Formal ethical reasoning and dilemma identification in a human-artificial agent system

Vincent Bonnemains

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Institut Supérieur de l'Aéronautique et de l'Espace (ISAE)*

Présentée et soutenue le *11/12/2019* **par :**

Vincent BONNEMAINS

Formal ethical reasoning and dilemma identification in a human-artificial agent system

---

JURY

Jean-Gabriel GANASCIA	Professeur des Universités	Président du Jury
Catherine TESSIER	Equivalent Professeur des Universités	Directrice de thèse
Claire SAUREL	Docteur	Codirectrice de thèse
Olivier BOISSIER	Professeur	Rapporteur
Virginia DIGNUM	Full Professor	Rapporteur
Matthias SCHEUTZ	Full Professeur	Examinateur
Jérôme PERRIN	Directeur Scientifique	Examinateur

École doctorale et spécialité :
EDSYS : Informatique 4200018

Unité de Recherche :
Office national d'études et de recherches aérospatiales (ONERA)

Directeur(s) de Thèse :
Catherine TESSIER et *Claire SAUREL*

Rapporteurs :
Olivier BOISSIER et *Virginia DIGNUM*

Table of contents

List of figures

List of tables

Preamble

This work does not attempt at proposing scientific improvements in the domain of moral philosophy and ethics. It has been realized by computer scientists sometimes helped by philosophers, and with a significant study of ethics literature.

Remerciements/Aknowledgement

Ce travail est dédié à toutes les personnes m'ayant entouré et accompagné durant ces trois années de thèse:

- En premier lieu, je tiens à remercier mes deux directrices de thèse Catherine Tessier et Claire Saurel, dont le professionnalisme n'a eu d'égal que leur sympathie à mon égard. Si j'ai pu évoluer et m'épanouir au cours de ces années, ce fut sans nul doute grâce à leur encadrement constant, aidant mais jamais oppressant. Ce travail est en grande partie le fruit de leur enseignement pertinent et bienveillant.
- Je remercie aussi ma mère, pour cela ainsi que pour tant d'autres choses. Si j'ai pu réussir, c'est parce que j'ai su pouvoir m'appuyer sur son soutien et son amour infaillible, un phare m'éclairant dans le brouillard, sans lequel je n'aurais pas toujours su me relever.
- À Colline, dont les encouragements et l'amour m'ont permis de m'accomplir, dans ce travail comme dans la vie en règle général.
- Merci à mon père, s'il n'a pas toujours eu les mots, il a su à sa manière m'accompagner et m'aider quand j'en avais besoin.
- À mes sœurs, que j'ai su toujours disponibles pour m'épauler et m'aider à donner le meilleur de moi-même.
- À Lucien, à ta manière aussi tu m'as aidé. Depuis 15 ans épaules contre épaules, on se marre bien!

- À tous mes amis, que je ne peux remercier individuellement sous peine d'avoir des remerciements plus long que le contenu de cette thèse: Cassandre, Tom, Nelson, Arnaud, Amandine, Léa, Mathilde, Émilie, Colin, André et tous les autres, je pense à vous.
- Adèle,Pauline et Mehdi, les meilleurs collègues que l'on puisse rêver, et aussi mes amis, merci.
- Merci aussi à: Betty, Laurence, Valentin, Sébastien, Lucien, Sovanna, Antoine, Mathieu, Gustav, Guillaume, Charles, Cédric et tous les autres.
- I would like to acknowledge Matthias Scheutz and his team for their hospitality.
- Je remercie enfin l'ONERA, l'ISAE-SUPAERO, la région Occitanie, l'INSA et EDSYS pour m'avoir permis de réaliser dans les meilleures condition cette thèse.

Chapter 1

Introduction

1.1 Context

Waymo, Romeo, Pepper, Buddy, drones, etc, it is a fact: so-called autonomous machines are out of the box and they bring with them a bunch of ethical concerns. These issues can be divided into three fields:

- ethical thoughts concerning research on autonomous machines and the design of autonomous machines,
- or ethical thoughts concerning the use and misuse of autonomous machines and how autonomous machines can be part of society,
- or technical approaches aiming at imbuing ethics into an autonomous machine.

This thesis focuses on the third field: from embedded knowledge and perceived and interpreted data, a machine computes decisions related to actions to perform to satisfy goals and criteria . Moreover, this knowledge embeds elements pertaining to ethics and axiology[1], which enables the machine to compute judgements on decisions and justify these judgements for an operator or a user. Indeed, autonomous machines will be put in contexts where computed decisions will require more than a mere multi-criteria optimization to reach a goal, i.e., knowledge including ethical considerations (Malle et al., 2015; The EthicAA team, 2015). For instance, while monitoring elderly people, Romeo will have to deal with privacy, benevolence, respect for the patient's autonomy, and other medical values. In the same vein, an autonomous car or an Unmanned Air Vehicle (UAV) should be able to assess the consequences of a crash.

[1] the philosophical study of values

1.2 Related work in short

1.2.1 Modelling reasoning embedding ethical concerns

The literature dealing with ethical knowledge embedding autonomous machines is composed of different approaches, and among them:

(Cointe et al., 2016) have proposed a global model based on a Belief-Desire-Intention architecture in order to allow an agent to select among all possible actions the one which is "just", thanks to the assessment of ethical principles organized through an order of preference. This model has been built from the perspective of an agent inside a multi-agent system. Hence, the aim was mainly to propose a general architecture to judge an agent's decisions and the decisions of other agents, and not go further into the study of ethical principles.

(Berreby et al., 2015) have extended event calculus in order to model ethical reasoning through ethical principles which is not so far from (Cointe et al., 2016). The idea was to allow an agent to assess whether an action is right or wrong thanks to a set of predicates modelling the Doctrine of Double Effect and other ethical frameworks. This work however does not tackle the decisions that are related to non-action and does not discuss the inherent subjectivity of the model.

