main resulting cases are as follows:

- CASE 1: The utility and rational orderings agree with each other on the ranking of worlds.
- CASE 2: The utility and rational orderings disagree with each other on the ranking of worlds.

In CASE 1 the result of merging two orderings is obvious. The problematic case is CASE 2. Suppose the specific case of a conflict between the two orderings with maybe one of them being more justifiable by human values. For instance, if there is conflict among the advices from the teacher (or the school) and the mother, surely one might prefer one to the other. Imagine you have exam and you are sick, too. The school rules obliges you to do the test, but your mother obliges you to stay at home. The orderings enforces by theories, i.e. deontology and utilitarianism, are both equally conductive but in the case of conflict one would follow the theory that is most believed. We propose in our methodology to consider one of the two orderings as the main ordering. We take the advantages from the other theory in the case of conflict. Thus we propose a solution to the problematic cases while we are still abiding to our definition.

Intuitively, a world w' is ethically preferred to the world w if and only if, w' has a higher ethical order than w coded by rationality/utility. If the rational/utility status between the world is undistinguishable then w' should be compared to w regarding the other code of ethics.

In fact, following our intuition, an automated entity would consider the utility order as primitive if it is in favor of deontology. When it prefers utilitarianism, it would consider the rational order. Perhaps there are some specific cases where ethical justifications are generated by deontology reasoning while in some other cases it might be utilitarianism reasoning that supports ethical values. We will implement these two types of incorporation of deontology and utilitarianism reasoning procedures in IDP and we advise the applicant to take one as primitive due to his needs. Before giving the official definition of truth condition for E-STIT, in example 1 we clarify the idea behind achieving the ethical order with respect to rational and utility orders.

Example 1 (Robots in ICU [6]). Figure 2 illustrates a simple example for the choices of two autonomous agents in a Kripke semantics of STIT framework. The example is taken from [6] where, two robots i and j are designed to work overnight in ICU. Suppose that, there are two humans; H_1 under the care of i and H_2 under the care of j, both are recovering in ICU. H_1 requires the life supporting equipments, but is expected to be gradually weaned from it, gradually. H_2 suffers from extreme pain and requires very expensive medication to control his pain. Suppose, i can terminate H_1's life providing enough organs to save n other humans (according to the information received from the hospital data base) and j is able to delay the delivering of pain medication to H_2 with the utilitarian goal of economically strapped resources in the hospital. If p and q are two atomic propositions indicating "terminating H_1's life" and "delaying the delivering of pain medication", respectively, then one can illustrate the scenario as the one in figure 2.

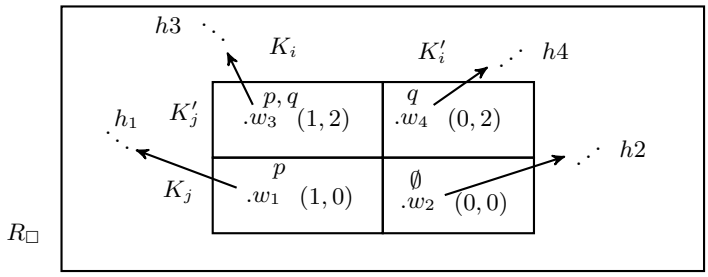

Fig. 2. Temporal Kripke semantics of STIT logic for two agents in a single moment.

In figure 2, the choices available for robots i and j are $Choice(i) = \{K_i = \{w_1, w_3\}, K_i' = \{w_2, w_4\}\}$ which terminate or do not terminate H_1's life (p or $\neg p$), and $Choice(j) = \{K_j = \{w_1, w_2\}, K_j' = \{w_3, w_4\}\}$ which delay or do not delay the delivering of pain medicine to H_2 (q or $\neg q$). According to the discussion

above, the two ethical theories, i. e. deontology and utilitarianism, allow to distinguish the moral order between the worlds. The latter comes from the numbers indicated in figure 2 while the former is obtained by existing deontological codes.

According to deontological theories, there are some actions one just shouldn't perform [25] such as killing and causing harm in example 1. Therefore, the worlds with less harm have a higher rationality. $w_3 <_r w_1 \approx_r w_4 <_r w_2$ is the rational order one can obtain from example 1 between the worlds. According to utilitarianism, however, an action cannot really be immoral if it generates only happiness for a majority [25]. Hence, the action that provides more happiness has higher utility. So, killing a person who is about to die to save 5 other people justifies itself to be moral. Thus, the world w_3 has the highest utility order.

