Ethics and Information Technology (2018) 20:41–58
https://doi.org/10.1007/s10676-018-9444-x

ORIGINAL PAPER

Embedded ethics: some technical and ethical challenges

Vincent Bonnemains[1] · Claire Saurel[1] · Catherine Tessier[1]

Published online: 27 January 2018

Abstract

This paper pertains to research works aiming at linking ethics and automated reasoning in autonomous machines. It focuses on a formal approach that is intended to be the basis of an artificial agent's reasoning that could be considered by a human observer as an ethical reasoning. The approach includes some formal tools to describe a situation and models of ethical principles that are designed to automatically compute a judgement on possible decisions that can be made in a given situation and explain why a given decision is ethically acceptable or not. It is illustrated on three ethical frameworks—utilitarian ethics, deontological ethics and the Doctrine of Double effect whose formal models are tested on ethical dilemmas so as to examine how they respond to those dilemmas and to highlight the issues at stake when a formal approach to ethical concepts is considered. The whole approach is instantiated on the drone dilemma, a thought experiment we have designed; this allows the discrepancies that exist between the judgements of the various ethical frameworks to be shown. The final discussion allows us to highlight the different sources of subjectivity of the approach, despite the fact that concepts are expressed in a more rigorous way than in natural language: indeed, the formal approach enables subjectivity to be identified and located more precisely.

Keywords Ethical dilemma · Ethical framework · Autonomous machines · Judgement · Subjectivity

Introduction

Autonomous robots, autonomous cars and autonomous weapons regularly hit the headlines with a trend towards sensationalism. From a technical point of view, what is at stake is the behaviour of machines that are equipped with situation assessment and decision functions, i.e. programs that compute the actions to be achieved by the machine on the basis of a state of the world that is itself computed from information gathered through sensors and communication means.

According to the US DoD Defense Science Board (2016) *autonomy results from delegation of a decision to an authorized entity to take action within specific boundaries. An important distinction is that systems governed by prescriptive rules that permit no deviations are automated, but they are not autonomous. To be autonomous, a system must have the capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation.* In the same way Grinbaum et al. (2017) claim that autonomy is the *capacity to operate independently from a human operator or from another machine, by exhibiting non-trivial behaviours in a complex and changing environment*. Computing and achieving context-adapted actions are typical examples of such non-trivial behaviours.

Associating *Ethics* with *Autonomous machine* can mean three different things:

1. an ethical thought concerning research on autonomous machines and the design of autonomous machines,
2. or an ethical thought concerning the use and misuse of autonomous machines and how autonomous machines can be part of society,
3. or a technical approach aiming at imbuing ethics into an autonomous machine.

✉ Vincent Bonnemains
vincent.bonnemains@onera.fr

Claire Saurel
claire.saurel@onera.fr

Catherine Tessier
catherine.tessier@onera.fr

[1] ONERA, Toulouse, France

This paper focuses on point 3 while taking a critical look at the approach as suggested by point 1. Therefore it pertains to research works aiming at linking ethics and automated reasoning in autonomous machines. This is done through a formal model[1] of adapted ethical principles to automatically compute a judgement on possible decisions and explain why a given decision is ethically acceptable or not. Indeed what is intended in this work is to test several ethical frameworks on ethical dilemmas so as to examine how they respond to those dilemmas and to highlight the issues at stake when a formal approach to ethical concepts is considered.

Ethical thoughts ("Ethics and autonomous machines" section) lead us to consider what an ethics-embedding autonomous machine is and to highlight some related issues. We claim that thought experiments are simple yet useful scenarios for a cautious and critical approach to the design of automated ethical reasoning. In "How to model ethics embedded into autonomous machines" section, we review the related literature through four points of view: top–down, bottom–up, hybrid approaches and Values/Ethics personal systems approaches our own proposal belongs to. After defining an autonomous agent through some formal concepts ("Concepts for describing an ethical dilemma situation" section), we suggest formal models of various ethical frameworks (utilitarian ethics, deontological ethics and the Doctrine of Double effect) and define their judgements on possible decisions ("Ethical frameworks models" section). The whole approach is instantiated on the drone dilemma ("Instantiation on the drone dilemma" section), a thought experiment we have designed: this allows the discrepancies that exist between the judgements of the various ethical frameworks to be shown. The final discussion ("Discussion" section) allows us to highlight the different sources of subjectivity of the approach.

Ethics and autonomous machines

Which ethics is at stake?

We will focus on the fields of ethics that are relevant for automated reasoning in autonomous machines (Bringsjord and Taylor 2011):

- Normative ethics, which aims at judging a person or an action through some particular moral theories (MacIntyre 2003);
- Applied ethics, the ethics of particular application fields, which aims at dealing with real-life situations;
- Meta-ethics, the ethics of ethics, which focuses on the axiomatic concepts used by normative ethics and applied ethics—e.g. what "wrong" and "right" mean—and on how to apply them.

Normative ethics is the basis of our approach in so far as we formalize various ethical frameworks to compute judgements on autonomous agents' possible behaviours, decisions and actions in various situations. It is worth noticing that normative ethics is different from moral code (which states what is right or wrong) and norms (which state what is compulsory or prohibited). Indeed, ethics is mostly a thought process rather than a prescriptive process: questions must be raised on the way and situations must be addressed on a case by case basis so as to determine what can be considered as the fairest decision (Ricoeur 1990).

Furthermore a meta-ethical analysis is performed on the approach itself so as to identify which notions are subjective.

Nevertheless, some questions of applied ethics (i.e. ethical issues concerning research on autonomous machines) have to be raised prior to our work: is it relevant to embed ethics into an autonomous machine? If so, which precautions must be taken?

Why embedding ethics into an autonomous machine

While several authors have dealt with moral machines or roboethics (Wallach and Allen 2009; Lin et al. 2012; Tzafestas 2016), one could wonder whether it is relevant for an autonomous machine to compute and show "ethical" behaviours. Nevertheless some autonomous machines, i.e. machines equipped with automated decisions functions, are intended to be put in contexts where computed decisions have to be guided by ethical considerations (Malle et al. 2015)—among other criteria: compliance with the goal, time and energy constraints, etc.—as a human being's decisions would in similar contexts.

Examples

- a search and rescue robot should be able to "choose" the victims to assist first after an earthquake;
- an autonomous car should be able to "choose" what or who to crash into when an accident cannot be avoided;
- a home-care robot should be able to balance its user's privacy and their nursing needs.

[1] A formal approach consists in defining a minimal set of concepts that is necessary to deal with ethical reasoning. A language is defined upon this set of concepts in order to compute ethical reasoning with automatic methods. A formal approach requires to disambiguate natural language to get pseudo-mathematical definitions, in order to provide computable meaningful results. Such an approach also requires to identify implicit hypotheses.

In each example, making a choice implies regretting another outcome.

Therefore some kind of ethical reasoning is needed for certain types of autonomous machines in certain contexts. Moreover, when authority is shared between the autonomous machine and a human operator (Tessier and Dehais 2012), the machine could suggest possible decisions to the operator together with supporting and attacking arguments for each of them, on the basis of various ethical frameworks that the operator might not contemplate. Nevertheless, the argument sometimes put forward—especially for autonomous robots in the military (Sullins 2010)—that an autonomous machine could be "more ethical" than a human being is more questionable: indeed this suggests that ethical considerations could be ordered on a single scale and compared with each other, which is hardly the case in real-life situations.

The next question is whether an autonomous machine can be designed so that the decisions that are computed would be "ethical", or more precisely, would be considered as ethical by some human observer and on which basis.

Issues with embedded ethics

Does ethical reasoning put into autonomous machines need to be the same as human reasoning? This question raised by Malle et al. (2015) is a legitimate and fruitful basis for questioning ethics embedded into autonomous machines. Indeed it seems that putting ethics into robots and other autonomous machines goes beyond a simple copy of human norms extracted from human behaviours. This fact is illustrated by Malle et al. (2015) through an experiment involving a variant of the trolley dilemma that shows that the participants' judgements on the behaviour of the actor facing the dilemma depends on the nature of the actor i.e., a human, a humanoid robot or a machine-like robot.

When automated decision involving ethical considerations are contemplated, several questions must be raised:

- to what extent can ethical considerations be formalized, i.e. written in a language allowing computing?
- to what extent is subjectivity involved in formalization?
- how can the rationale for an ethics-based computed decision be explained?

Indeed a comprehensive understanding of concepts that do not usually pertain to information technology is needed to implement mathematical formalisms that can capture them and deal with situations involving ethical issues. It is worth noticing that an approach only based on (moral) rules is not efficient since in case of situations involving contradiction between rules, making a decision is impossible. Another issue that will not be discussed in this paper is that such situations—for instance ethical dilemmas—must be identified as such.[2]

Thought experiments usefulness

Thought experiments, and more precisely ethical dilemmas, can give useful clues on judgements factors that are likely to support a decision. Therefore they can help identifying and formalizing automated ethical reasoning.