Furthermore, considering (Hunyadi, 2019), we argue that an autonomous agent cannot be a fully ethical agent. This is why we support the idea that autonomous agents embedding ethical considerations should be part of an operator/user-artificial agent system. Therefore, we have designed an experiment in order to assess the impact an agent embedding ethical concerns would have on an operator in such a system. As far as we know, experiments with artificial agents embedding ethical considerations are not so common.

1.2.2 Moral reasoning experiments with artificial agents

(Greene et al., 2001) have performed an experiment to identify the properties of moral dilemmas influencing human decision making. The results have shown that people are subject to multiple influences, one of them being the "personal/impersonal" property of actions to perform in a moral dilemma. For instance, as "pushing fatman" in the footbridge dilemma requires a physical contact with a potential victim (i.e. fatman), the action is personal and tends to make people not choose this action, even if the

consequences are the same as ones of the trolley dilemma. Nevertheless, this experiment does not include artificial agents in the decision process.

(Malle et al., 2015) have highlighted the fact that people do not judge artificial agents' behaviors the same way as human behaviors when they act similarly when facing ethical dilemmas. Indeed, it seems that people tend to blame people more for action and artificial agents more for inaction. This experiment tends to show that reasoning embedding ethical concerns should not be a copy of human ethical reasoning. Nonetheless, the decision making resulting from an interaction between an operator and an artificial agent was not studied.

1.3 Positioning with regard to literature

On the one side, the literature presents several methods of modelling reasoning embedding ethical concerns. Most of them adapt computer science methods to model ethics. While this is a relevant approach to study the limits of the tools, we have decided to start from the study of ethical concepts expressed in natural language. The aim of the work is then to reduce the subjectivity of natural language through our formalization. Moreover, modelling ethical frameworks allows to highlight the complexity of ethical reasoning, and particularly to locate sources of subjectivity more precisely.

The review of the literature has revealed that ethics modelling usually assumes that a situation has been identified as requiring moral reasoning. Nonetheless this requires first a formal definition of such a situation, i.e., a formal definition of a moral dilemma, which, as far as we know, has not been studied previously[2].

On the other side, experimental works studying moral reasoning have highlighted the properties of moral dilemmas influencing decision making. Moreover, experiments with artificial agents have shown that people have different expectations from humans and from artificial agents when facing moral dilemmas. Nevertheless, an experiment aiming at studying decision making of an operator-artificial agent system facing moral dilemmas has not been performed yet.

Consequently our work aims first at modelling ethical frameworks in order to allow an artificial agent both to judge decisions and to identify a situation as a moral dilemma. Second it focuses on the way an artificial agent embedding ethical concerns may affect an operator's decision making when facing moral dilemma.

The contribution will be detailed in the four following chapters.

[2]deontic dilemmas have been studied through deontic logic. Nonetheless we argue that deontic dilemmas do not fit our moral dilemma notion which is wider

1.4 Summary of the contribution

1.4.1 Chapter 3 : the formalism

This chapter focuses on the cornerstone of our work, i.e., the model which allows us to represent a situation and judge possible decisions regarding this situation.

Goal 1: *build a situation model allowing to compute judgements on possible decisions, and encompassing the notion of avoidance and non-action.*

The first step is to define a situation model. A first study of major ethical frameworks (i.e., deontogism and consequentialism) has led us to consider that representing a situation requires at least a concept of world state allowing to assess the evolution of the world according to the decision computed by the agent. Nonetheless, it is important to keep in mind that agent's decisions may lead to agent's inaction and an evolution of the world through an event. This is why we represent a situation through an initial *world state*, possible *decisions* of the agent facing the situation and for each decision an *event* that may modify the initial world state to obtain a new world state called the *effect* of the decision. Moreover, we need to be able to evaluate in each world state what is "good" and what is "bad", this is why a world state is composed of *facts*. Finally, because ethics requires to evaluate facts that are "obtained" as well as facts that are "not obtained", we define a fact as a variable with binary values.

Goal 2: *Formalize decision judgement methods*

The second step is to formalise concepts and rules for ethical frameworks based on the situation model in order to judge decisions related to a situation as (ethically) "acceptable" or "unacceptable" depending on the ethical framework used. Therefore, we have made several modelling choices to model ethical frameworks:

- ***Consequentialism*** has been formalised as a combination of negative and positive utilitarianism. Hence, functions allowing to identify positive and negative facts according to agents have been defined, as well as a preference relation between facts. The main idea here was to formalise a rule such that the decision with the "best" effect (i.e., the effect with preferred facts) is judged as "acceptable" while other decisions are judged "unacceptable".

- ***Deontology*** has been formalised through a rule focusing on the nature of a decision in itself. A decision is then judged "acceptable" if its nature is neutral or good, and "unacceptable" if its nature is bad.

- ***Doctrine of Double Effect (DDE)*** has been studied as a promising way of judging decisions ethically. Indeed, DDE is defined in literature with several rules combining both the properties of the decision and the properties of the effect. Hence, we have studied and formalised the concepts of *causality* and *proportionality* to define three rules by which a decision is judged "acceptable" or not.

Once ethical frameworks are formalized, we need to observe how they apply on different moral dilemmas.

1.4.2 Chapter 4 : formalism illustration on three examples

Goal: *Apply ethical frameworks judgements on a complex situation in order to illustrate the model and highlight the limits of the formalization*

This chapter focuses on three moral dilemma formalizations:

- The footbridge dilemma is a classical thought experiment which fits well the aim of illustrating how our model runs.
- The "drone dilemma" is a dilemma we have designed. It highlights how computing a preference between facts may be complex when facts belong to very different fields.
- The "doctor dilemma" comes from (The EthicAA team, 2015). This situation is different from the two previous ones as the agent has to make a decision among three possible decisions.

