How ethical frameworks answer to ethical dilemmas: towards a formal model

Vincent Bonnemains[1] and **Claire Saurel**[2] and **Catherine Tessier**[3]

Abstract.

This paper is a first step towards a formal model that is intended to be the basis of an artificial agent's reasoning that could be considered by a human as an ethical reasoning. This work is included in a larger project aiming at designing an authority-sharing manager between a robot and a human being when the human-robot system faces decision making involving ethical issues. Indeed the possible decisions in such a system will have to be considered in the light of arguments that may vary according to each agent's points of view.
The formal model allows us to translate in a more rigourous way than in natural language what is meant by various ethical frameworks and paves the way for further implementation of an "ethical reasoning" that could put forward arguments explaining one judgement or another. To this end the ethical frameworks models will be instantiated on some classical ethical dilemmas and then analyzed and compared to each other as far as their judgements on the dilemmas are concerned.

1 INTRODUCTION

Let us consider two classical ethical dilemmas. How would you react?

1. The crazy trolley
   A trolley that can no longer stop is hurtling towards five people working on the track. They will die hit by the trolley, unless you decide to move the switch to deviate the train to another track only one person is working on. What would you do? Sacrifice one person to save the other five, or let five people die?
2. The "fatman" trolley
   A trolley that can no longer stop is hurtling towards five people working on the track. This time you are on a bridge, a few meters before them, with a fat man. If you push this man on the track, he is fat enough to stop the trolley and save the five people, but he will die. Would you push the "fatman" ?

There is no really "right" answer to those dilemmas, nevertheless ethics may be used to guide reasoning about them. Therefore we will start by general definitions about ethics and related concepts.

Definition 1 (Ethics) *Ricoeur [9] defines* ethics *as compared to norm in so far as norm states what is compulsory or prohibited whereas ethics goes further and defines what is fair and what is not, for oneself and for others. It is this judgement that leads the human through their actions.*

As far as ethical dilemmas are concerned, one builds a decision on normative ethics.

Definition 2 (Principle or moral value) Principles *or* moral values *are policies, ways of acting. Example: "Thou shalt not lie".*

Definition 3 (Ethical dilemma) *An* ethical dilemma *is a situation where it is impossible to make a decision without overriding one of our principles.*

Note that the definition used (based on [11]) is the usual one, not the logic one.

Definition 4 (Normative ethics) Normative ethics *aims at building a decision through some norm established by a particular ethical framework.[3]*

Definition 5 (Ethical framework) *An* ethical framework *gives us a way for dealing with situations involving ethical dilemmas thanks to principles, metrics, etc. For example utilitarianism focuses on the consequences of a decision, the best being the one which provides the most good or does the least harm.*

We will consider that the *agent* is the entity that has to make a decision in an ethical dilemma.

In this paper, our aim is to formalize different kinds of judgements according to various ethical frameworks, in order to provide an artificial agent with the decision-making capability in front of an ethical dilemma, together with the capability to explain its decision, especially in a user/operator-robot interaction context [10]. It is inspired by two papers, [4] and [7], whose goals are close from ours, i.e. to find a way to judge how ethical is an action regarding the agent's believes.
The work of [7] is based on a model of believes, desires, values and moral rules which enables the agent to evaluate, on a boolean basis, whether each action is moral, desirable, possible, etc. According to preferences between those criteria, the agent selects an action. The main goal of this model is to allow an agent to estimate the ethics of other agents in a multi-agent system. However, the way to determine whether an action is right, fair or moral is not detailed. Moreover the paper does not question the impact of an action on the world, nor the causality between events.
The work of [4] is based on the crazy trolley dilemma, and intends to formalize and apply the Doctrine of Double Effect. The agent's responsibility, and the causality between fluents and events are studied (for example an event makes a fluent true, a fluent is

[1] ONERA and University Paul Sabatier, France, email: Vincent.Bonnemains@onera.fr
[2] ONERA, France, email: Claire.Saurel@onera.fr
[3] ONERA, France, email: Catherine.Tessier@onera.fr

necessary for an event occurrence, etc.) Nevertheless, some concepts are not deepened enough: for example, the proportionality concept is not detailed and is only based on numbers (i.e. the number of saved lives).
Both approaches have given us ideas on how to model an ethical judgement, starting from a world representation involving facts and causality, so as about some modelling issues: how to determine a moral action? how to define proportionality? As [4], we will formalize ethical frameworks, including the Doctrine of Double Effect. Moreover the judgements of decisions by the ethical frameworks are inspired by [7]. Nevertheless we will get multi-view judgements by using several ethical frameworks on the same dilemma.

We will first propose some concepts to describe the world and the ethical dilemma itself. Then we will provide details about ethical frameworks, tools to formalize them and how they judge possible choices in the ethical dilemmas. Choice (or decision) is indeed the core of our model, since it is about determining what is ethically acceptable or not according to the ethical framework. We will show that although each ethical framework gives different judgements on the different ethical dilemmas, similarities can be highlighted.