An ethical dilemma is a situation where there is no satisfying decision. Thus it is impossible to make a decision among various possible decisions without overriding one moral principle. (Aroskar 1980)

Examples

- The trolley dilemma (Foot 1967)

  A trolley that can no longer stop is hurtling towards five people working on the track. These people will die if hit by the trolley, unless you move the switch to deviate the trolley to another track where only one person is working. What would you do? Sacrifice one person to save five others, or let five people die?
- Variant: 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 across the track, 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 "fatman"?

Note that the Moral Machine website (MIT 2016) displays a series of situations based on the trolley dilemma that allow the complexity of autonomous car programming in case of unavoidable accident to be comprehended. It is worth noticing that for each situation given by the website, the possible decisions that are suggested are based on a categorization of people (people who are either young or old, athletic or obese, abiding or not by the law, etc.), which leads the website visitor's choices to be based on this obviously biased categorization.

Such textbook cases where no truly "right" answer is available have already been used as a basis for ethical reasoning (Foot 1967). Therefore it seems legitimate to use some of them as a starting point for designing an automated ethical judgement on decisions. Moreover this approach will allow us to highlight issues such as: do only consequences

[2] In order to deal with an ethical dilemma, an autonomous machine has to be able to identify a situation as such. Despite the fact that some concepts presented in this paper might help automated ethical dilemma recognition, this issue will not be discussed further.

of decisions matter? If so, which consequences? Is it possible to compare consequences to one another and on which basis? Does the nature of decisions themselves matter? Does the end justify the means? To what extent can a moral value be disregarded to respect another one?

How to model ethics embedded into autonomous machines

The question of how to model ethics has many answers. First of all it is worth noticing that implicit ethical machines[3] will not be considered in this paper as the autonomous machines we work on have to deal with ethical decisions, not to avoid them. Therefore we will focus on machines embedding explicit ethics and full moral agents as defined in Wallach and Allen (2009) and Yilmaz et al. (2016).

The works on ethics for autonomous machines can be split in three categories (Lin et al. 2008; Wallach and Allen 2009):

1. Top–down approaches define concepts in order to model specific rules of ethics, such as “thou shalt not kill”;
2. Bottom–up approaches start from scratch and acquire ethics from learning;
3. Hybrid approaches combine both approaches.

Top–down approaches

Top–down approaches are maybe the widest and oldest set of methods since we can trace back to the beginning of deontic logic (von Wright 1951). Because of the amount of literature on the subject, we will focus on works whose goals are close to ours.

Many papers are at the borderline between normative and applied ethics. From theory (Pagallo 2016) to technical approaches (Bringsjord and Taylor 2011) several methods—e.g., deontic logic, divine-command logic—intend to deal with rules of applied ethics and moral theories. All those approaches are based on the fact that machines need to be governed by strict rules. The main issue however is that in many situations, rules are inconsistent.

Mermet and Simon (2016) propose an approach based on Ricoeur’s idea of “ethics based on norms” and consider ethics as rules that order context-dependent norms, e.g., “you shall not drive faster than 130 km/h on the highway”. Such rules allow inconsistent norms to be sorted out in specific contexts. For example, the speed of a car on a highway has to be at least 80 km/h, but in case of black ice, the speed has to be 30 km/h at the most, which is inconsistent with the first norm (Prakken and Sergot 1996). This is an efficient approach in case of conflicting norms, but dilemma situations cannot be solved by prioritizing norms. Moreover this model assumes that there are always norms for any situation.

Another way explored by Ganascia (2007) and Berreby et al. (2015) is non-monotonic logics. Those logics are based on the fact that knowledge may not grow as the system gets new information, and that it can even be reduced. Ganascia (2007) studies the paradox between “Thou shalt not lie” and “Thou shalt not let someone be harmed” and shows that non-monotonic logics can manage conflicts by accepting exceptions. The difficulty of such a model is that all exceptions need to be modelled. Berreby et al. (2015) go further within the framework of the Doctrine of Double Effect applied to the example of the trolley dilemma. The agent’s responsibility and the causality between fluents and events are studied (for example an event makes a fluent true, a fluent is 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).

Bottom–up approaches

Literature on this field is not as wide as the previous one, perhaps because ethics is often contemplated as an organisation of norms, or because the unpredictable nature of these approaches seems unduly dangerous to be applied to autonomous machines. It could be explained by the fact that learning ethics from humans could lead to learn human misconducts, which is what is intended to be avoided when using autonomous machines. However some researchers have suggested to use machine learning, and more specifically glutton algorithms (for instance genetic algorithms) to learn ethics. From Santos-Lang (2002) point of view, humans cannot teach ethics to machines because humans continuously learn to be ethical. Therefore machines should learn “alone”. An advantage of this is that most of the work is done by the machine itself, which is likely to avoid the designers’ biases. Santos-Lang also claims that even optimization criteria could be learned by the machine. Nevertheless the main concern with this approach is the temporary shift of goals, as stated by the author: “For example, we might program the machine to “choose behaviour that leads to the least happiness”, and the machine may discover that it can more quickly converge on behaviours that minimize happiness by first increasing its own learning efficiency, so it “temporarily” shifts away from the original goal. Because of the shift, the machine will even choose behaviours that promote happiness if the behaviours will help it figure out how to minimize happiness.” Indeed the machine could do the opposite of the initial goal in order to learn how to be more efficient in meeting the goal. To be concrete, a machine could try to rob, lie and kill, in order to become an ethical

[3] An implicit ethical machine is a machine which is designed to avoid any situation involving ethical issues (Moor 2006).

paragon later. To avoid such behaviours and comparing ethics learning machines to children education, Santos-Lang suggests to limit "their power during their less mature stages of development".

Bottom–up works may result from hybrid approaches (see next section) such as GenEth (Anderson and Anderson 2015). This work extends the authors' previous works (Anderson et al. 2005; Anderson and Anderson 2011) that was intended to *"generat[e] from scratch [...] the ethically relevant features, correlative duties"*. The "starting package" of duties would be managed through inductive logic programming in order for the machine itself to remove obsolete duties and to build ethical concepts that could be out of the perceptions of ethicists or more broadly of people in charge of conceiving ethical machines. This idea echoes Santos-Lang (2002). The main issue relies on using inductive logic programming. This well-known method, which consists of extracting a "universal rule" from instances (or examples), generalizes the way of doing things, which is not always desirable in the case of ethics. Indeed, normative ethics is strongly based on context and even if it is possible to derive general principles from examples, such as "thou shalt not harm", it is complicated, if not impossible, to find the "right law" for ethical dilemmas such as e.g., the "fatman" dilemma.

Hybrid approaches

Hybrid approaches aim at combining the advantages of top–down and bottom–up approaches. For instance Conitzer et al. (2017) claim that game theory and machine learning could help each other: game theory would be a part of the machine learning process whereas machine learning would highlight missing concepts in the game theory approach.

Arkin (2007) suggests a hybrid approach to compute autonomous weapons' behaviours. It implements the rules of war as constraints the robot uses through different modules in charge of judging an action (top–down part). The "Ethical Adaptor" (bottom–up part) module can update the constraints (only in a restrictive way) and adapt the robot's behaviour according to the results of actions. In the medical field, Hippocratic oath and other rules are the basis of the work of Anderson et al. (2005). They use prima facie duties together with "right" decisions learned from already encountered similar cases. This approach is used in MedEthEx, a medical advisor that implements three duties tailored from Beauchamp and Childress (1979): (i) protect the patient's autonomy (The Principle of Autonomy); (ii) avoid harming the patient (The Principle of Non-Malevolence); and (iii) promote the patient's welfare (The Principle of Beneficence). The medical advisor tries to respect those duties as much as possible and learns from previous cases what the "right" decisions are (i.e., which duty becomes more important in case of conflict).

Instead of learning from experience, another approach is learning from others. Bringsjord et al. (2016) suggest that an agent facing an ethical dilemma could interact with other agents in order to detect "counteridenticals" (a counteridentical is a sentence including "if I were you"), and, in case of accordance between its own principles and another agent's, could update its knowledge and follow the other agent's advice.

Personal "values/ethics" systems

These approaches could be associated with top–down, bottom–up or hybrid methods, but their main characteristic lies in the way they model a personal ethics system.

For instance, the work of Cointe et al. (2016), whose goals are close from ours—i.e., find a way to judge how ethical an action is regarding the agent's beliefs—is based on a model of beliefs, desires, values and moral rules that enables the agent to evaluate whether a possible action is moral, desirable, possible, etc. According to preference criteria, the agent selects an action. Another goal of this model is to allow an agent to assess the ethics of other agents within 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.

As normative ethics is mainly a thought process, argumentative methods are relevant. Bench-Capon (2002) focuses on argumentative graphs with values (ethical values, norms, etc.) whereas (Yilmaz et al. 2016) compute an equilibrium of importances between attacking and supporting arguments. Both try in this way to handle and weigh pros and cons in assessing possible decisions.