In figure 2, the numbers representing the utility value of each world are individual agent-independent utility values. Following John Harsanyi [14], we would like to obtain the individual agent-dependent utility value (said to be group utilities) for each specific world.

In fact, the relationship between individual utilities and group utilities is a subject open to pitfalls. We adopt John Harsanyi's proposal and conceive group utility as the arithmetical mean of the individual utilities of the agents involved [14].

Definition 14 (Group Utility $GU(w)$). *Let i be an agent, w a world and u_i be an individual utility for the agent i and $|Agent|$ be the cardinality of the all agents. then, the Group Utility of the world$w$($GU(w)$) is defined as follow:*

$$GU(w) = \frac{1}{|Agent|} \sum_{i \in Agent} u_i$$

In figure 2, the individual utility of agents is indicated as a pair of real numbers for each world, (u_i, u_j), where the first number u_i represents the *individual utility* for agent i and the second number u_j represents the *individual utility* for agent j. For the world w_3 in figure 2 (where p and q terminate H_1's life and delay the delivering pain medicine holds), for instance, number $u_i = 1$ is given as the utility of performing p for agent i and number $u_j = 2$ as the utility of performing q for agent j. One can easily obtain the *group utility* of each world. $GU(w_1) = 0.5, GU(w_2) = 0, GU(w_3) = 1.5, GU(w_4) = 1$ (Obviously, one can allocate the utility values to the corresponding histories). Therefore, in example 1 contrary to the rational order, w_3 has the highest utility order and w_2 has the lowest. In short, we have the following utilitarian and deontological orders:

- Utilitarian order: $w_2 <_u w_1 <_u w_4 <_u w_3$.
- Deontological order: $w_3 <_r w_1 \approx_r w_4 <_r w_2$

The utilitarian order suggests the action p for robot i and the action q for robot j whereas the deontological order proposes the actions $\neg p$ and $\neg q$. Now, several questions arise as for instance *a) What is the best choice an ethical autonomous agent can take? b) How can we represent and formalize the ethical order between the worlds? c) Does any defined ethical order of an agent's preference limit it's auonomy?* In order to respond to these questions in E-STIT, the ethical order between the worlds can be founded by justifying the human worth. It is ethically desired that robot i refrains from ending H_1 life and robot j also delivers appropriate pain relief to H_2 (indicated by w_2 in figure 2). The next preferred ethical action is when robot i sustains life support, but robot j withholds the medicines to save money (w_4). So when i performs p and j performs q the ethical results is unsatisfactory because one of the moral rules is ignored (represented as w_1). The worst possible outcome is when i kills H_1 and j withholds (in w_3) [6]. The ethical order obtained from the justification above follows the deontological theories and therefore one can conclude the following order via E-STIT:

- Ethical order: $w_3 <_e w_1 <_e w_4 <_e w_2$

Once we have the ethical order between the worlds, we apply Definition 8 to specify the ethical choice for the agents. In example 1 the choices for agents i and j which contain w_2 are ethically preferred to the other choices. The ethical preferences relation between the choices for agent i and j are respectively as follows:

$$K_i \prec K_i' \quad \text{and} \quad K_j' \prec K_j$$

Although, the ethical order of the choices is acceptable from the human's ethical preferences in a similar real situation, the utility and rational orders is completely sensitive to the source and the way they are defined. In this paper, we are not verifying the wrongness or rightness of the utility and rational orders. This is another vast topic to discuss. In this paper we are just about to regulate the agent's preferences when the rationality and utility orders of the worlds are proved in advance. Therefore, a robot who follows utilitarianism would show a completely different behavior based on its ethical preferences. The actions of $\neg p$ and $\neg q$ appear to be ethical according to the greatest happiness that they are going to generate.

Following Horty [16], the conceptual analysis of *"ought"* operator, representing what an agent ethically obliged to choose, is based on the optimal ethical choice of the agent that is defined with Definition 9.