2 CONCEPTS

2.1 Assumptions

For this work we will assume that:

- The agent decides and acts in a complex world which changes.
- The ethical dilemma is studied from the agent's viewpoint.
- For each ethical dilemma, the agent has to make a decision among all possible decisions. We will consider "doing nothing" as a possible decision.
- In the context of an ethical dilemma, the agent knows all the possible decisions and all the effects of a given decision.
- Considerations as *good/bad*[4] and *positive/negative*[5] are defined as such from the agent's viewpoint.

Moreover, as some dilemmas involve the human life question, we will make the simplifying assumption:

- A human life is perfectly equal to another human life, whoever the human being is.

In the next sections we will define some concepts to represent the world and its evolution. Those concepts and their interactions are illustrated in figure 1.

2.2 World state

We characterize the environment around the agent by *world states*.

Definition 6 (World state - Set $\mathcal{S}$) *A* world state *is a vector of state components (see definition below). Let $\mathcal{S}$ be the set of world states.*

[4] A decision is good if it meets the moral values of the agent; a bad decision violates them.
[5] A fact is positive if it is beneficial for the agent; it is negative if it is undesirable for the agent.
[6] This model is not quite far from event calculus and situation calculus. As things currently stand, fluents are close to state components, and events and actions modify values of them through functions (such as $Consequence$ in this paper).

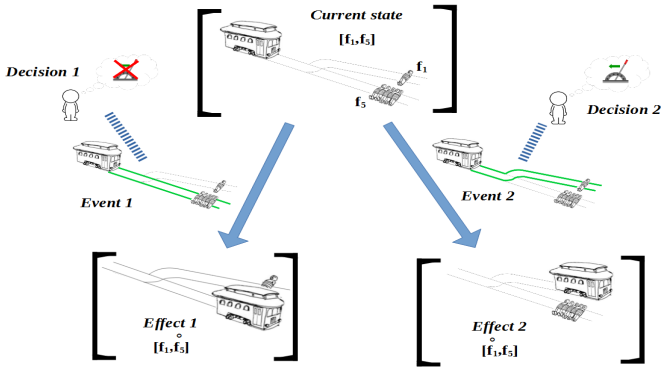

Figure 1. The world and concepts[6]

Definition 7 (State component / fact - Set $\mathcal{F}$) *A* state component, *also named* fact, *is a variable that can be instantiated only with antagonist values. We consider antagonist values as two values regarding the same item, one being the negation of the other. An item can be an object (or several objects), a living being (or several living beings), or anything else which needs to be taken into account by the agent. Let $\mathcal{F}$ be the set of state components.*

Example:

- f_5 = five people are alive
- $\overset{\circ}{f}_5$ = five people are dead

Because two values of a fact concern the same item, f_5 and $\overset{\circ}{f}_5$ concern the same five people.
Depending on the context "$\circ$" will not have exactly the same meaning. This notation allows us to consider antagonist values such as gain/loss, gain/no gain, loss/no loss, etc. Those values have to be defined for each fact.
Consequently an example of a world state is:

$$s \in \mathcal{S},\ s = [f_1, \overset{\circ}{f}_5],\ f_1, \overset{\circ}{f}_5 \in \mathcal{F} \tag{1}$$

2.3 Decision, event, effect

Definition 8 (Decision - Set $\mathcal{D}$) *A* decision *is a choice of the agent to do something, i.e. perform an* action*, or to do nothing and let the world evolve. Let $\mathcal{D}$ be the set of decisions.*

When the agent makes a decision, this results in an event that modifies the world. Nevertheless an event can also occur as part of the natural evolution of the world, including the action of another agent. Consequently we will differentiate the *event* concept from the agent's *decision* concept.

Definition 9 (Event - Set $\mathcal{E}$) *An* event *is something that happens in the world that modifies the world, i.e. some states of the world. Let $\mathcal{E}$ be the set of events.*

Let $Event$ be the function computing the event linked to a decision:

$$Event\ :\ \mathcal{D} \rightarrow \mathcal{E} \tag{2}$$

The consequence of an event is the preservation or modification of state components. The resulting state is called *effect*.

Definition 10 (Effect) *The* effect *of an event is a world state of the same dimension and composed of the same facts as the world state before the event; only the values of facts may change.* $Effect \in \mathcal{S}$. *Let* $Consequence$ *be the function to compute the effect from current state:*

$$Consequence \ : \ \mathcal{E} \times \mathcal{S} \rightarrow \mathcal{S} \tag{3}$$

Example:

$$f_1, f_5, \overset{\circ}{f_5} \ \in \ \mathcal{F} \tag{4}$$

$$e \ \in \ \mathcal{E} \tag{5}$$

$$i \in \mathcal{S}, i \ = \ [f_1, f_5] \tag{6}$$

$$Consequence(e, i) \ = \ [f_1, \overset{\circ}{f_5}] \tag{7}$$

In the case of the crazy trolley dilemma, if the agent's decision is to "do nothing" (no action of the agent), the trolley will hit the five people (event) and they will be killed (effect). If the agent's decision is to "move the switch" (decision), the trolley will hit one person (event); and they will be killed (effect).