Through these examples, two classical subjects of ethics are treated: war and the medical field. Furthermore, this chapter allows us to show where subjectivity lies in the models, which is unavoidable when ethics is at stake.

Previous works are based on the assumption that the agent is aware of facing a situation requiring ethical reasoning. Nonetheless, identifying a situation as such is far from being obvious. This is why the following chapter aims at formally defining a situation of moral dilemma.

1.4.3 Chapter 5 : identification of a situation as a moral dilemma

Goal: *propose a formal definition of a moral dilemma based on natural language definitions in order to allow an autonomous machine to identify a situation as such.*

Definitions of a moral dilemma in natural language are numerous. Studying them has allowed us to highlight some properties which are necessary for a situation to be a moral dilemma. Then, these properties need to be formalized in order to propose a formal definition of a moral dilemma. This is proposed thanks to an extension of the formalism defined previously. As for the most important property, a situation is considered as a moral dilemma if all the possible decisions are morally unsatisfactory. Two ways are proposed for a decision to be unsatisfactory: either by the nature of the decision in itself or by the consequences (i.e., the effect) of the decision. Afterwards, this definition is applied on several examples in order to evaluate which situations are identified as moral dilemmas depending on the agent and references that are considered.

We have built a situation model and formalized ethical frameworks in order to judge decisions. This way, artificial agents may embed ethical concerns in their reasoning. Still, it must be kept in mind that artificial agents cannot be fully ethical agents and thus should not be placed in situations of moral dilemmas alone, but rather should be supervised by operators or users. Nonetheless, it is for now very difficult to assess how artificial agents may impact the ethical decision making of an operator facing a moral dilemma. Consequently we have performed an experiment to that end.

1.4.4 Chapter 6 : study of the impact of an artificial agent embedding decision judgement capabilities on an operator

Goal: *Design and conduct an experiment aiming at highlighting the impact of an agent embedding moral decision judgement capabilities on a human (i.e., an operator).*

This chapter presents the experiment we have performed and its results: participants were asked to make a decision in front of moral dilemmas with and without an artificial agent supporting a decision (either a deontological or a consequentialist decision) with an argument. In order to have the best possible material, we have first performed two

online pre-experiments. The first pre-experiment aimed at selecting the best possible argument for each decision in each moral dilemma selected for the experiment. The second pre-experiment aimed at categorizing the moral dilemmas for the experiment. Indeed, the literature highlights the fact that the properties of moral dilemmas have a great influence on the participants' decision makings. Therefore, we have selected eighteen dilemmas, splitted into two categories: moral dilemmas where people mainly tend to make a deontological decision, and moral dilemmas where people mainly tend to make a consequentialist decision.

The results of this experiment have allowed us to identify two standard profiles of participants: people who mainly tend to make consequentialist decisions, and people who mainly tend to make deontological decisions. This profile categorization has allowed us to highlight the fact that deontological people are more sensitive to the artificial agent's choices and arguments. Indeed, deontological people tend in a significant manner to let the artificial agent "decide" (i.e., they neither actively confirm nor cancel the artificial agent's choice but wait until time to make a decision is over, which have been explicitly noticed at the beginning of the experiment as agreeing with the artificial agent) when it agrees with their first decision.

Chapter 2

State of the art

This chapter will lay the foundations of our work by describing the main concepts used later in this manuscript. Starting from distinguishing Ethics and Moral, we then describe normative ethics, applied ethics and meta-ethics. After that, because we assume normative ethics as best fitting our aim, we focus on different ethical frameworks belonging to it. Moral dilemmas are briefly defined before rising some issues about so called "moral machines". Then, we present different models of automated reasoning embedding ethics and categorize them. The aim of this chapter is not to be exhaustive and particularly not to cover the flourishing number of philosophical theories about moral and ethics, but to provide a basis for chapter 3. Moreover, the orientation of this chapter is underpinned by choices which will be justified further.

N. B. :

1. Some references might seem to be associated with a surprising year, such as d'Aquin (1853). Indeed, the year specified here is the date of the publication that is cited, and not the original date.

2. The vocabulary used in the following sections refers to the articles cited and does not mean that the author of this thesis agrees with it.

2.1 Ethics and Moral

The roots of moral philosophy are still in debate. Nonetheless, it is possible to find a consensus on the idea that it was created, or developed while living in groups was necessary for humans to survive. Indeed, from divine interventions of Shamash, exhibited at the Louvre Museum, or Zeus in Plato's Protagoras (Singer, 2005), to

Levinas' hypotheses of human face-to-face birth of morality (Bergo, 1999), it seems legitimate to identify moral philosophy as the science aiming at identifying clues of a good way of living with others.
Ethics and *moral* are two terms widely used in moral philosophy, often interchangeably. While these uses are relevant as their respective Greek and Latin roots mean the same thing (Downie, 1980), in this work we will subscribe to Ricoeur's point of view (Ricoeur, 1990), assuming that *moral* and *ethics* can depict different fields. We will follow Cointe's work (Cointe, 2017), associating *moral* with the theory of the good, and *ethics* with the theory of the right (Timmons, 2012):

- Moral is the ability to categorize an action or a fact as good or bad, based on society, education, religion and/or other sources of moral assessment.
- Ethics is a thought process using moral assessments to judge an action as right or wrong. For instance, an action might be right even if it is morally assessed as bad.