Definition 15 (The Kripke-style semantics for E-STIT [20,21]). *Let* $M = (W, R_i, R_\Box, R_X, \leq_r, \leq_u, V)$ *be a Kripke model of E-STIT logic then;*

$M, w \models p$ *iff* $w \in$*V(p);*
$M, w \models \neg\varphi$ *iff it is not the case that* $M, w \models \varphi$*;*
$M, w \models \varphi \wedge \psi$ *iff* $M, w \models \varphi$ *and* $M, w \models \psi$*;*
$M, w \models \Box\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_\Box w$*;*
$M, w \models [i]\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_i w$*;*
$M, w \models \bigcirc\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in K$ *such that* $K \in Optimal_i$*;*
$M, w \models X\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_X w$*;*
$M, w \models Y\varphi$ *iff* $M, w' \models \varphi$ *for all* $w' \in W$ *such that* $w' R_X^{-1} w$*;*

The *validity* and *satisfiability* for the Kripke semantics are define as usual.

4 Knowledge-Base Specification and IDP for E-STIT

Generally, a knowledge base system consists of two main sections: the section which specifies the domain knowledge and the section with Lua prescription which provides the programming options and inferences tasks. The specification of domain knowledge consists of a vocabulary, a theory and a structure. In the case of E-STIT, the specification regarding vocabulary and theory of a knowledge base system contains the information about the number of agents and worlds, the morality and the utility of each world and also the alternative choices for each agent. The inferences tasks are either posing a query to ask each agent about the ethical status of each world, or they expand all possible models out of the theory T specifications in the knowledge base and the specific structure S for that unique situation.

4.1 Vocabulary specification

Type symbols, predicate symbols and (/partial) function symbols are the main elements of vocabulary in IDP. We specify the terms $world$, $agent$, $moment$, $prop$ as the type symbols for the concepts of world, agent, moment and atomic propositions in E-STIT, $uval$ and $rval$ as types of numbers to characterize the utility and rationality orders between the worlds. $(R-agent(agent, moment, world, world))$ is a predicate symbol for equivalence relation on a set of worlds with respect to each agent and the predicate symbol $(Val(wold, prop))$ is for the truth values of each world. Ethical orders between the worlds $(Ethical - order(world, world))$ are expressed as predicate symbols. Formulas and utility (/rationality) relations between the worlds are given as partial functions. The terms $neg(prop)$, $and(prop, prop)$, $stit(agent, prop)$, $estit(agent, prop)$ and $next(moment, prop)$ respectively stand for $\neg\varphi$, $\varphi \wedge \varphi$, $[i]\varphi$, $\bigcirc_i\varphi$ and $X\varphi$. In IDP we can not define an infinitely set of formulas. We restrict ourselves to the finite set of formulas using partial functions. This point is realizable in the structure specifications that requires to clearly specify formulas and relations between them. Finally, some auxiliary symbols such as$R - agentT(agent, moment, world, world)$ are given.

Listing 1.1. Vocabulary specification for E-STIT

```
vocabulary E–STIT–V{
    /*
     * Types contain the complete domain of the problem.
     *       In this case being the propositions,worlds, agents and moments */
    type moment isa nat
    type world
    type agent
    type prop
    type uval isa int
    type rval isa nat

     partial neg(prop): prop
     partial and(prop,prop):prop
     partial next(moment,prop):prop
     partial stit(agent,prop):prop
     partial estit(agent,prop):prop


     //agents poissible choices
  R_agent(agent, moment,world, world)
     // R_agent is the equivelant classes(partitions)on all set of worlds
  R_Tagent(agent, moment, world, world)

      //R_next(m1,w1,m2,w2) says that w2 which is in m2 is in the next moment of
          w1 which in m1.
  R_next(moment,world,moment,world)

     //lifting preferences relation over set of world(better choice)
   Better_Choice(agent,moment,world,world)

     //optimal choice
   Optimal_Choice(agent,moment,world)

   Val(world,prop)
   eval(world,prop)

     partial Rationality(world):rval
     partial Utility_ind(agent,world):uval
     partial Utility(world):uval
    //Ethical_order(wi,wj) expresses that wi has higher ethical order than wj.
     Ethical_order(world,world)
     }
```

Listing 1.1 represents the vocabulary specification in IDP for E-STIT logic. Now we are ready to define the semantics of the logic and the relations between type, predicate and function symbols in theory via inductive definitions and aggregation functions.