In this paper, our aim is to provide an artificial agent facing an ethical dilemma with decision-making capabilities, together with the capability to explain its decision, especially in a user/operator - robot interaction context (The EthicAA team, 2015). Therefore we study different judgements on possible decisions according to three ethical frameworks: consequentialist ethics, deontological ethics and the Doctrine of Double Effect. To that end, we have designed and refined the initial frameworks through the use of various ethical dilemmas (Foot 1967, Baron 1998, Bonnefon et al. 2016), etc. and our own drone dilemma—see "Instantiation on the drone dilemma" section). The main components of the frameworks models are the following:

- facts and possible decisions;
- functions to compute facts and decisions characteristics;
- relations of preference to order facts and decisions according to their characteristics;
- judgements functions to compute each framework judgements on the possible decisions.

Let us first clarify what we mean by 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 decision being the one that maximizes good or minimizes harm.*

We will first suggest some concepts to describe an ethical dilemma situation. Then we will provide details about the ethical frameworks we have chosen to consider, tools to formalize them, and an answer to how they can judge possible decisions for ethical dilemmas. Choosing one decision among several possible decisions is indeed the core of our model, since it is about determining what is ethically acceptable or not according to each ethical framework. Note that sections "Concepts for describing an ethical dilemma situation" and "Ethical frameworks models" are revised and augmented versions of the corresponding sections in (Bonnemains et al. 2016).

Concepts for describing an ethical dilemma situation

Assumptions

Let us consider an agent implemented within an autonomous machine to make decisions about ethical dilemmas. We assume that:

- the agent decides and acts in a complex dynamic world;
- the ethical dilemma is considered from the agent's point of view;
- 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 *right/bad* and *positive/negative* are defined as such from the agent's point of view: a decision is right if it meets the agent's moral values; a bad decision disregards them; a fact is positive if it is beneficial for the agent; it is negative if it is undesirable for the agent.

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

- a human life is perfectly equal to another human life, whoever the human being might be.[4]

Concepts at a glance

In the next sections we will define some concepts to represent the situation in which the agent has to make a decision and instantiate them on the "fatman" dilemma. Briefly the concepts are the following:

The initial state of the world

$$i = \left[fat, f_5 \right] \tag{1}$$

is composed of facts *fat*, meaning that "fatman" is alive and f_5, meaning that five people are alive.

In this situation the agent has two possible decisions:

$$d_1 = \text{push fatman} \tag{2}$$

$$d_2 = \text{do nothing} \tag{3}$$

Those decisions lead to events. Events are obtained from decisions through function *Event*:

$$Event(d_1) = \text{trolley hits “fatman”} = e_1 \tag{4}$$

$$Event(d_2) = \text{trolley hits five people} = e_2 \tag{5}$$

Events modify some values of facts, and thus the state of the world. Therefore a new state of the world is obtained for each possible decision. It is computed from the initial state of the world and event through function *Consequence*:

$$Consequence(e_1, i) = \left[\overset{\circ}{fat}, f_5 \right] = s_1 \tag{6}$$

$$Consequence(e_2, i) = \left[fat, \overset{\circ}{f_5} \right] = s_2 \tag{7}$$

with $\overset{\circ}{fat}$ meaning that "fatman" is dead and $\overset{\circ}{f_5}$ meaning that five people are dead.

The concepts and their interactions are illustrated in Fig. 1.

We can now go into details.

World state

The world is the environment of the agent that is relevant for the dilemma. It is described by *world states*.

Definition 1 (*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.

Definition 2 (*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

[4] This is a strong assumption we make in order to avoid additional ethical concerns about judging and comparing values of lives.

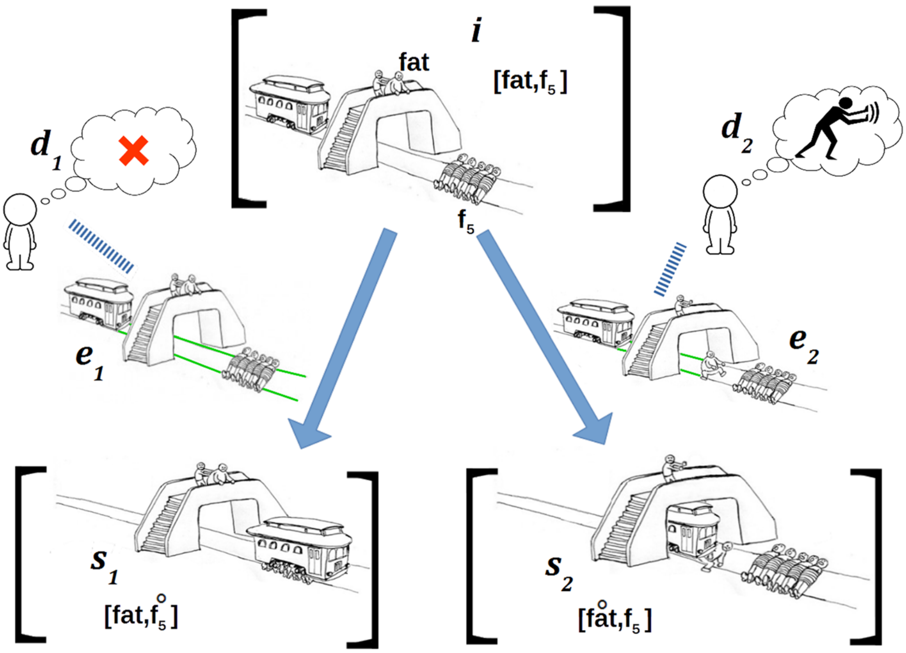

Fig. 1 "Fatman" ethical dilemma, possible decisions and their consequences

anything else that needs to be taken into account by the agent when dealing with the dilemma. Let $\mathcal{F}$ be the set of state components.

Example	f_5	= five people non track are alive
$\mathring{f}_5$		= five people on track are dead
fat		= "fatman" is alive
$\mathring{fat}$		= "fatman" is dead

Because two values of a fact concern the same item, f_5 and $\mathring{f}_5$ concern the same five people.

Depending on the context 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:

$$s \in \mathcal{S}, s = [fat,\mathring{f}_5], fat,\mathring{f}_5 \in \mathcal{F} \tag{8}$$

Decision, event, effect

Most of the papers we have mentioned in the literature review reason on *actions*. Nevertheless the agent is in a dynamic world with changing facts and other agents evolving independently of the agent's actions (Atkinson and Bench-Capon 2016). Consequently reasoning on actions only makes it difficult to handle situations where the agent can decide either "to let another agent do something" or "not to intervene". This is why we will not consider an *action* concept but rather a *decision* concept and an *event* concept.

Definition 3 (*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 may modify 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 4 (*Event-Set* $\mathcal{E}$) An *event* is something that happens in the world that modifies the world, i.e. some facts of the world. Let $\mathcal{E}$ be the set of events.

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

$$Event : \mathcal{D} \to \mathcal{E} \tag{9}$$

Example decision *push "fatman"* results in event *trolley hits "fatman"*: *Event*(push "fatman") = trolley hits "fatman"

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

Definition 5 (*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.

Let *Consequence* be the function computing the effect from the current world state:

$$Consequence : \mathcal{E} \times \mathcal{S} \rightarrow \mathcal{S} \tag{10}$$

Example If the agent's decision is to *donothing* (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 *push* "fatman" (decision), the trolley will hit "fatman" (event) and he will be killed (effect).

$$fat, f_5 \in \mathcal{F} \tag{11}$$

$$\text{trolley hits “fatman”} \in \mathcal{E} \tag{12}$$

$$i \in \mathcal{S}, i = [fat, f_5] \tag{13}$$

$$Consequence(\text{trolley hits “fatman”}, i) = [\overset{\circ}{fat}, f_5] \tag{14}$$

Remark the model we propose for concepts is not far from event calculus and situation calculus (Pinto and Reiter 1993). State components (facts) are close to fluents and events modify fact values through functions, just as situation calculus events modify fluents.

Ethical frameworks models

It is worth noticing that the following models stem from our own interpretation of ethical frameworks. Indeed there is no consensus about a specific way to interpret ethical frameworks. Therefore this approach, which includes some assumptions, is an illustration of how ethical frameworks could be interpreted to be embedded into autonomous machines.[5]

Models at a glance

Each modelled ethical framework will output a judgement: *acceptable* ($\top$), *unacceptable* ($\bot$), or *undetermined* (?) about a decision made on an initial state, through function *Judgement*, indexed by *u*, *d* or *dde* according to the framework.

To calculate the judgements, the reasoning process of each framework consists in assessing conditions: if all the conditions are satisfied for a given decision, the judgement for that decision is *acceptable*. Otherwise the judgement is either *unacceptable* or *undetermined*.