It is worth noticing that basic decision making[1] can be guided only with *moral*. Nevertheless, the world cannot be reduced to a binary assessment. This is why *ethics* was designed to answer issues at stake when facts or decisions can be both evaluated as good and bad, or when a choice has to be made between unsatisfactory alternatives (see "moral dilemmas" 2.4).

2.2 Normative ethics, metaethics and applied ethics

Ethics is a major and wide theme of philosophy. This is why it has been splitted into three main ethical theories (Cavalier, 2014; McCartney, S. and Parent, R., 2012):

2.2.1 Normative ethics

Normative ethics aims at judging a person, an action or a decision as right or wrong depending on an ethical principle (MacIntyre, 2003). Therefore, normative ethics is mainly used to answer the question "What ought I to do?" (Kagan, 2018). A classic example of normative principle is the Kantian categorical imperative (Paton, 1971), which evaluates a decision as right iff it can be translated to a universal moral principle. In this case, it is for instance always wrong to "lie". The literature is full of ethical frameworks of normative ethics, some of which will be described below (see 2.3).

[1] such as deciding to give a hand to someone or not

Ethical principle

The notion of ethical principle is widely employed while rarely defined. In this work, we define an ethical principle as: "a rule or set of rules aiming at judging the rightness of a decision or an action". For instance, Aristotle's equality principle *"treat like cases alike" (from Barnes et al. (2014))* fits this definition well.

2.2.2 Metaethics

A definition of metaethics can be found in (Pollock, 2014):

"a discipline that investigates the meaning of ethical systems and whether they are relative or are universal, and are self-constructed or are independent of human creation."

Indeed, *ethics* preceded by *meta* induces intuitively a notion of "bird's eye view" on ethics itself. It deals with semantic and epistemic issues of ethics. Basically, we can find in metaethics questions such as *what is the meaning of "good/bad" ?*, or *"is ethics culturally relative ?"* (Evans and MacMillan, 2014). Metaethics is also at stake while studying how normative principles are applied and interpreted. Indeed, the notions of "good" and "bad" are deeply related to some normative frameworks and then, can modify the way an action is judged as *right* or *wrong*. Another influence of metaethics relies on basic concepts used in ethical frameworks, such as consequences in *consequentialism* (see below section 2.3.2).

2.2.3 Applied ethics (Singer et al., 1986)

Applied ethics is more and more in the spotlight considering the evolution of technologies, such as in medical care or robotics. It consists in analysing controversial moral issues of real-life (abortion, euthanasia, robots in society, etc.), using ethical deliberation and thought to answer them. This is often illustrated through codes of ethics for different professions. It might seem that applied ethics is only the implementation of normative principles. Nonetheless, several normative frameworks may be relevant for such complex issues, with different and opposed judgements for each of them. Moreover, as it was said in the metaethics paragraph (see 2.2.2), the judgement may vary depending on how normative principles are interpreted through metaethics.

2.3 Ethical frameworks of normative ethics

Ethical frameworks have a common property: they are single or set of principle(s) which allow a decision to be *judged*.

2.3.1 Judgement

The ethical judgement is the assessment of a decision as right or wrong depending on the moral principles used. It is worth noticing that it does not guarantee that a decision will be able to be selected. Indeed, cases where several decisions are right, or all decisions are wrong, are frequent (see *moral dilemmas*, section 2.4). Still, it is a first step toward ethical behaviour as it helps distinguishing what is *right* from what is *wrong*.

Among these frameworks it is possible to identify some common ways of judging, allowing to group them into categories. Let us present some of the most well-known categories of ethical frameworks of normative ethics.

2.3.2 Consequentialism

Consequentialism stems from teleologism, from the Greek "telos" (end) and "logos" (reason). Considering this paradigm, behaviours are justified by fulfilling purposes and pursuing ends (Woodfield, 1976), which is well illustrated by the maxim supported by the Italian humanist Machiavel: "The end justifies the means".

For consequentialism, decisions are judged as *right* or *wrong* depending solely on the morality of their consequences. The main issue with consequentialism is the way to evaluate the consequences and to assess their "good/bad nature". There are many consequentialist frameworks that propose different ways of consequence assessment. Among them, the most popular are the following (Sinnott-Armstrong, 2015):

- *Utilitarianism* is probably one of the most common frameworks of consequentialism. It originally consists in quantifying the amount of "good" among the consequences in order to judge the decision maximizing the good as "right", while the others are "wrong" (Bentham and Mill, 1961). One criticism of this approach is the fact that the amount of "bad" is ignored. This is why "negative utilitarianism" appeared, promoting the minimization of the "bad"(Acton and Watkins, 1963), as the only criterion or as a supplementary criterion added to the maximization of the "good".

- *Hedonism* gives a way to evaluate the "good" and "bad" consequences through notions of pleasure and pain. Therefore, generating pleasure is associated with the "good" while generating pain is associated with the "bad".
- *Egoism/altruism* are frameworks which question the target of the "good"/"bad". Egoism focuses on the "good/bad" for the agent making the decision, while altruism focuses on "good/bad" for the others.
- *Direct Consequentialism* questions the consequences to be taken into account. This framework proposes to take into account only the direct consequences[2] of the action.

It is worth noticing that these frameworks are not mutually exclusive and can be combined. For instance, a framework might be a negative altruist-hedonist utilitarianism, focusing on minimizing the pain of others. These frameworks are interesting because they rise metaethics questions. Indeed, minimizing "bad" and maximizing "good" require to define these "good" and "bad". Moreover, because consequentialism focuses on consequences, it is relevant to think about which consequences are considered. Direct consequences are likely to have consequences themselves, which will induce other consequences, etc. This is a classical issue about transitive closure.