4.2 Theory specification

Theory specifies the semantics of a Kripke model for E-STIT. It is composed of the usual evaluation for common connective operators (e.g. $\neg$, $\wedge$, $\vee$ and $\rightarrow$), the evaluation for the modal operator (i.e.$[i]$ and $\bigcirc_i$) and the agent's choices according to the ethical status of the worlds. Here we only define the evaluation function, but we ask the reader to click on the link in figure 3 to see the theory specification.

http://dtai.cs.kuleuven.be/krr/idp-ide/?src=473bdaf7a004b5342f38b3fe9e0fdb67

Fig. 3. IDP specification for ICU example

We define the evaluation functions recursively via inductive definitions. These definitions are the main part of a theory specification for a modal logic. The predicate *eval* which stands for evaluation expresses the fact that a formula is evaluated as being true in a world. Listing 1.2 displays the inductive definitions in the theory T-E-STIT that we define as ruth table for E-STIT semantics. We comment the evaluation function for two connective operators; $\vee$ and $\rightarrow$ in IDP, due to the fact that, they are inter-definable via the other connective operators; $\neg$ and $\wedge$.

Listing 1.2. Theory specification for E-STIT truth table

```
theory T-E-STIT:V-E-STIT{
        // EVAL Definition
    eval(w,P) <- Val(w,P).
    eval(w,neg(P)) <- ~eval(w,P).
    eval(w,and(P,Q)) <- eval(w,P) & eval(w,Q).
    //eval(OR(P,Q)) <- eval(P) | eval(Q).
   //eval(IMPL(P,Q)) <- ~eval(P) | eval(Q).
   eval(w,next(m,P))<- ?m1,w1: m1~=m & w1~=w & R_next(m,w,m1,w1)& eval(w1,P) &
       m1>m.

        //stit evaluation functions
   eval(w,stit(a,P)) <-  !w1:(R_agent(a,m,w,w1) => eval(w1,P)).

         //ETHICAL evaluation function
   eval(w,estit(a,P))<- !w1: (Optimal_Choice(a,m,w1)=> eval(w1,P)).
        }
```

4.3 Creating a Structure

In order to solve a concrete problem, we specify an instance (a structure). A structure contains information about the actual value of the concepts in a vocabulary. At least the types should be fully interpreted in any structure. Listing 1.3 shows the structure for the example 1.

Listing 1.3. Structure specification of health care system example (example 1)

```
structure S:V-E-STIT{
   agent={i;j}
   world={w1;w2;w3;w4}
   moment={1}
   uval={0..5}
   rval={1..4}

    prop = {p; q; negp; negq; pandq; qandp; istitp; jstitq;istit_negp;iestitp;
        i_estit_negp;jstitq;jstit_negq;jestitq;j_estit_negq}


    Val={w1,p;w3,p;w3,q;w4,q}

    R_agentT={i,1,w1,w3;i,1,w2,w4; j,1,w1,w2; j,1,w3,w4}

    and = {p,q -> pandq; q,p -> qandp}
```

```
    neg={p -> negp; q -> negq}
 //istitp=[i]p
    stit={i,p->istitp; i,negp->istit_negp; j,q->jstitq;j,negq->jstit_negq }
//  iestit= Oip, agent i ought to see to it that p from the ethical point of view
    .
    estit={i,p->iestitp; j,q->jestitq; i,negp->i_estit_negp; j,negq->
        j_estit_negq}


    Rationality={w2->4; w1->2;w4->2;w3->1}

    Utility_ind={i,w1->0;i,w2->1; i,w3->1;i,w4->0;j,w1->0;j,w2->0; j,w3->2;j,w4
        ->2}
 }
```

4.4 Inference tasks

Now we have declaratively specified a total structure with a finite domain for the E-STIT logic. We are searching for an assignment to the symbols in the vocabulary V-E-STIT that expands the structure S and satisfies T-E-STIT. This can be done by the $modelexpand(T, S)$ inference. Given T-E-STIT and S-E-STIT as an input, $modelexpand$ will return either a model M such that $S \leq_p M$ and M is a totally specified structure or $UNSAT$ if such a model does not exist. The simplest way to print them is by using *printmodels()*. Also, we are able to print a specific part of the model we are more interested in. For instance, listing 1.4 illustrates the main procedure in Lua in a way that one can only print the evaluation function out of obtained model from the *modelexpand(T,S)* inference rule. The reader can find the experimental work in IDP regarding example 1 in figure 3.