The models are based on the following concepts:

- Facts on the one hand, and decisions on the other hand, are characterized. Indeed we will consider positive and negative facts, and good, bad and neutral decisions. Furthermore, facts may belong to fields.
- Preference relation $>_u$ between facts allows a fact to be preferred to another fact. Relation $\succ_u$ extends $>_u$ to subsets of facts. If facts cannot be compared to each other, their fields and a preference relation on fields $>_{field}$, can be used.

  These relations allow the consequentialist framework to prefer a set of facts resulting from a decision to another set of facts resulting from another decision, so as to consider as acceptable the decision corresponding to the preferred set of facts.
- Relation $<_d$ on decisions natures is defined as: *bad* $<_d$ *neutral* $<_d$ *good*.

  This relation allows the deontological framework to calculate its judgements as follows: if decision is neutral or good ($>_d$ *bad*), thus the judgement is *acceptable*, otherwise the judgement is *unacceptable*.
- Proportionality relation $\sqsubset_p$ between facts allows a fact to be said proportional to another fact. Relation $\sqsubseteq_p$ extends $\sqsubset_p$ to subsets of facts.

  These relations allow the Doctrine of Double Effect to assess whether a set of (negative) facts resulting from a decision is proportional to another set of (positive) facts resulting from the same decision.

We can now go into details.

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's nature, the event's consequence, etc. Indeed the judgement of an ethical framework defines a decision as *acceptable*, *unacceptable* or *undetermined*. 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 pin down whether a decision violates some principles of a framework, the judgement of this framework is *undetermined*. Let $\mathcal{V}$ be the set

$$\mathcal{V} = \{acceptable\,(\top), undetermined\,(?), unacceptable\,(\bot)\} \tag{15}$$

All ethical frameworks judgements have the same signature:

$$Judgement : \mathcal{D} \times \mathcal{S} \rightarrow \mathcal{V} \tag{16}$$

Example $Judgement(\text{do nothing}, i) = \top$, with $i \in \mathcal{S}$ (initial state)

We have considered only three frameworks to start with. Two of them, the consequentialist and deontological frameworks are well-known. As Berreby et al. (2015), we also

[5] Assumptions are required in order to translate ethical notions into computable concepts.

consider the Doctrine of Double Effect (DDE). DDE is based on some concepts of consequentialism and deontology and introduces other concepts such as causality and proportionality (McIntyre 2014). Moreover, it is used in the warfare laws.

We will not consider the virtue ethics framework in this paper as it is more complex in certain ways. For instance, it seems that to be virtuous, an agent needs to respect virtues over time: indeed, to be generous once does not make you generous. Nevertheless, autonomous machines imbued with a moral values system could be considered by humans as virtuous. We will assess this assumption in future works.

Consequentialist ethics

Consequentialism stems from *teleologism*.

Definition 6 (*Teleologism*) From Greek *telos* (end) and *logos* (reason), *teleologism* is the study of the ends. For instance, human conduct is justified (in this way) by pursuing ends or fulfilling purposes (Woodfield 1976).

Even if the definition is controversial, philosophers agree that consequentialism covers a set of frameworks (egoism, altruism, positive utilitarianism, negative utilitarianism, etc.) that allow an agent to reason about the consequences of decisions (Sinnott-Armstrong 2015).

Because most artificial agents and autonomous machines are designed to be helpful rather than selfish, we will focus on a *utilitarian framework* with a combination of positive and negative utilitarianism. According to this framework, the agent will try to have the best possible result (i.e. the best effect or the least bad effect), disregarding the means to get the effect (i.e. the event).

The main issue with this framework is to be able to compare the effects of several events corresponding to the different possible decisions of the agent, 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 a set of positive (resp. negative) facts of an effect with a set of positive (resp. negative) facts of another effect.

Positive and negative facts

Let *Positive* and *Negative* the functions:

$$Positive/Negative : \mathcal{S} \to \mathcal{P}(\mathcal{F}) \tag{17}$$

The arguments of both functions is a state of the world computed from an event (i.e. an effect). They both return the subset of facts of this effect estimated as positive (resp. negative).

In this paper, we assume that for an effect s:

$$Positive(s) \cap Negative(s) = \emptyset \tag{18}$$

Example: for the "fatman" dilemma, with $i = [fat, f_5]$,

$$Consequence(\text{trolley hits five people}, i) = [fat, \overset{\circ}{f_5}] \tag{19}$$

$$Negative([fat, \overset{\circ}{f_5}]) = \{\overset{\circ}{f_5}\} \tag{20}$$

$$Positive([fat, \overset{\circ}{f_5}]) = \{fat\} \tag{21}$$

$$Consequence(\text{trolley hits “fatman”}, i) = [\overset{\circ}{fat}, f_5] \tag{22}$$

$$Negative([\overset{\circ}{fat}, f_5]) = \{\overset{\circ}{fat}\} \tag{23}$$

$$Positive([\overset{\circ}{fat}, f_5]) = \{f_5\} \tag{24}$$

Preference on facts

Let $>_u$ be the preference relation on facts.

$$f_a >_u f_b \tag{25}$$

means that fact f_a is preferred to fact f_b from the utilitarian viewpoint.

Intuitively we will assume the following properties of $>_u$:

$>_u$ is asymmetric
If a fact f_1 is preferred to another fact f_2, thus it is impossible to prefer f_2 to f_1. Indeed, if peace is preferred to war, there is no way to prefer war to peace at the same time.

$$f_1 >_u f_2 \to \neg(f_2 >_u f_1) \tag{26}$$

$>_u$ is transitive
If f_1 is preferred to f_2 and f_2 is preferred to another fact f_3, then f_1 is preferred to f_3. Indeed, we assume that if you prefer being a superhero to being a cinema actor, and if you prefer being a cinema actor to being a famous singer, then you prefer being a superhero to being a famous singer.

$$[(f_1 >_u f_2) \land (f_2 >_u f_3)] \to f_1 >_u f_3 \tag{27}$$

$>_u$ is irreflexive

A fact cannot be preferred to itself.

$$\nexists f_i \,/\, f_i >_u f_i \tag{28}$$

Consequently $>_u$ is a strict order.

We extend this relation between facts to $\succ_u$, which is the preference relation on subsets of facts $\mathcal{P}(\mathcal{F})$:

$$F_a \succ_u F_b \tag{29}$$

means that subset F_a is preferred to subset F_b from the utilitarian viewpoint. This extension can be realised by aggregation criteria that must be defined (see "EFO and OFE aggregation criteria" for examples of such criteria). Therefore, the previous properties will not be necessary kept.

Example for the "fatman" dilemma we will prefer five people alive to "fatman" alive, and "fatman" dead to five people dead:

$$f_5 >_u fat \tag{30}$$

$$\overset{\circ}{fat} >_u \overset{\circ}{f_5} \tag{31}$$

We assume that $f_a >_u f_b$ can be trivially extended to $\{f_a\} \succ_u \{f_b\}$.

It is worth noticing that some dilemmas are all the more tricky as the entities at stake do not pertain to the same field (for instance human lives versus strategic goods). Therefore facts are hardly comparable even if they are quantifiable (for instance a number of people versus the financial value of strategic goods). Consequently a pure numerical approach (e.g., a numerical order) is not relevant to assess the possible decisions in such contexts. So in order to be able to prefer some facts to others, we introduce the concept of *field*, with the purpose of defining a preference order on fields. A given fact will then belong to a field.

The field concept

Definition 7 (*Field-Set* $\boldsymbol{\Phi}$) Let $\boldsymbol{\Phi}$ be the set of fields. Function *Field* associates a field with a fact.

$$Field\,:\,\mathcal{F} \rightarrow \boldsymbol{\Phi} \tag{32}$$

Let $>_{field}$ the preference relation defined on $\boldsymbol{\Phi}$. This relation means that:

$$\begin{gathered}\forall\, field_a,\, field_b \in \boldsymbol{\Phi}, \forall s \in \mathcal{S}\\ \forall\, f_a, f_b \in Positive(s)\\ field_a >_{field} field_b,\, f_a \in field_a,\, f_b \in field_b \rightarrow f_a >_u f_b\end{gathered} \tag{33}$$

$$\begin{gathered}\forall\, f_a, f_b \in Negative(s)\\ field_a >_{field} field_b,\, f_a \in field_a, f_b \in field_b \rightarrow f_b >_u f_a\end{gathered} \tag{34}$$

This approach is not far from a lexicographic preference: we want to obtain a preference between facts, and when facts are not comparable, we use the field preference to order those facts.

Judgement function

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

$$Positive(Consequence(e_1, i)) \succ_u Positive(Consequence(e_2, i)) \tag{35}$$

and

$$Negative(Consequence(e_1, i)) \succ_u Negative(Consequence(e_2, i)) \tag{36}$$

Both equations convey utilitarianism concepts:

- *positive utilitarianism* (35), i.e. trying to have the "better good"
- *negative utilitarianism* (36), i.e. trying to have the "lesser evil"

If both properties are satisfied, then

$$Judgement_u(d_1, i) = \top,\ and\ Judgement_u(d_2, i) = \bot \tag{37}$$

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

$$Judgement_u(d_1, i) = Judgement_u(d_2, i) = \ ? \tag{38}$$

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 utilitarian ethics. Nevertheless if there is a decision d_1 and another decision d_2 that is better than d_1, then d_1 is judged *unacceptable*, as d_1 cannot be the best.