3 ETHICAL FRAMEWORKS

3.1 Judgement

The agent will make a decision according to one or several ethical frameworks. Each ethical framework will issue a judgement on a decision, e.g. on the decision nature, the event consequence, etc. When several ethical frameworks are considered by the agent, their judgements may be confronted to compute the agent's resulting decision, see figure 2:

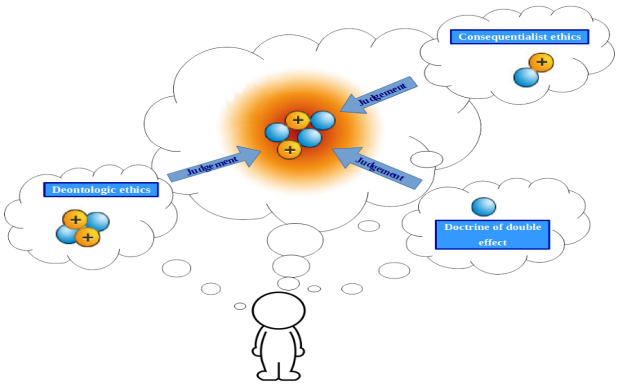

Figure 2. Decision computing from ethical frameworks judgements

Indeed the judgement of an ethical framework determines whether a decision is $acceptable$, $unacceptable$ or $undetermined$ as regards this ethical frame. A decision is judged $acceptable$ if it does not violate the principles of the ethical framework. A decision is judged $unacceptable$ if it violates some principles of the ethical framework. If we cannot determine whether the decision violates principles or not, it is judged $undetermined$. Let $\mathcal{V}$ be the set

$$\mathcal{V} = \{acceptable(\top), undetermined(?), unacceptable(\bot)\} \tag{8}$$

All judgements have the same signature:

$$Judgement \ : \ \mathcal{D} \times \mathcal{S} \rightarrow \mathcal{V} \tag{9}$$

The literature highlights three major ethical frameworks [8]: consequentialist ethics, deontological ethics and virtue ethics.
As far as virtue ethics is concerned, it deals with the agent itself in so far as the agent tries to be the best possible agent: through some decisions, some actions, it becomes more or less virtuous. Virtues could be: honesty, generosity, bravery, etc.[5]. However it seems difficult to confer virtues on an artificial agent as they are complex human properties. Consequently, according to [2], we will not consider an artificial agent as virtuous or not in this paper.
By contrast, and according to [4], we will consider the Doctrine of Double Effect although it is not one of the three main frameworks. Indeed it uses some concepts of them and introduces some other very relevant concepts such as causality and proportionality [6].

3.2 Consequentialist ethics

This ethical framework focuses only on the consequences of an event. According to consequentialist ethics, the agent will try to have the best possible result (i.e. the best effect), disregarding the means (i.e. the event). The main issue with this framework is to be able to compare the effects of several events, i.e. to compare sets of facts. Consequently

- we will distinguish between positive facts and negative facts within an effect;
- we want to be able to compute preferences between effects, i.e. to compare set of positive (resp. negative) facts of an effect with set of positive (resp. negative) facts of another effect.

3.2.1 *Positive/Negative facts*

Let $Positive$ and $Negative$ the functions:

$$Positive/Negative \ : \ \mathcal{S} \rightarrow \mathcal{P}(\mathcal{F}) \tag{10}$$

returning the subset of facts estimated as positive (resp. negative) from an effect.
In this paper, we assume that for an effect s:

$$Positive(s) \cap Negative(s) = \emptyset \tag{11}$$

3.2.2 *Preference*

Let $\succ_c$ be the preference relation on subsets of facts ($\mathcal{P}(\mathcal{F})$).
$F_1 \succ_c F_2$ means that subset F_1 is preferred to subset F_2 from the consequentialist viewpoint. Intuitively we will assume the following properties of $\succ_c$:

- if a subset of facts F_1 is preferred to another subset F_2, thus it is impossible to prefer F_2 to F_1.

$$F_1 \succ_c F_2 \rightarrow \neg(F_2 \succ_c F_1) \tag{12}$$

- if F_1 is preferred to F_2 and F_2 is preferred to another subset of facts F_3, then F_1 is preferred to F_3.

$$[(F_1 \succ_c F_2) \ \wedge \ (F_2 \succ_c F_3)] \ \rightarrow \ F_1 \succ_c F_3 \tag{13}$$

- A subset of facts cannot be preferred to itself.

$$\nexists \, F_i \ / \ F_i \succ_c F_i \tag{14}$$

Consequently $\succ_c$ is a strict order (irreflexive, asymmetric and transitive).

3.2.3 *Judgement function*

A decision d_1 involving event e_1 ($Event(d_1) = e_1$) is considered better by the consequentialist framework than decision d_2 involving event e_2 ($Event(d_2) = e_2$) iff for $i \in \mathcal{S}$:

$$Positive(Consequence(e_1, i)) \succ_c Positive(Consequence(e_2, i)) \tag{15}$$

and

$$Negative(Consequence(e_1, i)) \succ_c Negative(Consequence(e_2, i)) \tag{16}$$

Those equations are both consequentialism concepts:

- *positive consequentialism* (15), trying to have the "better good"
- *negative consequentialism* (16), trying to have the "lesser evil"

If both properties are satisfied, then

$$Judgement_c(d_1, i) = \top, \ and \ Judgement_c(d_2, i) = \bot \tag{17}$$

If at least one property is not satisfied, there is no best solution:

$$Judgement_c(d_1, i) = Judgement_c(d_2, i) = ? \tag{18}$$

In the case of a dilemma with more than two possible decisions, the best decision is the decision that is judged better than all the others. If such a decision does not exist, it is impossible to determine an $acceptable$ solution with consequentialist ethics. Nevertheless if there is a decision d_1 with another decision d_2 better than d_1, then d_1 is judged $unacceptable$, as d_1 cannot be the best.