Listing 1.4. Printing the result of evaluation function from the generated model

```
    //main procedure
procedure main(){
   //printmodels(modelexpand(T-E-STIT,S))
    local models = modelexpand(T-E-STIT,S)[1]
    print("The evaluation results are: ",models[V-E-STIT::eval].ct)
}
```

In order to implement a query mechanism for E-STIT in IDP we define a Lua procedure, named *query-eval*, in which we can query a world for a specific formula. We can query and print its results using a simple print command, *print(querry-eval("w1","iestitq"))* where the pair *("w1","iestitq")* evaluation is read as "agent i ought to see to it that p, from the ethical point of view". The IDP specification for query inference tasks and can be found by clicking on the link in figure 3.

5 Case Study

A Case Study is defined as a research strategy, i.e. an empirical inquiry that investigates a phenomenon within its real-life context. Case Study research can mean single or multiple case studies which can include quantitative evidence. It can rely on multiple sources of evidence or the benefits from the prior development of theoretical propositions [29]. In the field of AI, it is always helpful to use a Case Study to outline some of the key issues around the application of different ethical theories in designing and implementing ethics for machines. Here we use a Case Study to investigate two cases in our methodology for Machine Ethics.

Basically, the cases can be distinguished according to the number of agents and moments in the system but the examples are given in different fields of autonomous machines and robotics. We name each case according to the scientific field of the specified example. After a short introduction to that specific scientific

area, we present an example where a machine deals with ethical decision making. The first case is called Social Robotics and it illustrates the situation of a single autonomous agent that is making an ethical choice between possible alternatives at a single moment. In the first case, the given example is a physical assistive robot in the field of Social Robotics. The second case illustrates the situation of an autonomous agent that is obliged to violate a norm in a certain moment under the constraints, what is the best ethical action in the next moment. For the second case, the example is an autonomous driving vehicle deciding under which circumstances the crash is certain to happen (a case that is widely discussed nowadays).

5.1 Case1: Social Robotics

Social robots are robots that are designed to facilitate human interaction and communication following social rules according to their given tasks. The definition of social necessitates the presence of human beings whom the robot is supposed to interact and communicate with, which requires social values, norms and standards. A suggested ultimate requirement for social robots is to determine these social values and norms for the robot's communication skills according to the ethical theories. That's to say using for instance deontological approaches such as Isaac Asimov's three laws of robotics[6] or using the utilitarian approaches [6].

Social robots as for example educational robots, elderly care robots, museum guide robots, etcetera will soon be in charge of making ethical decisions regarding their tasks and their level of autonomy. Nobody wishes to spend money on robots who exhibit and train bad manners to children, improve the sense of infantilization for elderly people or behave impolite in the museum. In the IDP specification for E-STIT, we can investigate the behavior of robots with regard to the ethical theory which their decision process is built upon.

In this section, we discuss an example in which a single robot has to make decisions while taking care of elderly at home. Definitely, there are many different types of ethical dilemmas that a robot might encounter in this situation. According to our system, once robots recognize the moral status of their available choices, they will regulate their ethical preferences for their choices. The choices with the higher degree of ethics will be their preference over all other choices.

Example 2 (Elderly care robots). Example 2 (see figure 4) illustrates a simple example where the task of giving care to an elderly woman is given to a robot i. The robot's duties may include giving the person her medicine at the correct time, making a video call to her doctor when she is unwell or needs help beyond the robot's capabilities. If there is an emergency situation as for example a fire alarm. the robot is programmed to make contact with the fire brigade and then try to extinguishing the fire. Suppose now that such an alarm occurs at the time when the person is supposed to get her medication, but there is no apparent sign of fire. There are three options for robot i to chose among:

1. p: To give the elderly woman her medication.
2. q: To call the fire department and then try to extinguishing the fire.
3. r: To do something totally unrelated.

If i delays to give the elderly woman her medication, she might be in high risk of a heart attack but if i delays on calling for help, all might perish in a blaze before the firemen arrive. Doing nothing at all, as described in the third option, is the worst action with the lowest ethical order.