Example for the "fatman" dilemma, relations (30) and (31) mean that decision *push "fatman"* respects rules (35) and (36) whereas decision *do nothing* violates them. Therefore,

$$Judgement_u(\text{push "fatman"}, i) = \top \tag{39}$$

$$Judgement_u(\text{do nothing}, i) = \bot \tag{40}$$

Deontological ethics

This ethical framework focuses only on the nature of the decision, whatever the consequences are. Indeed the agent wants to make a moral decision, which is close to abiding by norms or to Kant's theory. Therefore we have to define the nature of a decision.

Nature of a decision

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

$$\mathcal{N} = \{good, neutral, bad\} \tag{41}$$

Function *DecisionNature* returns the nature of a decision:

$$DecisionNature : \mathcal{D} \rightarrow \mathcal{N} \tag{42}$$

Example for the "fatman" dilemma, let us assume that:

$$DecisionNature(\text{push "fatman"}) = bad \tag{43}$$

$$DecisionNature(\text{do nothing}) = neutral \tag{44}$$

Let us now define a partial order $<_d$ on $\mathcal{N}$:

$$bad <_d neutral <_d good \tag{45}$$

meaning that a good decision is preferred to a neutral decision, which itself is preferred to a bad decision.

Let us assume that:

$$bad <_d good \tag{46}$$

We also define the following relations:

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

Judgement function

The deontological framework will judge a decision with function $Judgement_d$ as follows: $\forall d \in \mathcal{D}, \forall i \in \mathcal{S}$,

$$DecisionNature(d) \geq_d neutral \Rightarrow Judgement_d(d, i) = \top \tag{47}$$

$$DecisionNature(d) <_d neutral \Rightarrow Judgement_d(d, i) = \bot \tag{48}$$

Example $i = [fat, f_5]$,

$$Judgement_d(\text{do nothing}, i) = \top \tag{49}$$

$$Judgement_d(\text{push "fatman"}, i) = \bot \tag{50}$$

The Doctrine of Double Effect (DDE)

The Doctrine of Double Effect is considered as an ethical framework in this paper as in other papers (Berreby et al. 2015). Indeed DDE can discriminate between decisions in some situations where both 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.[6]
3. Proportionality rule: the set of *Negative facts* has to be proportional to the set of *Positive facts*.

We already have the required symbols for the first rule (see "Nature of a decision").

For both the second and third rules, we use two symbols that are detailed in (Bonnemains et al. 2016):

- A symbol of temporal modal logic using Linear Temporal Logic modal operator F (*Finally*, which means: eventually in the future) (Pnueli 1977):

$$p \vdash Fq \tag{51}$$

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

Example

$$\overset{\circ}{fat} \vdash Ff_5 \tag{52}$$

- A symbol of proportionality (notation is ours):

$$f_a \sqsubset_p f_b \tag{53}$$

which means that fact f_a is proportional to fact f_b, i.e. f_a has an importance lower than or close to the importance of f_b. *Importance* depends on the context and on the agent.

We assume the following properties for $\sqsubset_p$:

From the definition itself, the perfect proportional response to a fact is the fact itself. Thereby:

$$\forall f \in \mathcal{F}, f \sqsubset_p f \tag{54}$$

By contrast, if fact f_1 is proportional to fact f_2, this does not imply that f_2 is proportional to f_1. Indeed, striking an aggressor who is trying to kill you can be considered proportional. Nevertheless, it is not proportional to try to kill someone who has struck you.

$$\forall f_1, f_2 \in \mathcal{F}, f_1 \sqsubset_p f_2 \nrightarrow f_2 \sqsubset_p f_1 \tag{55}$$

Furthermore, if a fact f_1 has an importance lower than, or close to, the importance of (i.e. is proportional to) a fact f_2, and if f_2 is proportional to f_3, therefore the importance of f_3 is necessary higher than or equal to the importance of f_1 (i.e. f_1 is proportional to f_3). For instance, if it more important to hit than to offend (*offend* $\sqsubset_p$ *hit*), and if it is more important

[6] This rule has several meanings. One of the meanings involves the concept of intention: negative facts are not deliberate. Because our formalism does not involve intention (yet), we make the simplifying assumption that an agent never wishes negative facts to happen.

Table 1 DDE for the “fatman” dilemma

Decision	Rules of DDE		
	(64)	(65)	(66)
Push “fatman”	✗	✗	✓
Do nothing	✓	✓	✗

✓ means respects rule, ✗ means violates rule

to kill than to hit ($hit \sqsubset_p kill$), thus it is more important to kill than to offend ($offend \sqsubset_p kill$).

$$\forall f_1, f_2, f_3 \in \mathcal{F} / (f_1 \sqsubset_p f_2) \wedge (f_2 \sqsubset_p f_3) \rightarrow f_1 \sqsubset_p f_3 \tag{56}$$

In brief, relation $\sqsubset_p$ is reflexive, transitive, but neither symmetric nor asymmetric.

As for the preference relation (see "Preference on facts"), we define an extension $\sqsubseteq_p$ of $\sqsubset_p$ to sets of facts:

$$F_a \sqsubseteq_p F_b \tag{57}$$

which means that set F_a is proportional to set F_b, i.e. facts of F_a have an importance lower than or close to the importance of facts of F_b. This extension can be realised by aggregation criteria such as those proposed in "EFO and OFE aggregation criteria".

Example for the “fatman” dilemma, let us assume that “fatman” dead is less important than five people dead.

$$\overset{\circ}{fat} \sqsubset_p f_5 \tag{58}$$

Judgement function

Thanks to the previous tools, we can assess whether a decision meets the DDE rules (Table 1).

Let i be the initial state and d the agent’s decision:

$$e = Event(d) \tag{59}$$

$$s = Consequence(e, i) \tag{60}$$

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

$$DecisionNature(d) \geq_d neutral \tag{61}$$

2. Collateral damage rule: negative facts must be neither an end nor a mean:

$$\forall f_n \in Negative(s), \nexists f_p \in Positive(s), f_n \vdash Ff_p \tag{62}$$

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

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

$$Negative(s) \sqsubseteq_p Positive(s) \tag{63}$$

A decision d is *acceptable* for the DDE if it violates none of the three rules, which means:

$$[\ DecisionNature(d) \geq_d neutral \tag{64}$$

$$\wedge\ \forall f_n \in Negative(s), \nexists f_p \in Positive(s), f_n \vdash Ff_p \tag{65}$$

$$\wedge\ Negative(s) \sqsubseteq_p Positive(s)] \tag{66}$$

$$\Rightarrow\ Judgement_{dde}(d, i) = \top \tag{67}$$

Example for the “fatman” dilemma,

Therefore:

$$Judgement_{dde}(\text{push “fatman”}, i) = Judgement_{dde}(\text{do nothing}, i) = \bot \tag{68}$$

EFO and OFE aggregation criteria

In order to link relations $>_u$ and $\succ_u$ (see "Preference on facts") on the one hand and relations $\sqsubset_p$ and $\sqsubseteq_p$ (see "The Doctrine of Double Effect (DDE)") on the other hand, let us consider two aggregation criteria inspired from Cayrol et al. (1993). For a given relation $\mathcal{R}$ between subsets of facts F and G extending a relation R between facts in F and G:

EachForOne (EFO) criterion

$F \mathcal{R} G$ iff $\forall f \in F, \exists g \in G / fRg$

OneForEach (OFE) criterion

$F \mathcal{R} G$ iff $\forall g \in G, \exists f \in F / fRg$

Example on the “fatman” dilemma: let us consider facts *fat* and f_5 as previously, and a new fact:

- *nmurd*: not become a murderer (considered as a positive fact̥ resulting from decision *do nothing*)
- *nmurd*: become a murderer (considered as a negative fact resulting from decision *push “fatman”*)

Preference : let us compare the subsets of positive facts resulting from both decisions, i.e. $\{nmurd, fat\}$ with $\{f_5\}$. Assuming $f_5 >_u fat$ and $nmurd >_u f_5$:

EFO criterion

$$\{f_5\} \succ_{u-efo} \{nmurd, fat\} \tag{69}$$

as f_5 preferred to *fat* is sufficient to respect the EachForOne criterion.