3.3 Deontological ethics

This ethical framework focuses only on the nature of the decision, no matter the consequences. Indeed the agent wants to make a moral decision, which is close to abide by norms or to Kant's theory. Therefore we have to define the nature of a decision.

3.3.1 *Decision nature*

A decision may be good, neutral, bad or undetermined from the agent's point of view. Let $\mathcal{N}$ be the set

$$\mathcal{N} = \{good, neutral, bad, undetermined\} \tag{19}$$

There is a partial order $<_d$ in $\mathcal{N}$:

$$bad <_d neutral <_d good \tag{20}$$

Meaning that a good nature is preferable to a neutral which is preferable to a bad. $undetermined$ cannot be ordered, because it represents a lack of information.
We assume intuitively that:

$$bad <_d good \tag{21}$$

Likewise, we admit that $good <_d bad$ is false. We also define the following relations:

- $=_d$, for example $good =_d good$
- $\leq_d$: $a \leq_d b$ iff $a <_d b$ or $a =_d b$.

Function $DecisionNature$ allows the nature of a decision to be obtained:

$$DecisionNature : \mathcal{D} \to \mathcal{N} \tag{22}$$

Example: $DecisionNature(to\ kill) = bad$. We will not explain further here how this function works but it is worth noticing that judging a decision from the deontological viewpoint is quite complex and depends on the context. For example denunciate a criminal or denunciate someone in 1945 are likely to be judged differently. It is even more complex to estimate the nature of a decision which is not linked to the agent's action. For example if the agent witnesses someone is lying to someone else, is it bad "to not react"?

3.3.2 *Judgement function*

The deontological framework will judge a decision with function $Judgement_d$ as follows: $\forall d \in \mathcal{D}, \forall i \in \mathcal{S}$ (Indeend initial state doesn't matter in this framework)

$$DecisionNature(d) \geqslant_d neutral \Rightarrow Judgement_d(d, i) = \top \tag{23}$$
$$DecisionNature(d) =_d undetermined \Rightarrow Judgement_d(d, i) = ? \tag{24}$$
$$DecisionNature(d) <_d neutral \Rightarrow Judgement_d(d, i) = \bot \tag{25}$$

3.4 The Doctrine of Double Effect(DDE)

The Doctrine of Double Effect is considered here as an ethical framework, as in other papers [4]. Indeed DDE allows some distinctions between decisions to be highlighted whereas other frameworks cannot. DDE can be described by three rules:

1. **Deontological rule:** the decision has to be *good* or *neutral* according to deontological ethics.
2. **Collateral damage rule:** Negative facts must be neither an end nor a mean (example: collateral damages).
3. **Proportionality rule:** the set of Negative facts has to be proportional to the set of Positive facts.

We already have the tools required for the first rule (see 3.3.1).
The second rule involves something else as until now, the difference between causal deduction (e.g. if I unplug the computer, it turns off) and temporal deduction (e.g. if I erase a file on the boss's computer, I will be fired) has not been considered. Only a function between an event and its effect has been defined and it does not any difference between an event preventing the occurrence of a fact which would happened as a natural evolution and an event inducing a fact by causality. As for the third rule, we need to define what proportional means.

3.4.1 *Causality*

Let us consider two facts that are causally connected, what does it mean? This link is not always a logical implication. Indeed it could be an inference, but such an inference is not always direct or instant. That is why we will use a symbol of temporal modal logic:

$$p \vdash Fq \tag{26}$$

which means the occurrence of p induces the occurrence of q (in all possible futures): fact p is a way to obtain fact q.
Example:

$$buy\ candy \vdash Fpossess\ candy \tag{27}$$

3.4.2 Proportionality

First of all, it is necessary to define which meaning of proportionality is needed. Indeed the concept is complex as it is a relation between positive and negative facts.
Examples:

1. It is proportional, in response to a cockroaches invasion, to set traps in a house. But it is not proportional to drop an A-bomb on the house to eliminate cockroaches.
   Nevertheless proportionality is less obvious in other cases, for instance :
2. Someone will consider that it is proportional to give a certain amount of money for exchange of a thing or a service, while someone else will think that it is not (e.g. too expensive).
3. Even if it is "easy" to compare the loss of one life to the loss of several lives, what about the comparison between the loss of one life and the safeguard of several lives?