2.3.3 Deontology

Let us imagine a situation where the death of a patient could save five others. Nonetheless, the first patient is in good shape and must be killed to save the five others. In this case introduced by (Foot, 1967) and many others, consequentialist reasoning leads to the idea that, assuming the life value of one person is equal to the life value of anyone else, killing one person to save other lives is always justified. This assertion is highly questionable, and has induced the emergence of the deontological framework. Deontology is often defined as opposed to consequentialism. Indeed, this framework is based on the respect of some moral rules (from Greek *deon*, "the duty"), disregarding the consequences (Alexander and Moore, 2007). This ethical framework judges a decision as right iff it respects one or several principles. The categorical imperative mentioned earlier (see *normative ethics* paragraph 2.2.1) is a perfect example of this framework, for instance: it is wrong to lie, whatever the consequences. Nonetheless, respecting moral principles (in this case "not lying") is subject to interpretation, because notions

[2]Direct consequences are the consequences induced by the action itself without intermediary. This notion remains subject to interpretation, as evoked in paragraph 2.6.1

at the core of moral principles are likely to be defined in different ways (e.g. lying by omission, being vague, etc.). Moreover, principles are often in conflict when a complex decision has to be made. Which principle/s is/are relevant in a particular situation and how to apply it/them are the main issues at stake while considering deontological frameworks.

2.3.4 Virtue Ethics

The third main ethical framework (Hursthouse and Pettigrove, 2016) is slightly different from the two previous ones, because its affiliation to normative ethics is ambiguous. Indeed, while deontology and consequentialism guide the decision through decisions and consequences, virtue ethics guides decision by an indirect way: it aims at defining "how to be" instead of "what to do" and then focuses on the agent. Nevertheless, some philosophers claim that this nuance does not make Virtue ethics application different from the way deontology is applied (Hursthouse, 2012). Indeed, moral principles of deontology require to be interpreted in order to lead the decision. This can be compared to the way virtues require to be linked to decisions promoting them. This idea could be extended to any ethical framework: even "the greatest good" of utilitarianism has to be interpreted in order to apply. Hence, an important distinction has to be made between an implementing rule and the principle in the name of what it has been defined, just as a law is not applicable until its implementing decree is defined. Another reason of virtue ethics singularity lies on the properties of virtues. Consequentialism and deontology can be applied at time t, without regarding the past. On the contrary, it seems that virtues come from habit and/or learning, and deal with a notion of persistence and evolution (Aristotle, 2012). For instance, somebody is generous not because they give one time, but because they fulfil this virtue all along their life.

2.3.5 Doctrine of Double Effect (DDE)

The Doctrine of Double Effect, even if not officially cited as a normative framework, is often used to evaluate a decision as "right" or "wrong" and then fits perfectly our definition of normative ethics. This doctrine was first introduced by Thomas d'Aquin (d'Aquin, 1853) while arguing that killing in a self-defence situation was acceptable. It can be defined as four rules to respect in order for a decision to be judged as "right"(Connell, 1967): "

- *The act itself must be morally good or at least indifferent.*

- *The agent may not positively will the bad effect but may permit it. If he could attain the good effect without the bad effect he should do so. The bad effect is sometimes said to be indirectly voluntary.*
- *The good effect must flow from the action at least as immediately (in the order of causality, though not necessarily in the order of time) as the bad effect. In other words the good effect must be produced directly by the action, not by the bad effect. Otherwise the agent would be using a bad means to a good end, which is never allowed.*
- *The good effect must be sufficiently desirable to compensate for the allowing of the bad effect*

"

It is worth noticing that this definition includes rules concerning the decision in itself, and rules concerning consequences. The Doctrine of Double Effect then shares both deontologist and consequentialist properties.

The following table (figure 2.1) summarizes the objects that each category of frameworks focuses on:

Table 2.1 Summary table of framework categories

Category	**Focuses on**		
	Decision	**Consequences**	**Agent**
Consequentialism		×	
Deontology	×		
Virtue			×
DDE	×	×	

2.4 Moral dilemmas

Ethical theories are debated and criticized as soon as they are conceived. Moral dilemmas, as complex situations where decisions are morally unsatisfying (see chapter 3: Contribution), are useful tools to depict how moral and ethics apply in situation and then to highlight the difficulties and failures of each theory. The most famous dilemma is probably the trolley dilemma (Foot, 1967):

"*...he [the agent] is the driver of a runaway tram which he can only steer from one*

narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed."

The question of the very existence of moral dilemmas has been debated for a long time. Thomas d'Aquin and Kant argued for instance that, if there is a way to determine the right decision to make in any situation (through their respective doctrines (d'Aquin, 1853; Kant, 1795)), moral dilemmas cannot exist. Nonetheless, Williams claims that dilemmas exist because facing a case of two opposite obligations, an agent will inevitably feel regrets by choosing one of the possible decisions (Williams and Atkinson, 1965).

2.5 Moral machines

Buddy, Romeo, Nao, autonomous cars, Tay, Watson and others share a common property: autonomy. It is worth noticing that the use of this kind of term, such as "intelligence" or "learning" is subject to interpretation and is often exploited to hit the headlines. The technical term "autonomy" applied to machines has a narrower definition than the usual term. According to the US DoD Defense Science Board (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."

This definition matches CERNA's (Grinbaum et al., 2017):
"[the] capacity to operate independently from a human operator or from another machine, by exhibiting non-trivial behaviours in a complex and changing environment"

It remains that autonomous machines will be put in contexts where computed decisions will require more than a simple multi-criteria optimization to reach a goal, i.e., knowledge including ethical considerations (Malle et al., 2015; The EthicAA team, 2015). Medical values[3] (Beauchamp and Childress, 1979) are at the core of many decision makings while monitoring elderly people or caring for patients. In the same vein, any autonomous vehicle should be able to assess the consequences of a crash (BBC, 2015). Moreover, such as any machine, autonomous systems are sensitive to bugs, hacks and other failures. Even if these accidents are statistically rare, because

[3]Privacy, Benevolence, Respect for the patient's autonomy, etc.

these machines will be widespread with a large range and potentially with lives at stake, the impact of a default might be devastating (Lin et al., 2008). Adding ethical concepts in the automated reasoning could be helpful to counter some of these risks, by weighing bad outcomes of any decision.