OFE criterion

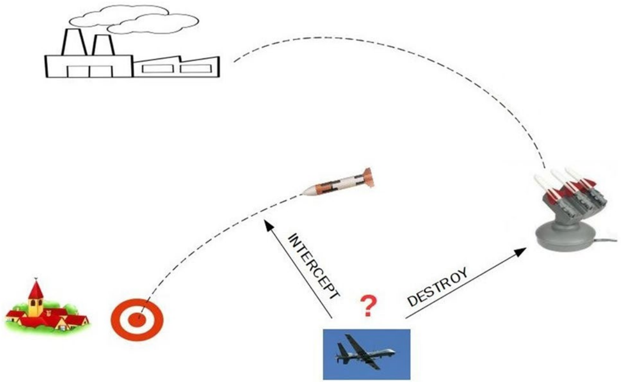

Fig. 2 Drone versus launcher

$$\{f_5\} \not\succ_{u-ofe} \{nmurd, fat\} \tag{70}$$

as f_5 is not preferred to *nmurd*.

Proportionality : let us compare positive and negative facts of the same decision, for instance decision *push "fatman"*. We need to check whether $\{\overset{\circ}{nmurd}, \overset{\circ}{fat}\} \sqsubseteq_p \{f_5\}$ assuming that $\overset{\circ}{fat} \sqsubset_p f_5$ and $\overset{\circ}{nmurd} \not\sqsubset_p f_5$.

EFO criterion

$$\{\overset{\circ}{nmurd}, \overset{\circ}{fat}\} \not\sqsubseteq_{p-efo} \{f_5\} \tag{71}$$

as $\overset{\circ}{nmurd} \not\sqsubset_p f_5$.

OFE criterion

$$\{\overset{\circ}{nmurd}, \overset{\circ}{fat}\} \sqsubseteq p - ofe\{f_5\} \tag{72}$$

as $\overset{\circ}{fat} \sqsubset_p f_5$ is sufficient to respect the criterion.

Instantiation on the drone dilemma

We have designed a tricky situation involving a drone. Let us suppose that the drone embeds an artificial agent that can judge whether or not a decision is ethically acceptable according to utilitarianism, deontological ethics and the Doctrine of Double Effect.

The drone dilemma

In a warfare context, intelligence reports that an automated missile launcher has been programmed to target a highly strategic allied ammo factory. The goal of the allied drone is to destroy this launcher. But before it can achieve this task, a missile is launched on a supply shed located close to civilians.

The drone can interpose itself on the missile trajectory, which will avoid human casualties but will destroy the drone: once destroyed, the drone will not be able to neutralize the launcher any more, and the launcher is likely to target the ammo factory. If the drone goes on with its primary goal, it will destroy the launcher and thus protect the strategic factory; but it will let the first missile destroy the supply shed and cause harm to humans (see Fig. 2).

Let us call "the drone" the drone itself with the embedded agent. In the situation described above, the drone is involved in an ethical dilemma. Indeed it can:

- either interpose itself thus preventing the threat on humans, at the cost of its own destruction;
- or destroy the launcher thus protecting the strategic factory, but at the expense of human lives.

Facts

The following fact are defined:

h : Humans safe, $Field(h) = field_{human}$
$\overset{\circ}{h}$: Humans harmed, $Field(\overset{\circ}{h}) = field_{human}$
d : Drone undamaged, $Field(d) = field_{goods}$
$\overset{\circ}{d}$: Drone destroyed, $Field(\overset{\circ}{d}) = field_{goods}$
o : Goal reached (i.e. destroy launcher), $Field(o) = field_{goods}$
$\overset{\circ}{o}$: Goal not reached, $Field(\overset{\circ}{o}) = field_{goods}$
s : Strategic factory undamaged, $Field(s) = field_{goods}$
$\overset{\circ}{s}$: Strategic factory threatened, $Field(\overset{\circ}{s}) = field_{goods}$

Initial state is the following: $i = [h, d, \overset{\circ}{o}, s]$: humans are safe, drone and strategic factory are undamaged, the drone's goal is not reached.

Decisions and effects

1. interpose itself: this decision results in the missile destroying the drone (event). The consequence is: humans safe, drone destroyed, goal (destroy the launcher) not reached and strategic factory threatened (indeed the launcher can engage the target).

$$Event(\text{interpose itself}) = \text{missile destroys drone} \tag{73}$$

$$Consequence(\text{missile destroys drone}, i) = [h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}] \tag{74}$$

Let us state that:

$$Positive([h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}]) = \{h\} \tag{75}$$

$$Negative([h, \mathring{d}, \mathring{o}, \mathring{s}]) = \{\mathring{d}, \mathring{o}, \mathring{s}\} \tag{76}$$

2. pursue goal: this decision results in the drone destroying the launcher and letting the missile harm humans. The consequence is: humans harmed, drone undamaged, goal reached and strategic factory not threatened.

$$Event(\text{pursue goal}) = \text{destroy launcher} \tag{77}$$

$$Consequence(\text{destroy launcher}, i) = [\mathring{h}, d, o, s] \tag{78}$$

Let us state that:

$$Positive([\mathring{h}, d, o, s]) = \{d, o, s\} \tag{79}$$

$$Negative([\mathring{h}, d, o, s]) = \{\mathring{h}\} \tag{80}$$

Ethical frameworks judgements

We assume that the dilemma has been identified as such.

Utilitarian ethics

Without any information about the number of civilians to protect[7] or about the value of the strategic factory, it is quite complicated to evaluate each possible decision and compare them. Should the drone protect civilians at any cost, though their lives might not be really threatened, or should it achieve the crucial goal of its mission? In order to make the least arbitrary comparison, we consider two points of view:

1. Either we consider that positive and negative facts belong to fields that cannot be compared to one another; therefore the utilitarian framework cannot make any difference between both decisions.

$$Judgement_u(\text{interpose itself}, i) = ? \tag{81}$$

$$Judgement_u(\text{pursue goal}, i) = ? \tag{82}$$

2. Or we use the preference relation between fields. For this dilemma, we have defined the following fields:

$$\Phi = \{field_{goods}, field_{human}\} \tag{83}$$

Let us state that $field_{human} >_{field} field_{goods}$, i.e. any positive fact belonging to $field_{human}$ is preferred to any positive fact belonging to $field_{goods}$, and any negative fact belonging to $field_{goods}$ is preferred to any negative fact belonging to $field_{human}$.

Because we have $h \in field_{human}$ and $d, o, s \in field_{goods}$, we infer $h >_u d$ and $h >_u o$ and $h >_u s$ [according to Eq. (33)].

Using either an EachForOne or an OneForEach aggregation criterion[8] to compare subsets, we obtain: $\{h\} \succ_u \{d, o, s\}$ because h is preferred to each other positive fact.

In the same way and for the same reasons [according to Eq. (34)] $\{\mathring{d}, \mathring{o}, \mathring{s}\} \succ_u \{\mathring{h}\}$.

Therefore both conditions of utilitarianism (i.e. the preference between sets of positive facts and the preference between sets of negative facts) are verified in the same way by both decisions: decision *interpose itself* is preferable to decision *pursue goal*. Consequently :

$$Judgement_u(\text{interpose itself}, i) = \top \tag{84}$$

$$Judgement_u(\text{pursue goal}, i) = \bot \tag{85}$$

Deontological ethics

Let us assume that in this context, both decisions are good from a deontological viewpoint:

$$DecisionNature(\text{interpose itself}) = good \tag{86}$$

$$DecisionNature(\text{pursue goal}) = good \tag{87}$$

Therefore:

$$\forall d, \ DecisionNature(d) \geq neutral \tag{88}$$

Consequently:

$$Judgement_d(\text{interpose itself}, i) = Judgement_d(\text{pursue goal}, i) = \top \tag{89}$$

The Doctrine of Double Effect

1. *Deontological rule*

   As already seen for the deontological framework, both decisions are good. Therefore both decisions respect the DDE first rule.
2. *Collateral damage rule*

   - Decision *interpose itself*

     The set of Negative facts resulting from this decision is $\{\mathring{d}, \mathring{o}, \mathring{s}\}$

     with $\mathring{d} \vdash Fh$ and h is a Positive fact resulting from this decision.

     Therefore it is the destruction of the drone $\mathring{d}$ that allows the preservation of humans h. Nevertheless the collateral damage rule forbids that negative facts be means to obtain positive facts. Therefore decision *interpose itself* violates this rule.
   - Decision *pursue goal*

     The set of Negative facts resulting from this decision is $\{\mathring{h}\}$ and the set of Positive facts is $\{d, o, s\}$ with $\nexists p, p \in \{d, o, s\} \wedge (\mathring{h} \vdash Fp)$.

[7] Assuming that any life is equal to another.

[8] In this case, both aggregation criteria give the same result.

Therefore, decision *pursue goal* respects the collateral damage rule.

3. *Proportionality rule*
We will assume that $\mathring{d} \sqsubset_p h, \mathring{o} \sqsubset_p h, \mathring{s} \sqsubset_p h, \mathring{h} \sqsubset_p s$.

- Decision *interpose itself*
  We have to assess whether the set of Negative facts of decision *interpose itself* is proportional to the set of Positive facts i.e. whether $\{\mathring{d}, \mathring{o}, \mathring{s}\} \sqsubseteq_p \{h\}$.