In this paper, proportionality is implemented by relation $\lesssim_p$ between facts ($\mathcal{F}$).
$f_1 \lesssim_p f_2$ means that f_1 is proportional to f_2, i.e. f_1 has an importance lower than or close to the importance of f_2. *Importance* depends on the context and on the agent.
There is no fact closer of a fact than the fact itself. For example the most equivalent response to a slap is another slap. Thereby we will assume that a fact is proportional to itself.

$$\forall f_i \in \mathcal{F} \rightarrow f_i \lesssim_p f_i \tag{28}$$

$\lesssim_p$ is therefore reflexive.
Furthermore if f_1 has an importance lower than or close to the importance of f_2 ($f_1 \lesssim_p f_2$), and the importance of f_2 is lower than or close to the importance of f_3 ($f_2 \lesssim_p f_3$), thus the importance of f_1 is necessary lower than or close to the importance of f_3 ($f_1 \lesssim_p f_3$). For example, if a murder is considered worse (i.e. more important) than a theft ($theft \lesssim_p murder$), and if a theft is considered worse than a lie ($lie \lesssim_p theft$), thus a murder is worse than a lie ($lie \lesssim_p murder$).

$$\forall f_1, f_2, f_3 \in \mathcal{F} \,/\, (f_1 \lesssim_p f_2 \wedge f_2 \lesssim_p f_3) \rightarrow f_1 \lesssim_p f_3 \tag{29}$$

$\lesssim_p$ is transitive.
By contrast, $f_1 \lesssim_p f_2$ does not mean that $f_2 \lesssim_p f_1$. It is true only if the importances of both facts are close. For example it is proportional to hit someone who threatens me with a gun, but it is not proportional to threaten someone with a gun if they hit me.
$\lesssim_p$ is neither symmetric nor asymmetric.

We extend the relation $\lesssim_p$ to a relation $\precsim_p$ between sets of facts, which means that the set of facts at the left of the symbol is proportional to the set of facts at the right. Two criteria can be considered to compute $\precsim_p$, they are inspired from [1]:

Democratic proportional criterion : a set of facts F is proportional to a set of facts G ($F \precsim_p G$) iff:

$$\forall f \in F, \exists g \in G / f \lesssim_p g \tag{30}$$

which means that every single element of F needs to be proportional to an element of G.

Elitist proportional criterion : a set of facts F is proportional to a set of facts G ($F \precsim_p G$) iff:

$$\forall g \in G, \exists f \in F / f \lesssim_p g \tag{31}$$

which means that every single element of G needs to have an element of F proportional to itself.

Example: Sam wants a candy, if he steals it, he will feel guilty, which he considers acceptable and proportional to have a candy, but he will be punished too, which is too bad for a candy, not proportional from his point of view. Another solution is to buy candy. Of course, he will have no more money after that but, to have a candy, it is proportional, and even better, the seller will offer him a lollipop, which is proportional to have no more money too! The last solution is to kill the seller to take the candy. By doing that, he will have candy, but he will go to jail, which is not proportional, and he will never have candy again, which is not proportional either.

To steal candy
Positive facts : $candy$
Negative facts : $guilty, punished$

$$guilty \lesssim_p candy \tag{32}$$

We want to know if $\{guilty, punished\} \precsim_p \{candy\}$. With the *elitist proportional criterion*, all facts of the set at the right of the symbol need to have (at least) a fact of the set at the left of the symbol proportional to themselves. Here this criterion is satisfied, $candy$ is the only fact at the right of the symbol, and $guilty$ at the left is proportional to $candy$ (32). But, with the *democratic proportional criterion*, all facts of the set at the left of the symbol have to be proportional to (at least) one fact of the set at the right of the symbol. And, even if $guilty$ is proportional to $candy$, $punished$ is not proportional to any fact. Thus, the democratic proportional criterion is not satisfied.

To buy candy
Positive facts : $candy, lollipop$
Negative facts : $no\ more\ money$

$$no\ more\ money \lesssim_p candy \tag{33}$$

$$no\ more\ money \lesssim_p lollipop \tag{34}$$

We want to know if $\{no\ more\ money\} \precsim_p \{candy, lollipop\}$. $no\ more\ money$ is proportional to $candy$ and $lollipop$ (33,34) therefore both criteria are satisfied.

To kill the seller
Positive facts : $candy$
Negative facts : $jail, no\ more\ candy\ for\ ever$
We want to know if $\{jail, no\ more\ candy\ for\ ever\} \precsim_p \{candy\}$. But in this case, there is no proportionality between negative and positive facts. Therefore no criterion is respected.

Therefore, it is possible to use the *democratic proportional criterion* or the *elitist proportional criterion* or both of them to determine whether a set of facts is proportional to another set of facts.

3.4.3 Judgement function

Thanks to the previous tools, we can now assess whether a decision meets the DDE rules.
Let i be the initial state and d the decision:

$$e = Event(d) \tag{35}$$

$$s = Consequence(e, i) \tag{36}$$

1. Deontological rule: decision d has to be good or neutral according to deontological ethics.

$$DecisionNature(d) \geqslant_d neutral \tag{37}$$

2. **Collateral damage rule:** negative facts must be neither an end nor a mean (such as collateral damages). It can be expressed as:

$$\forall f_n \in Negative(s), \nexists f_p \in Positive(s),\ (f_n \vdash F f_p) \quad (38)$$

The "evil wish" (negative fact(s) as a purpose) is not considered as we assume that the agent is not designed to make the evil.