We are not here in the position to discuss the morality or utility of the three potential actions. Only suppose that p has higher morality than q and r, due to the fact that any delay in medical care would cause health problems for granny. And q has the higher utility order than r, in case a real fire has triggered the alarm. The result will be serious damages to the house that might be covered by insurance. We can define the utility and the rationality orders between the worlds as follow:

[6] The three basic rules of Isacc Asimov;

1. A robot may not injure a human being, or through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second law.

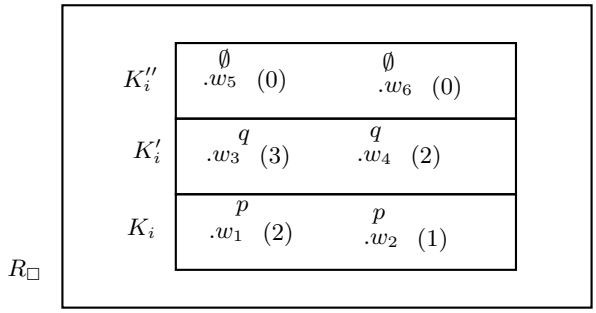

Fig. 4. Agent i's possible choices in E-STIT illustration, when the task of taking care of a person is delegated to him. At the same time of granny medical care time, the fire alarm notifies a fire.

- Deontological order between the worlds: $w_6 \approx_r w_5 <_r w_3 \approx_r w_4 <_r w_1 \approx_r w_2$
- Utilitarian order between the worlds: $w_6 \approx_u w_5 <_u w_2 <_u w_4 \approx_u w_1 <_u w_3$

From the deontology point of view, the tasks related to the human's life and health would have the highest ethical order. In our example where there is no sign of fire p receives the highest rational order in comparison to q and r. From the utilitarianism point of view, it might be the case that q has the highest utility order. Due to the alarm notification, one can obtain an ethical order following deontology as follows:

- Ethical order: $w_6 \approx_e w_5 <_e w_4 <_e w_3 <_e w_2 <_e w_1$

Regulating the preference order between worlds, we can lift it to regulate the preference between a set of worlds, representing the choice of the agent. In example 2 (figure 4), The possible choices for i are: $K_i = \{w_1, w_2\}$ performing task p, $K_i' = \{w_3, w_4\}$ performing task q and K_i'' do nothing. The ethical preference relation between the choices is as follows:

- Ethical preferences relation between choices: $K_i'' \prec K_i' \prec K_i$

In example 2, agent i prefers to give medicine to the lady first and then checks the problem with the fire. He definitely prefers the actions K_i and K_i' to K_i''. This preference relation only depends on how the applicant or programmer defines the ethical order between the worlds. Alternatively, one can easily consider K_i' (said to be informing the related institutions for the risk of fire) as the best choice. Certainly, these preferences are entirely sensitive to the society standards.

Clicking on the link in figure 5 give you an access to the online web-based IDP interpreter that contains the example 2 specifications. Example 2 is defined as a specific structure S for E-STIT specification in IDP theory, $T - E - STIT$. IDP inferences expand the model such that $S \models T - E - STIT$. Please, run the program and see the results by clicking on the link in figure 5.

http://dtai.cs.kuleuven.be/krr/idp-ide/?src=1ac22de5f7e469a622540e91e058bc70

Fig. 5. IDP specification for the social robotic case.

5.2 Case2: Autonomous vehicles

Nowadays, vehicle automation and self driving cars are progressing rapidly. The artificial automated cars are claimed to be more precise and predictable than human driving. Media reports often focuses on providing more reliable and safety expectations. Despite of the progression and precision in the automated technology and decision making process, the possibility of a crash still subsists. The question of the responsible for the

crash, i.e. the designer or the car itself, is not the concern of this work. Our consideration is mostly on the problem of ability, i.e. the problem of whether the automated vehicle has the ability to make ethically-complex decisions when driving, particularly prior to a crash.

The E-STIT logic is able to cast the optimal strategy for automated vehicles since the crash is in the next moment (in front). Goodall in [13,12] discusses the problems of autonomous driving vehicles. In this section we would like to illustrate and discuss the same example in E-STIT.

Example 3 (Autonomous vehicle [12]). Figure 7 illustrates a simple example of an autonomous vehicle which is taken from [12]. The vehicle is traveling on a two-lane bridge when a bus that is traveling in the opposite direction suddenly veers into its lane. If the bus does not correct itself, then a severe, two-vehicle crash results. The automated vehicle must decide how to react according to the ethical preference relations that have been proposed by E-STIT. The three alternatives in the next moment, when vehicle i sees to it that crash ($[i]\phi$), are as follows:

1. p: Avoid the frontal collision by off the bridge, which guarantees one-vehicle crash;
2. q: Crash head-on into the bus, which will result in a heavy, two-vehicle crash; and
3. r: Attempt to squeeze pass the bus on the right. If the bus suddenly corrects back toward its own lane a crash is avoided.