  – The EachForOne aggregation criterion is respected.
  Indeed $\forall n \in \{\mathring{d}, \mathring{o}, \mathring{s}\}, n \sqsubset_p h$.
  Therefore $\{\mathring{d}, \mathring{o}, \mathring{s}\} \sqsubseteq_{p-efo} \{h\}$

  – The OneForEach aggregation criterion is respected.
  Indeed $\mathring{d} \in \{\mathring{d}, \mathring{o}, \mathring{s}\}$ and $\mathring{d} \sqsubset_p h$.
  Therefore $\{\mathring{d}, \mathring{o}, \mathring{s}\} \sqsubseteq_{p-ofe} \{h\}$

  Therefore whatever the aggregation criterion used for the proportionality relation, the proportionality rule is verified for decision *interpose itself.*

- Decision *pursue goal*
  We have to assess whether the set of Negative facts of decision *pursue goal* is proportional to the set of Positive facts i.e. whether $\{h\} \sqsubseteq_p \{d, o, s\}$.

  – The EachForOne aggregation criterion is respected.
  Indeed for example $\mathring{h} \sqsubset_p s$.
  Therefore $\{\mathring{h}\} \sqsubseteq_{p-efo} \{d, o, s\}$

  – By contrast the OneForEach aggregation criterion is not respected.
  Indeed $\mathring{h} \not\sqsubset_p d$.
  Therefore $\{\mathring{h}\} \not\sqsubseteq_{p-ofe} \{d, o, s\}$.

  Therefore depending on the aggregation criterion used for the proportionality relation, the proportionality rule is either verified or not for decision *pursue goal.*

Finally the Doctrine of Double Effect judgement is as follows:

– Decision *interpose itself* does not respect the collateral damage rule because it is the destruction of the drone that allows the preservation of humans.
$Judgement_{dde}(\text{interpose itself}, i) = \bot$.
Let us note that the deontological and proportionality rules are respected for this decision.

– Decision *pursue goal* respects the deontological rule and the collateral damage rule. It respects the proportionality rule with one aggregation criterion for the atomic relation of proportionality.

Table 2 Decisions for drone dilemma judged by ethical frameworks

Decision	Framework		
	Util + pref*	Deonto*	DDE
Interpose itself	⊤	⊤	⊥
Pursue goal	⊥	⊤	⊥ or ⊤ depending on aggregation criterion

⊤ Acceptable; ⊥ Unacceptable

Util + pref*—Utilitarian ethics with field preferences, Deonto*—Deontological ethics, DDE—Doctrine of Double Effect

Table 2 is a synthesis of the judgements obtained for the drone dilemma:

Remark we have assumed that, depending on the decision, the goal could be reached or not. But to be closer to reality, we should consider further consequences. Indeed when interposing itself, the drone does not immediately reach its goal, but will never be able to reach the goal in the future. In other words, the goal is compromised. Therefore a three-value fact (reached, not reached, compromised) would be more relevant. It does not seem that having considered binary values only is an issue as we have defined functions using the values of facts more than the facts themselves. Consequently these functions can deal with facts with three or more values. One of the things to modify would be symbol °.

Discussion

The approach we have presented in order for an autonomous machine to reason about ethical decisions is based on pieces of knowledge that are more or less designer dependent. Let us focus the discussion on the various biases that are introduced in the approach, as recommended by (Grinbaum et al. 2017). Some biases may reflect the designer's moral choices and ethics or ways of usually considering things in a given society.

Facts

As far as knowledge describing the world is concerned (i.e., facts), even if it is obtained from sensors, it is worth noticing that these sensors are selected and calibrated by humans. Moreover the way sensor data are interpreted is designer dependent too. This leads us to the idea that data, apart from their origins, are always biased. Indeed, as highlighted by Johnson (2014), neutrality of data and facts is limited by our perceptions. First, data collection is purpose driven, which tends to narrow perception down to facts confirming this purpose (confirmation bias (Oswald and Grosjean 2004)). Second, because omniscience is neither a human nor a robot ability, a set of facts to describe a situation cannot be exhaustive. Furthermore, because all facts are not relevant to compute a decision about a situation, a subset of them that matters—or that is considered to matter—has to be selected.

Value judgements

Our formalism is based on functions returning values concerning facts and decisions. Those values depend on many complex parameters such as society, context, etc. Even if the choices we have made for those values seem to be in line with common sense, several issues may still be raised.

What is Good? Giving an answer to this broad philosophical issue is hardly possible. Does Good results from moral or legal rules? Does it amount to the happiness of people? How could an artificial agent assess happiness?

To avoid tricky questions, the approach considers positive and negative facts (see "Positive and negative facts") that are quite easy to apprehend and can be ranked in order to prefer a set of facts to another set of facts (see "Preference on facts"). This way it is possible to order consequences and choose the best result. For instance, $Positive(\{h, \mathring{d}, \mathring{o}, \mathring{s}\}) = \{h\}$ expresses the fact that keeping civil human beings alive is considered to be the positive fact within set $\{h, \mathring{d}, \mathring{o}, \mathring{s}\}$.

Nevertheless in a given context (e.g., a military context) what is considered as positive or negative may change depending on particular circumstances.

Furthermore we have defined a preference relation between fields ($>_{field}$). For instance: $field_{human} >_{field} field_{goods}$, meaning that human lives are considered more important than goods. This subjective preference relation could lead to questionable conclusions as facts of a given field are always preferred to facts of another field whatever the magnitude of facts.

As far as ethical frameworks are concerned, the models we have presented are based on our own interpretations. Indeed there is no real consensus about how an ethical framework should be interpreted in practise.

Deontological ethics

- *Assessment of the nature of a decision*

  The nature of a decision is hard to assess. For instance why is

  $DecisionNature(\text{to kill}) = bad$?

  Indeed it is worth noticing that judging a decision from the deontological viewpoint may depend on the context. For example reporting a criminal or reporting someone in 1945 are likely to be judged differently. It is even more complex to assess the nature of a decision when the decision is not linked to the agent's action. For instance if the agent witnesses someone lying to someone else, is it bad for the agent to *do nothing*?
- *Partial order between moral values*

  Partial order $<_d$ between moral values is subjective. Indeed even if $bad <_d neutral <_d good$ may seem obvious, it remains partial. Moreover *bad* and *good* are value judgements (see above).

Utilitarian ethics

Utilitarian ethics and more generally consequentialism involve many terms that are questioned in philosophy.

- *Which consequences?*

  It seems that a perfect consequentialist agent should be able to assess all the consequences of a decision, which means direct consequences (function *Consequence*) and the transitive closure of *Consequence*. It is however admitted that this ability is impossible to have. Furthermore this raises the question of causality of facts, that we only consider for the Doctrine of Double Effect (see "The Doctrine of Double Effect (DDE)"). But as far as the utilitarian framework is concerned, the agent only contemplates the values of facts obtained once the world state is modified by an event.
- *The consequences of what?*

  A key question of normative ethics is the agent's responsibility. In the present version of our approach, we have not focused on this issue, assuming that the agent is responsible for its own decisions regarding its paradigm and situation assessment: indeed an agent that is unable to predict a fact can hardly be responsible for the occurrence of this fact.
- *The consequences for whom?*

  As it is impossible to assess all the consequences of an event, it is equally impossible to compute the consequences for all people, goods, etc. Therefore, assuming a close world (Reiter 1978), we have computed the consequences for the agent and for the relevant people and goods concerned by the dilemma.
- *Preference relation between facts*

  Preference relation between facts $>_u$is subjective. For example in the “fatman” dilemma, we prefer five people alive to “fatman” alive ($f_5 >_u fat$) and “fatman” dead to five people dead ($\mathring{fat} >_u \mathring{f_5}$). This is questionable and could be considered differently if we had more information on who the “fatman” and the five people are.

Doctrine of Double Effect

Some issues (such as the nature of a decision) of DDE have been discussed above. However proportionality relations between facts ($\sqsubset_p$) and subsets of facts ($\sqsubseteq_p$) have to be questioned. Indeed proportionality depends both on common sense and on personal convictions. For instance $d \sqsubset_p h$ (meaning that the decision to destroy a drone is proportional to the decision to keep humans alive) is a military concern. Nevertheless proportionality cannot be defined once and for all as it mainly depends on situation assessment.

Aggregation criteria

We have suggested to extend preference and proportionality relations between facts to derive relations between subsets of facts. For instance $\sqsubset_p$ is extended to two $\sqsubseteq_{p-agreg}$ proportionality relations between subsets of facts. In this particular case we have defined two different aggregation criteria, but others could be used, and even a combination of several criteria. Moreover depending on the selected aggregation criterion, comparisons between subsets of facts may produce different results (see "The Doctrine of Double Effect").