3. **Proportionality rule:** the set of negative facts has to be proportional to the set of positive facts.

$$Negative(s) \precsim_p Positive(s) \quad (39)$$

A decision d is *acceptable* for the DDE if it violates no rule, which means:

$$[\quad DecisionNature(d) \geqslant_d neutral \quad (40)$$
$$\wedge \quad \forall f_n \in Negative(s), \nexists f_p \in Positive(s),\ (f_n \vdash F f_p) \quad (41)$$
$$\wedge \quad Negative(s) \precsim_p Positive(s) \quad] \quad (42)$$
$$\Rightarrow \quad Judgement_{dde}(d, i) = \top \quad (43)$$

4 INSTANTIATION: ETHICAL DILEMMAS

This section focuses on how our model can be instantiated on the ethical dilemmas that have been introduced at the beginning of the paper. For each dilemma the agent has to choose a decision. We will describe how consequentialist ethics, deontological ethics and the Doctrine of Double Effect assess the agent's possible decisions.

4.1 The crazy trolley

4.1.1 World, decisions, effects

Facts

- f_5: five people alive
- f_1: one person alive
- $\overset{\circ}{f_5}$: five people dead
- $\overset{\circ}{f_1}$: one person dead

Initial state : the six people are alive.

$$i = [f_5, f_1] \quad (44)$$

Decisions and effects

1. move the switch: this decision results in the train hitting one person (event). The consequence will be : five people alive, one person dead.

$$Event(move\ the\ switch) = train\ hits\ one\ person \quad (45)$$

$$Consequence(train\ hits\ one\ person, i) = [f_5, \overset{\circ}{f_1}] \quad (46)$$
$$Positive([f_5, \overset{\circ}{f_1}]) = \{f_5\} \quad (47)$$
$$Negative([f_5, \overset{\circ}{f_1}]) = \{\overset{\circ}{f_1}\} \quad (48)$$

2. do nothing: this decision is associated with the train hitting five people. The consequence is : five people dead, one person alive.

$$Event(do\ nothing) = train\ hits\ five\ people \quad (49)$$

$$Consequence(train\ hits\ five\ people, i) = [\overset{\circ}{f_5}, f_1] \quad (50)$$
$$Positive([\overset{\circ}{f_5}, f_1]) = \{f_1\} \quad (51)$$
$$Negative([\overset{\circ}{f_5}, f_1]) = \{\overset{\circ}{f_5}\} \quad (52)$$

4.1.2 Study under ethical frameworks

Consequentialist ethics

Facts can be compared with one another as they involve numbers of lives and deaths of people only.[7]
With consequentialist ethics we have

$$\{f_5\} \succ_c \{f_1\} \quad (53)$$

meaning that it is better to have five people alive than one person alive (numerical order $5 > 1$), and

$$\{\overset{\circ}{f_1}\} \succ_c \{\overset{\circ}{f_5}\} \quad (54)$$

meaning that it is better to lose one life than five lives (reverse numerical order $1 > 5$).
Therefore

$$Positive([f_5, \overset{\circ}{f_1}]) \succ_c Positive([\overset{\circ}{f_5}, f_1]) \quad (55)$$
$$Negative([f_5, \overset{\circ}{f_1}]) \succ_c Negative([\overset{\circ}{f_5}, f_1]) \quad (56)$$

Consequently (15,16)

$$Judgement_c(move\ the\ switch, i) = \top \quad (57)$$
$$Judgement_c(do\ nothing, i) = \bot \quad (58)$$

Deontological ethics

Let us assess the nature of both possible decisions:

$$DecisionNature(move\ the\ switch) = neutral \quad (59)$$

$$DecisionNature(do\ nothing) = neutral \quad (60)$$

No decision is unacceptable from the deontological viewpoint:

$$\forall d,\ DecisionNature(d) \geqslant neutral \quad (61)$$

Consequently

$$Judgement_d(move\ the\ switch, i) = Judgement_d(do\ nothing, i) = \top \quad (62)$$

Doctrine of Double Effect

Let us examine the three rules.

1. *Deontological rule*: we have seen above that both decisions are neutral. Therefore both of them satisfy the first rule.
2. *Collateral damage rule*:
   - move the switch:

$$Negative([f_5, \overset{\circ}{f_1}]) = \{\overset{\circ}{f_1}\} \quad (63)$$
$$\nexists f_p \in Positive([f_5, \overset{\circ}{f_1}]), \overset{\circ}{f_1} \vdash F f_p \quad (64)$$

   - do nothing:

$$Negative([\overset{\circ}{f_5}, f_1]) = \{\overset{\circ}{f_5}\} \quad (65)$$
$$\nexists f_p \in Positive([\overset{\circ}{f_5}, f_1]), \overset{\circ}{f_5} \vdash F f_p \quad (66)$$

Therefore both decisions respect the second rule.