The key question is: is it really possible to put ethics in robots? Philosophers claim that ethics is a prerogative of mankind, or at least of living beings. For some of them, ethics requires pain and pleasure to be experimented (Aristotle, 2012). For others, "counter-factual thinking" is the cornerstone of ethical reasoning (Wachter et al., 2017). Because this concept represents the intuition of "what ought to be", as opposed to what is, sensors and automated reasoning of a machine are far from being able to encompass such a notion (Hunyadi, 2019).

Therefore, even if they modify in an indirect way moral values of our society (Vallor, 2016), machines are not likely to be fully ethical agents (Moor, 2006). Thus, we will rather use in this work the term *reasoning embedding ethical concepts* instead of *ethical reasoning*. Moor distinguishes two categories of autonomous agents embedding ethical concepts:

- *implicit ethical agents* act ethically because they *"implicitly supports ethical behavior"*. It means that these agents are ethical thanks to the way they were designed and placed in situations to avoid unethical behaviors. In this way, our cars can be considered ethical because they are safe thanks to ABS (*"antiblockiersystem"* avoids wheels to be blocked while braking) and ESP (*Electronic Stability Program*). A computer is ethical for a child thanks to parental control. Nonetheless, these machines cannot analyse and reason on the decisions they compute.

- *explicit ethical agents*, on the contrary, manage ethical concepts while reasoning to make a decision, considering the "good/bad" and the "wrong/right" nature of these decisions. These machines, to the best of our knowledge, do not exist yet. Nonetheless, our work belongs to the research field aiming at modelling ethical concepts to judge a decision.

It is worth noticing that these categories have to be considered regarding the concept of "*fully ethical agent*", which is the term defining a truly ethical agent. The phrase (implicit/explicit) "ethical agent" brings confusion on the fact that *machines are not and will not be ethical agents* (at least in the near future).

2.6 Ethics modelling

Ethics modelling has become popular for a few decades. Some promising models have been described which allow ethical concepts to be manipulated in automated reasoning. To present them, we will use the classical classification proposed, for instance, in (Wallach and Allen, 2009):

- Top-down methods start from ethical principles, e.g., ethical frameworks, and identify the necessary concepts that are required to apply these ethical principles, e.g. "judgement","agent","consequences", etc. and represent them with relevant formalisms.
- Bottom-up approaches are based on experience and "learn" which decisions are ethical according to situations.
- Hybrid methods share both top-down and bottom-up approaches.

2.6.1 Top-down methods

Declarative models

The use of logic for modelling ethical principles is intuitive. Nonetheless, classical logics[4] fail to encompass ethics because of its versatility. Indeed ethics, as a complex philosophical field analysing the real world, cannot be simplified to universal laws without exceptions. Therefore, any model attempting to represent ethics requires the ability to catch exception and context-sensitivity. To that end, (Ganascia, 2007) stresses the relevancy of using non-monotonic logic. This is particularly well-illustrated in the work of (Berreby et al., 2015), which is based on a combination of Answer Set Programming (a non-monotonic framework (Brewka et al., 2011)) and Event Calculus (Kowalski and Sergot, 1989) to formalize the Doctrine of Double Effect (see 2.3.5) on the Trolley Dilemma (see 2.4). This work has been deepened in (Berreby et al., 2017), with a declarative modular framework (see figure 2.1), allowing the representation of a situation and the judgement of an action depending on an ethical framework.

This model is composed of four modules:

- The *Event Motor* allows to represent the current situation through fluents and events which will modify the situation when their preconditions are fulfilled.

[4]such as propositional and first order logics

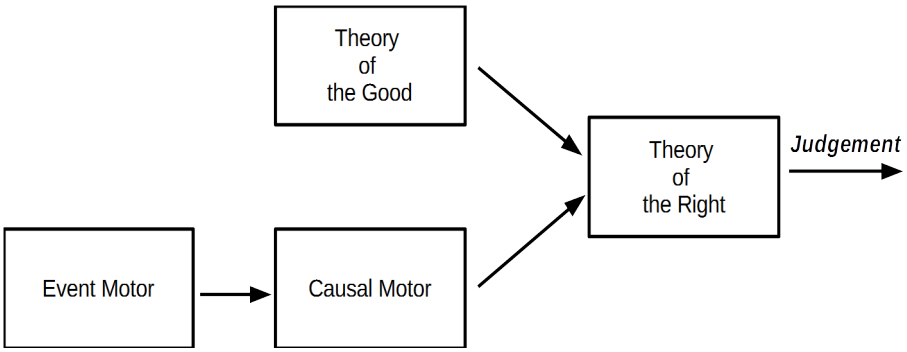


Fig. 2.1 Models and Modularity, based on Berreby et al. (2017)

- The *Causal Motor* computes which fluents are direct or indirect[5] consequences of an Event.
- *The Theory of the Good* is related to "good/bad" assessment of an event. The first step in this module is to qualify the "good/bad" nature of events thanks to different logical rules. The second step is to quantify the amount of events qualified as "good".
- *The Theory of the Right* is the final step of the reasoning. It consists in using the results of the Causal Motor and an assessment of the good given by the Theory of the Good to compute the events which are "right" depending on the ethical framework that is considered.