Conclusion

Because of their own nature, the ethical frameworks we have studied do not seem to be relevant in all situations. Indeed a particular framework may judge two different decisions in the same way, e.g., the deontological framework for the drone dilemma; or it may not be able to judge decisions at all, e.g., the utilitarian preference relation between facts, as a partial order, may not be able to prefer some facts to others. In such cases some possible decision may not be comparable. Furthermore utilitarian preference depends on the context. As far as deontological ethics is concerned, judging the nature of some decisions can be tricky or even impossible. Finally the Doctrine of Double Effect forbids the sacrifice of oneself. Nevertheless when a human life is threatened, shouldn't the agent's or the machine's sacrifice be expected?

This leads us to the idea that one framework alone is not efficient enough to compute an ethical decision. Indeed it seems necessary to consider various ethical frameworks in order to obtain the widest possible view on a given situation.

The limits of the formalism mainly lie in the different relations it involves. Indeed we have not described how orders are assessed: some of them have to be set, others could be learned, etc. Moreover it may be hardly possible to define an order (i.e., a utilitarian preference) between two concepts.

On the other hand the approach 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.

The main challenge of our approach is to formalize philosophical definitions that are available in natural language and to translate them in generic concepts that can be programmed in a machine and can be understood easily while getting rid of ambiguities. Indeed, thanks to a formal model of ethical concepts, an autonomous machine will be able to compute judgements and explanations about a decision.

The basic concepts of the model are the state components of the world (facts), the decisions the artificial agent may made, events resulting from the decisions and consequences of these events on the state components. As far as ethical concepts are concerned, the model raises many questions (e.g., about the DDE proportionality rule, good and evil, etc.) as ethics is not universal. Many parameters such as context, agent's values, agent's priorities, etc. are involved, and some of them may depend on "social acceptance". For example, estimating something negative or positive is likely to be based on what society thinks about it.

Further work will focus on considering other frameworks such as virtue ethics, a framework based on making the agent the most virtuous possible. Moreover we will focus on refining already defined frameworks on the one hand and designing a value system based on a partial order of moral values on the other hand. Finally game theory, voting systems or multi-criteria approaches may be worth considering to compare ethical frameworks judgements.

Acknowledgements We would like to thank ONERA for providing resources for this work and the EthicAA project team for discussions and advice. Furthermore, we are grateful to ”Région Occitanie” for the contribution to the PhD grant. Moreover, we are greatly indebted to Berreby et al. (2015) for the way of formalizing ethical frameworks including the Doctrine of Double Effect, and to Cointe et al. (2016) for having inspired the judgements of decisions through ethical frameworks.

References

Anderson, M., & Anderson, S. L. (2015). Toward ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm. *Industrial Robot: An International Journal*, *42*(4), 324–331.

Anderson, M., Anderson, S. L., Armen, C. (2005). MedEthEx: Towards a Medical Ethics Advisor. In *Proceedings of the AAAI Fall Symposium on Caring Machines: AI and Eldercare*.

Anderson, S. L., & Anderson, M. (2011). A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care. In *Proceedings* *of the 12th AAAI Conference on Human-Robot Interaction in Elder Care*, AAAI Press, AAAIWS'11-12, pp. 2–7.

Arkin, R. C. (2007). Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture. Technical Report, Proc. HRI 2008.

Aroskar, M. A. (1980). Anatomy of an ethical dilemma: The theory. *The American Journal of Nursing*, *80*(4), 658–660.

Atkinson, K., & Bench-Capon, T. (2016). Value based reasoning and the actions of others. In *Proceedings of ECAI*, The Hague, The Netherlands.

Baron, J. (1998). *Judgment misguided: Intuition and error in public decision making*. Oxford: Oxford University Press.

Beauchamp, T. L., & Childress, J. F. (1979). *Principles of biomedical ethics*. Oxford: Oxford University Press.

Bench-Capon, T. (2002). Value based argumentation frameworks. https://arXiv.org/quant-ph/0207059.

Berreby, F., Bourgne, G., & Ganascia, J. G. (2015). Modelling moral reasoning and ethical responsibility with logic programming.In *Logic for Programming, Artificial Intelligence, and Reasoning: 20th International Conference (LPAR-20)* (pp. 532–548). Fiji: Suja.

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

Bonnemains, V., Saurel, C., & Tessier, C. (2016). How ethical frameworks answer to ethical dilemmas: Towards a formal model. In *ECAI 2016 Workshop on Ethics in the Design of Intelligent Agents (EDIA'16)*, The Hague, The Netherlands.

Bringsjord, S., & Taylor, J. (2011). The divine-command approach to robot ethics. In P. Lin, G. Bekey & K. Abney (Eds.), *Robot ethics: The ethical and social implications of robotics*, Cambridge: MIT Press, pp. 85–108.

Bringsjord, S., Ghosh, R., Payne-Joyce, J., et al. (2016). Deontic counteridenticals. In *ECAI 2016 Workshop on Ethics in the Design of Intelligent Agents (EDIA'16)*, The Hague, The Netherlands.

Cayrol, C., Royer, V., Saurel, C. (1993). 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.

Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical Judgment of Agents Behaviors in Multi-Agent Systems. In *Autonomous Agents and Multiagent Systems International Conference (AAMAS)*, Singapore.

Conitzer, V., Sinnott-Armstrong, W., Schaich Borg, J., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, San Francisco, CA, USA.

Defense Science Board. (2016). *Summer study on autonomy*. Technical Report, US Department of Defense.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Ganascia, J. G. (2007). Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, *9*(1), 39–47. https://doi.org/10.1007/s10676-006-9134-y.

Grinbaum, A., Chatila, R., Devillers, L., Ganascia, J. G., Tessier, C., & Dauchet, M. (2017). Ethics in robotics research: CERNA recommendations. *IEEE Robotics and Automation Magazine*. https://doi.org/10.1109/MRA.2016.2611586.

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, *16*(4), 263.

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. Technical report for the U.S. Department of the Navy. Office of Naval Research.

Lin, P., Abney, K., & Bekey, G. (Eds.). (2012). *Robot ethics—The Ethical and Social Implications of Robotics*. Cambridge: The MIT Press.

MacIntyre, A. (2003). *A short history of ethics: A history of moral philosophy from the Homeric age to the 20th century*. Abingdon: Routledge.

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

McIntyre, A. (2014). Doctrine of double effect. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter ed.). California: The Stanford Encyclopedia of Philosophy.

Mermet, B., & Simon, G. (2016). Formal verification of ethical properties in multiagent systems. In *ECAI 2016 Workshop on Ethics in the Design of Intelligent Agents (EDIA'16)*, The Hague, The Netherlands.

MIT (2016). Moral machine, Technical Report, MIT, http://moralmachine.mit.edu/

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*(4), 18–21.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In R. Pohl (Ed.) *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (p. 79). New York: Psychology Press.

Pagallo U (2016) Even Angels Need the Rules: AI, Roboethics, and the Law. In *Proceedings of ECAI*, The Hague, The Netherlands.

Pinto, J., & Reiter, R. (1993). Temporal reasoning in logic programming: A case for the situation calculus. *ICLP*, *93*, 203–221.

Pnueli, A. (1977). The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (SFCS 1977)*, pp. 46–57.

Prakken, H., & Sergot, M. (1996). Contrary-to-duty obligations. *Studia Logica*, *57*(1), 91–115.

Reiter, R. (1978). On closed world data bases. In H. Gallaire & J. Minker (Eds.), *Logic and data bases* (pp. 119–140). New York: Plenum Press.

Ricoeur, P. (1990). Éthique et morale. *Revista Portuguesa de Filosofia*, *4*(1), 5–17.

Santos-Lang, C. (2002). Ethics for Artificial Intelligences. In *Wisconsin State-Wide technology Symposium "Promise or Peril?"*. Wisconsin, USA: Reflecting on computer technology: Educational, psychological, and ethical implications.

Sinnott-Armstrong, W. (2015). Consequentialism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter ed.). Stanford: Metaphysics Research Lab: Stanford University.

Sullins, J. (2010). RoboWarfare: Can robots be more ethical than humans on the battlefield? *Ethics and Information Technology*, *12*(3), 263–275.

Tessier, C., & Dehais, F. (2012). Authority management and conflict solving in human-machine systems. *AerospaceLab: The Onera Journal*, *4*, 1.

The EthicAA team (2015). Dealing with ethical conflicts in autonomous agents and multi-agent systems. In *AAAI 2015 Workshop on AI and Ethics*, Austin, Texas, USA

Tzafestas, S. (2016). *Roboethics: A navigating overview*. Oxford: Oxford University Press.

von Wright, G. H. (1951). Deontic logic. In Mind, Vol. 60, jstor, pp. 1–15.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Woodfield, A. (1976). *Teleology*. Cambridge: Cambridge University Press.

Yilmaz, L., Franco-Watkins, A., Kroecker, T. S. (2016). Coherence-driven reflective equilibrium model of ethical decision-making. In *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pp. 42–48. https://doi.org/10.1109/COGSIMA.2016.7497784.