[7] For the sake of simplicity in this paper, we will consider that $\{f_5\} >_c \{f_1\}$ if f_5 is preferred to f_1

3. *Proportionality rule*: we will assume in this context that the death of one person is proportional to the safeguard of the lives of the five other people, and conversely that the death of five people is not proportional to safeguard one life: $\overset{\circ}{f_1} \lesssim_p f_5$ and $\neg(\overset{\circ}{f_5} \lesssim_p f_1)$.

   Both the democratic and the elitist proportional criteria (3.4.2) give the same results as sets of facts are composed of one fact.

$$[Negative([f_5, \overset{\circ}{f_1}]) = \{\overset{\circ}{f_1}\}] \precsim_p [Positive([f_5, \overset{\circ}{f_1}]) = \{f_5\}] \tag{67}$$

   Move the switch is the only decision which respects the proportionality rule.

Consequently

$$Judgement_{dde}(move\ the\ switch, i) = \top \tag{68}$$
$$Judgement_{dde}(do\ nothing, i) = \bot \tag{69}$$

Synthesis

Table 1 is a synthesis of the judgements obtained for the crazy trolley dilemma:

Table 1. Decisions for crazy trolley judged by ethical frameworks

Framework / Decision	Conseq*	Deonto*	DDE
Move the switch	$\top$	$\top$	$\top$
Do nothing	$\bot$	$\top$	$\bot$

$\top$ Acceptable $\bot$ Unacceptable
Conseq*: Consequentialist ethics — Deonto*: Deontological ethics
DDE: Doctrine of Double Effect

4.2 "Fatman" trolley

We will just highlight what differs from the crazy trolley dilemma.

4.2.1 World, decisions, effects

Facts : Fact f_5 is the same whereas fact f_1 is replaced by fat.

- fat: "fatman" alive
- $\overset{\circ}{fat}$: "fatman" dead

Initial state : $i = [f_5, fat]$, the five people and "fatman" are alive.
Decisions and effects *Move the switch* is replaced by *push "fatman"*

1. push "fatman": this decision results in the train crashing on "fatman"(e).

$$Event(push\ "fatman") = e \tag{70}$$
$$Consequence(e, i) = [f_5, \overset{\circ}{fat}] \tag{71}$$
$$Positive([f_5, \overset{\circ}{fat}]) = \{f_5\} \tag{72}$$
$$Negative([f_5, \overset{\circ}{fat}]) = \{\overset{\circ}{fat}\} \tag{73}$$

2. *do nothing* is equivalent to the same decision in the crazy trolley.

4.2.2 Study under ethical frameworks

Decision *do nothing* has same judgements as in the previous case. Let us study the judgements for decision *push "fatman"*.

Consequentialist ethics

The result in terms of human lives is the same as in the first dilemma. Consequently we have exactly the same judgement.

$$Judgement_c(push\ "fatman", i) = \top \tag{74}$$

Deontological ethics

Let us consider decision nature of *push "fatman"* as bad.

$$DecisionNature(push\ "fatman") = bad \tag{75}$$
$$Judgement_d(push\ "fatman", i) = \bot \tag{76}$$

Doctrine of Double Effect

1. *Deontological rule*: decision *push "fatman"* does not respect the first rule.
2. *Collateral damage rule*:
   - push "fatman":

$$Negative([f_5, \overset{\circ}{fat}]) = \{\overset{\circ}{fat}\}$$
$$\overset{\circ}{fat} \vdash F f_5$$

     and

$$f_5 \in Positive([f_5, \overset{\circ}{fat}])$$

     It is because "fatman" is pushed that the five people are alive. Therefore

$$Judgement_{dde}(push\ "fatman", i) = \bot \tag{77}$$

3. *Proportionality rule*: if we assume that:

$$\overset{\circ}{fat} \lesssim f_5 \tag{78}$$
$$\neg(\overset{\circ}{f_5} \lesssim fat) \tag{79}$$

   with the same reasoning as for the crazy trolley, *push "fatman"* respects the proportionality rule.

Consequently *push "fatman"* only respects one rule out of three:

$$Judgement_{dde}(push\ "fatman", i) = \bot \tag{80}$$

Synthesis

Table 2 is a synthesis of the judgements obtained for the "fatman" trolley dilemma:

Table 2. Decisions for "fatman" trolley judged by ethical frameworks

Framework / Decision	Conseq*	Deonto*	DDE
Push "fatman"	$\top$	$\bot$	$\bot$
Do nothing	$\bot$	$\top$	$\bot$

This variant of the first dilemma is interesting because it allows us to distinguish some ethical frameworks particularities. We can see for example the usefulness of collateral damage rule for the DDE. Furthermore, the consequentialist framework does not make any difference between both dilemmas, contrary to the deontological framework or the DDE.

5 ANALYSES

Once the judgements are computed, we can analyse the similarities between ethical frameworks. Two frameworks are similar if they have common judgements values on the same decisions compared to the total number of decisions.