This work is based on a previous article (Berreby et al., 2015), modelling the Doctrine of Double Effect (see 2.3.5) on the Trolley dilemma. This work discusses the notion of causality and responsibility . These notions are relevant features at the core of many ethical frameworks, such as for direct/indirect consequences for Consequentialism (paragraph 2.3.2) or the causal relation required for the DDE. It is pointed out in the article that *causality* and *responsibility* are subtle and complex notions. Indeed, even if intuition would have led us to define causality as: *"α is a cause of β, if, had α not happened, then β would not have happened"* (citation from (Berreby et al., 2015)), this definition fails to encompass many issues such as "preventing something". Indeed, while an agent avoids an effect by preventing its cause, what is the causal relation of this avoidance? what is the responsibility of the agent concerning the effect avoided? This work supports the idea that a formalism attempting at modelling ethics requires to take causal and preventing relations into account.

[5]as for consequences (see paragraph 2.3.2), direct fluents are fluents resulting from the event without any intermediary.

Another attempt to model ethics through logic has led to the use of a category of modal logic: deontic logic (von Wright, 1951). The very basic logic of this category, usually called Standard Deontic Logic (SDL), is a classical modal logic with only one modality O, Oq meaning that q is obligatory. From this statement, two convenient short-cut modalities can be defined:

- Fq equivalent to $O\neg q$, q is forbidden.
- Pq equivalent to $\neg O\neg q$, q is permissible.

Plenty of deontic logics were derived from SDL in order to solve philosophical weaknesses such as Chisholm's paradox (Chisholm, 1963; McNamara, 2014). This paradox highlights the issue of "contrary to duty" obligations, when we ought to do something which is forbidden:

1. It ought to be that Jones goes (to the assistance of his neighbors) (Ogo).
2. It ought to be that if Jones goes, then he tells them he is coming ($O(go \rightarrow tell)$).
3. If Jones doesn't go, then he ought not tell them he is coming ($\neg go \rightarrow O(\neg tell)$).
4. Jones doesn't go ($\neg go$).

This situation which seems correct leads to two inconsistent sentences: $Otell \wedge O(\neg tell)$. Some clues have been then proposed to solve such issue by adding complexity to SDL. For instance, Nute (2012) argues for a differentiation between "factual detachment" and "deontic detachment". The first one being a derivation from a factual statement, such as $O(\neg tell)$ derived from $\neg go$ and $\neg go \rightarrow O(\neg tell)$, and the second one derived from a deontic statement, such as $Otell$ derived from Ogo and $O(go \rightarrow tell)$. Therefore, the solution would be to make a distinction between these two detachments which could be done through temporal indexes, obtaining $O_{t1}tell \wedge O_{t2}(\neg tell)$. Nonetheless, it is said in this work that temporal factors are sometimes inefficient to solve some inconsistencies, such as with general moral principles, which are not affected by time.

Hence, dyadic system is proposed to encompass this issue. In this case, the predicate $O(p)$ becomes $O(q/\phi)$, meaning that "It ought to be that q given that ϕ". Hence, the conditional property can handle the inconsistency through conditioning a statement to its feasibility.

Bringsjord and Taylor (2011) have proposed LRT*, a deontic logic to formalize the *divine command approach*. This ethical framework can be categorized as a deontological

framework. Indeed, it consists in considering an action as "right" if it fulfills a principle commanded by God. Even if this framework is debatable because of its subjective origin (Wainwright, 2006), this work shows the possibility to model ethical frameworks through deontic logic rules.

BDI models

Belief Desire Intention model is probably *"the best known and best studied model of practical reasoning agents" (from (Georgeff et al., 1999))*. A BDI architecture includes three components:

- *Belief* ($\mathcal{B}$) is the set of data representing the world.
- *Desire* ($\mathcal{D}$) is the set of states of the world or agent's states the agent desires to reach.
- *Intention* is the result of a decision computation based on *Belief* and *Desire* to obtain the subset of *Desire* the agent intends to reach.

Cointe et al. (2016) have improved this model by adding components related to moral and ethics in order to represent the reasoning process of an agent embedding ethical considerations.

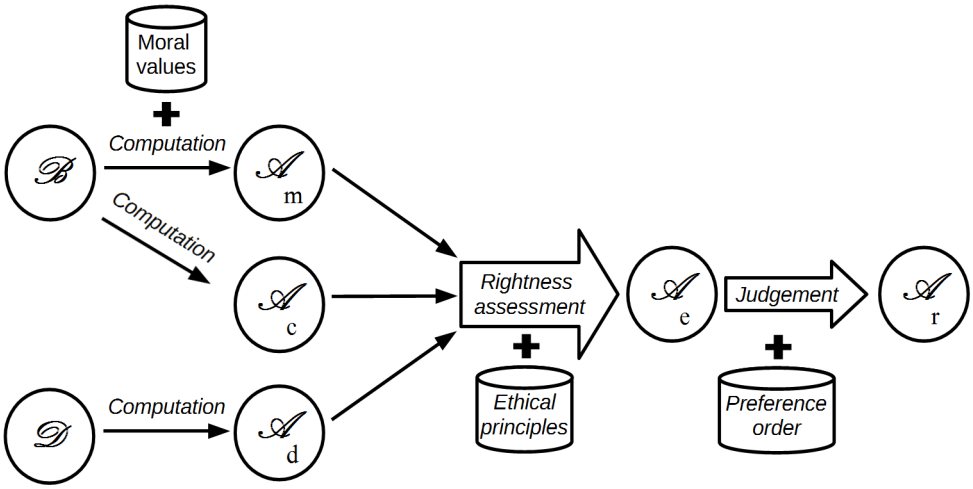


Fig. 2.2 Ethical judgment process (based on Cointe et al. (2016))

Basically, it consists in considering three sets of actions:

- $\mathcal{A}_c$ is the set of feasible actions, obtained from the set of all actions $\mathcal{A}$ filtered by an awareness process based on *Belief*.