The crash in alternative 1 (p) would be a smaller offset crash which carries a lower risk of injury than the full frontal collision in alternative 2 (q). Therefore, one interpretation would allocate p the higher utility order and q due to the possibility of having less injury. Given that, r could have the highest rational order according to the deontology theories. The E-STIT Kripke model of the example is illustrated in figure 6.

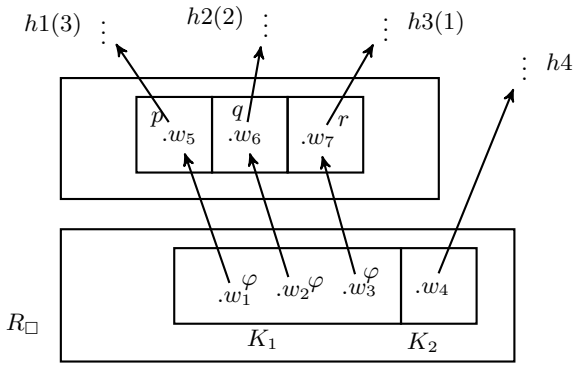

Fig. 6. Autonomous vehicle i choice when crash is inevitable

Fig. 7. Autonomous vehicle and the bus

Suppose that two different utility and rationality orders exist between the worlds is as follows:

- Utilitarian ethical order: $w_7 <_u w_6 <_u w_5$.
- Deontological ethical order: $w_6 <_r w_5 <_r w_7$

A formula such as $M, w5 \models \bigcirc_i p$ is true and a formula such as $[i]\phi \wedge X \bigcirc_i p$ [7] is valid according to the ethical order induced by utilitarianism and is false in the other ethical order.

We invite the reader to find the IDP experiment and implementation for this specific temporal Kripke model in the link represented in figure 8.

[7] material implication is not able to capture the concept of conditional, i.e. the formula: $[i]\phi \rightarrow X \bigcirc_i p$ is not valid in E-STIT. The modal conditional operator (conditional ought) is defined via dominance relation in Horty's logic. For the sake of simplicity we didn't discuss conditional modality in this paper and we leave it as a future work.

http://dtai.cs.kuleuven.be/krr/idp-ide/?src=c407a5e7d18ad4ec6628473360b18b56

Fig. 8. IDP specification for the autonomous vehicle case.

6 Summary

We believe that a logical-based approach to AI sounds promising [6,3] . Applying logical approaches to ethical theories ensures that the agents follow permissible actions by these theories. Bringsjord et. al. in [6] propose strong arguments for logical-based approached in AI. The argument of Bringsjord et. al. [6] and the deficits of individual ethical reasoning, motivate us to investigate the combination of two ethical theories reasoning in a deontic logic of agency (named E-STIT).

In E-STIT, agents make choices and the choice with the higher ethical order is preferred to any other choices. There are two metrics to evaluate the ethical order between the choices, the utility order and the rationality order. From the deontological point of view, a choice is ethically preferred to the other choice if and only if, it has a higher rational order. If the rationality order between two choices is indistinguishable then the machine decides ethically based on the utility order between concurrent choices. Similarly, from the utilitarianism point of view, this is the utility order between the worlds that has to be taken as primitive.

Knowledge base languages are examples of intermediate languages which are designed to model a formal approach on the meta level. We use IDP [22] as a such a knowledge base language to model E-STIT at the meta level. IDP integrates declarative specifications in FO$(\cdot)$ with imperative management of the specifications via Lua language [8] to solve different problems that arrise in the domain of Machine Ethics. Using the $Query(S, E)$ inference task in IDP the agent can ask about the morality of each world. Using the $Modelexpand(T, S)$ inference rule, the agent will receive all the possible models out of the theory T specifications in the knowledge base and the specific structure S for a concrete problem.