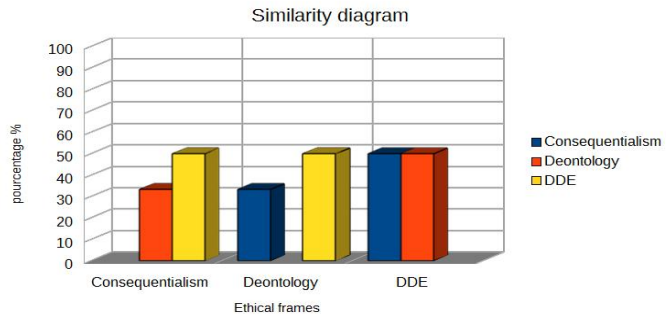

Figure 3. Similarity diagram between ethical frameworks. Each bar illustrates similarity between the framework whose name is under the bar, and the framework whose color is in the caption. The higher the bar, the more similar the frameworks.

Figure 3 is based on three dilemmas (the crazy trolley, the "fatman" trolley, and another one – *UAV vs missile launcher* – that is not described here).

We can notice that the consequentialist and deontological frameworks are quite different and that the DDE is close to the two others. This can be explained by the rules of the DDE, which allow this framework to be both deontological (deontological rule) and close to consequentialism (proportionality rule).

6 DISCUSSION

Because of their own natures, the three ethical frameworks that we have studied do not seem to be appropriate in all situations. For example we have seen that consequentialist ethics does not distinguish between crazy trolley and "fatman" trolley dilemmas. Moreover the consequentialist preference relation between facts is a partial order, which means that it is not always possible to prefer some facts to others. Consequently judging a decision is sometimes impossible with consequentialist ethics. Furthermore consequentialist preference depends on the context: preferring to feel pain in order to stop the fall of a crystal glass with one's foot does not mean that you prefer to cut your finger to get back a ring. As far as deontological ethics is concerned, judging the nature of some decisions can be tricky (see 3.3.1). Finally the Doctrine of Double Effect forbids the sacrifice of oneself. Nevertheless if a human life is threatened, shouldn't the agent's sacrifice be expected?

This leads us to the idea that one framework alone is not efficient enough to compute an ethical decision. It seems necessary to consider as much ethical frameworks as possible in order to obtain the widest possible view.

The limits of the model lie mainly in the different relations it contains. Indeed, we have not described how orders are assessed. Moreover it may be hardly possible to define an order (i.e. consequentialist preference) between two concepts. On the other hand the model is based on facts that are assumed to be certain, which is quite different in the real world where some effects are uncertain or unexpected. Furthermore, the vector representation raises a classical modelling problem: how to choose state components and their values? The solution we have implemented is to select only facts whose values change as a result of the agent's decision.

7 CONCLUSION

The main challenge of our model is to formalize philosophical definitions described with natural language and to translate them in generic concepts that can be easy-to-understand by everyone. The interest of such a work is to get rid of ambiguities in a human/robot, and more broadly human/human, system dialog and to allow an artificial agent to compute ethical considerations by itself. This formalism raises many questions because of ethical concepts themselves (DDE's proportionality, the good, the evil, etc.). Indeed ethics is not universal, that is why it is impossible to reason on fixed preferences and calculus. Many parameters such as context, agent's values, agent's priorities, etc. are involved. Some of those parameters can depend on "social acceptance". For example, estimating something negative or positive (or computing a decision nature) can be based on what society thinks about it, as on agent's values.

Further work will focus on considering other frameworks such as virtue ethics on the one hand and a value system based on a partial order on values on the other hand. Furthermore game theory, voting systems or multicriteria approaches may be worth considering to compare ethical frameworks judgements.

ACKNOWLEDGEMENTS

We would like to thank ONERA for providing resources for this work, the EthicAA project team for discussions and advice, and reviewers who gave us relevant remarks.

REFERENCES

[1] V. Royer C. Cayrol and C. Saurel, 'Management of preferences in assumption-based reasoning', in *4th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 13–22, (1993).
[2] T. de Swarte, 'Un drone est-il courageux ?', *Lecture Notes in Computer Science*, (2014).
[3] Encyclopædia Britannica, 'Normative ethics', Encyclopædia Britannica Inc., (2016).
[4] G. Bourgne F. Berreby and J-G. Ganascia, *Logic for Programming, Artificial Intelligence, and Reasoning: 20th International Conference, (LPAR-20 2015)*, chapter Modelling Moral Reasoning and Ethical Responsibility with Logic Programming, Springer, Suja,Fiji, 2015.
[5] R. Hursthouse, 'Virtue ethics', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, fall edn., (2013).
[6] A. McIntyre, 'Doctrine of Double Effect', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, Winter edn., (2014).
[7] G. Bonnet N. Cointe and O. Boissier, 'Ethical Judgment of Agents Behaviors in Multi-Agent Systems', in *Autonomous Agents and Multiagent Systems International Conference (AAMAS 2016)*, 2016, Singapore.
[8] R. Ogien, 'Les intuitions morales ont-elles un avenir ?', *Les ateliers de l'éthique/The Ethics Forum*, **7**(3), 109–118, (2012).
[9] P. Ricoeur, 'Ethique et morale', *Revista Portuguesa de Filosofia*, **4**(1), 5–17, (1990).
[10] The ETHICAA team, 'Dealing with ethical conflicts in autonomous agents and multi-agent systems', in *AAAI 2015 Workshop on AI and Ethics*, Austin Texas USA, (January 2015).
[11] CNRS TLFi.