7 Related Work and Future Work

One of the primary candidates for Machine Ethics is deontic logic's formalization for the notations of obligatory, permissible and prohibited actions. Bringsjord et. al. in [3,6] utilize a logic of agency (deontic based utilitarian logic) to obtain obligatory and permissible action given the set of ethical principles into the decision procedure of an autonomous system. They argue that the proof system ensures us that 1) robots only take permissible actions, and 2) all actions that are obligatory for robots are actually performed by them. They claim that their work is free of any ethical theories while it abides to ethical principles. Ethical principles are originated from ethical theories.

In our proposed approach, we use the same logic of agency. But the concept of obligatory action is fertilized by synthesization of reasoning procedures from two ethical theories (deontology and utilitarianism). Therefore the obligatory actions are the ones that are filtered by two ethical standards. We discuss the same example in [3] (robots and softbots in ICU) to resolve it in E-STIT. We show that according to the ethical theories there might be different results for the same problem. As computer scientists we are not in the position of making ethical principles. This is a vast topic that has been discussed for a long time among ethicists.

References

1. Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford Scholarship Online, 2009.
2. Michael Anderson and Susan Leigh Anderson. *Machine ethics*. Cambridge University Press, 2011.
3. Konstantine Arkoudas, Selmer Bringsjord, and Paul Bello. Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, 2005.
4. Philippe Balbiani, Andreas Herzig, and Nicolas Troquard. Alternative axiomatics and complexity of deliberative STIT theories. *Journal of Philosophical Logic*, 37(4):387–406, 2008.
5. Nuel D Belnap. Facing the future: agents and choices in our indeterminist world. 2001.
6. Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Systems, IEEE*, 21(4):38–44, 2006.

7. Jan Broersen. Deontic epistemic *stit* logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):127 – 152, 2011.
8. Marc Denecker and Eugenia Ternovska. A logic of nonmonotone inductive definitions. *ACM transactions on computational logic (TOCL)*, 9(2):14, 2008.
9. Marc Denecker and Joost Vennekens. Building a knowledge base system for an integration of logic programming and classical logic. In *Logic Programming*, pages 71–76. Springer, 2008.
10. Dov Gabbay, Jeff Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre. Handbook of deontic logic and normative systems. 2013.
11. Lou Goble. Multiplex semantics for deontic logic. *Nordic Journal of Philosophical Logic*, 5(2):113–134, 2000.
12. Noah Goodall. Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 0(2424):58–65, 2014.
13. Noah J Goodall. Machine ethics and automated vehicles. In *Road Vehicle Automation*, pages 93–102. Springer, 2014.
14. John C Harsanyi. *Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility*. Springer, 1976.
15. Andreas Herzig and François Schwarzentruber. Properties of logics of individual and group agency. *Advances in modal logic*, 7:133–149, 2008.
16. John Horty. *Agency and Deontic Logic*. Oxford University Press, New York, 2001.
17. John Horty and Nuel Belnap. The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24:583–644, 1995.
18. Deborah G Johnson. Computer ethics. *the Philosophy of Computing and Information*, page 65, 1985.
19. Jérôme Lang and Leendert van der Torre. From belief change to preference change. In *ECAI*, volume 178, pages 351–355, 2008.
20. Emiliano Lorini. Temporal logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics*, 23(4):372–399, 2013.
21. Emiliano Lorini and Giovanni Sartor. A stit logic analysis of social influence. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 885–892. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
22. Maarten Mariën, Johan Wittocx, and Marc Denecker. The idp framework for declarative problem solving. In *Search and Logic: Answer Set Programming and SAT*, pages 19–34, 2006.
23. James M Moor. The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE*, 21(4):18–21, 2006.
24. Arthur N Prior. *Past, present and future*, volume 154. Clarendon Press Oxford, 1967.
25. Russ Shafer-Landau. *Ethical theory: an anthology*, volume 13. John Wiley & Sons, 2012.
26. Spyros G Tzafestas. Roboethics: A branch of applied ethics. In *Roboethics*, pages 65–79. Springer, 2016.
27. Johan Wittocx, Broes De Cat, and Marc Denecker. The idp system. In *Proceedings of the 22nd Benelux conference on artificial intelligence*, 2010.
28. Michael Wooldridge. Computationally grounded theories of agency. In *MultiAgent Systems, 2000. Proceedings. Fourth International Conference on*, pages 13–20. IEEE, 2000.
29. Robert K Yin. Case study research design and methods third edition. *Applied social research methods series*, 5, 2003.