- $\mathcal{A}_d$ is the set of desirable actions, filtered from $\mathcal{A}$ through *Desire*.
- $\mathcal{A}_m$ is the set of moral actions, based on moral rules and values supports knowledge.

Then, the "rightness" of these actions is assessed according to a set of ethical principles, giving $\mathcal{A}_e$. Finally, the judgement allows to select the rightful action(s) to perform $\mathcal{A}_r$, considering a set of preferences among ethical principles.

Value Based Reasoning

Many models embed values as a way to consider ethics. (Atkinson and Bench-Capon, 2016) for instance, base their work on Action-based Alternating Transition Systems (ATTS), an extension of State Transition Diagram (STD). In this work, values are symbolized by numerics on transitions, representing how much a transition (i.e., an action) supports a value. Then, the decision is computed through utility function over the values.

2.6.2 Bottom-up methods

Compared to top-down methods, bottom-up methods allow the agent to adapt its behavior over time. This is a major benefit of these approaches, though they often fall down under the requirements of safe provability and justification (Castelvecchi, 2016; Winfield and Jirotka, 2017).

Case-based reasoning

GenEth, the general ethical dilemma analyzer of (Anderson, 2015), is interesting because, even if its learning comes from previous cases, its design preserves an ethical meaning. This approach uses *inductive logic programming* (ILP) to compute a general rule constructed as a combination of Horn clauses. ILP is an inductive algorithm, extracting from previous cases categorized as negative or positive for a specific property a complete (meaning that all the positive cases are encompassed by the hypothesis) and consistent (none of the negative cases are encompassed) hypothesis. The algorithm is based on five concepts:

- *ethical features* are the relevant features to take into account by the ethical reasoning. They are depicted as integers for a specific action, representing how much the feature is present while performing this action (positive number if the feature is present, negative if absent).

- a *duty* is attributed to each ethical feature. It represents the fact that an ethical feature has to be minimized or maximized. For an action, the value of a duty will be as positive as the duty is respected or as negative as the duty is violated.
- *action* is therefore a tuple of integers, each of them symbolizing the degree of respect/violation of a duty.
- a *case* is a tuple resulting from the subtraction of one action's tuple to another. They consider positive and negative cases. Positive case is when the tuple of less preferable action is substracted from the tuple of more preferable action and vice versa for negative case.
- a *principle* is the resulting rule computed from cases. It allows to compute which action is preferable between two actions.

This approach is defined as a *case-supported principle-based behavior paradigm* (CPB), because it uses a principle computed from previous cases to determine which action to perform. Moreover, the initial cases used to compute the principle are, for *GenEth*, determined by ethicists. Nonetheless, as generalization approaches tend to extract rules from common properties, they might fail to encompass the context-sensitivity of ethics which is evoked in paragraph 2.6.4.

Another approach provided in (Kuipers, 2018) allows to avoid generalization from cases. The approach is to compare previously encountered cases with the current case the agent is facing. Thanks to a *"Rich description"* of the context, it would be possible to find the most similar known cases and the previous outcomes while facing these cases. Formally, the knowledge is composed of numerous cases represented as:

$$\langle S, A, S', v \rangle \tag{2.1}$$

meaning that facing situation S, the agent has done action A and obtained resulting situation S'. This case is morally assessed with value v. It is pointed out that this method is not applicable for now because of some difficulties such as the description of contexts, and has some failures such as known cases both similar to the current case and indicating that agent should do A and $\neg A$.

Reinforcement learning

Reinforcement learning (RL) is another bottom-up approach that allows to get rid of the requirement for previous encountered situations. Moreover, it avoids generalization.

(Abel et al., 2016) claim that POMDP might be of great help considering several ethical issues. This method is based on a reward function which is updated while observing the result of actions. The article ends with some of the main issues at stake with RL and some particular problems of ethical RL. For instance, this method requires "teachers" who aim at defining the norms linked to the ethical utility function. Therefore the issues are to select the accurate teachers and to specify what to do when teachers are in conflict.

2.6.3 Hybrid methods

Hybrid methods usually propose a starting model of ethics that can evolve over time thanks to incoming data. In that way, ethical rules can be formalised (top-down) and be updated next (bottom-up).

Ethical Governor

Arkin's work (Arkin, 2007) is often cited as the foundation stone for agents embedding ethical concepts. Specifically dedicated to warfare field, his research is based on an added module to a classical architecture of automated reasoning. This module called *"Ethical Governor"* aims at controlling the ethical conformity (in accordance to war laws) of actions related to the use of lethal force. According to Arkins, there are three main sets of behaviours:

- $P_{l-ethical}$: the set of lethal ethical behaviours
- P_{lethal} : the set of lethal behaviours
- P : the set of behaviours

$$\text{with } P_{l-ethical} \subset P_{lethal} \subset P \tag{2.2}$$

Therefore the ethical governor is designed to identify the belonging of a behaviour to these sets. Then, the bottom-up part of the approach, called *"Ethical Adaptor"* is in charge of *"After-action reflection"*. This means that its evaluation of the *"performance of the system, triggered either by a human specialized in such assessments or by the system's post-mission cumulative internal affective state, will provide guidance to the architecture to modify its representations and parameters"* (from (Arkin, 2007), page 72). This work was a great step toward machines embedding ethical reasoning and has given relevant guidelines for future works. Nonetheless, as an initial work not

concretely implemented, it rises several issues, such as safe provability of the learning module and non-consistency of constraints management facing a situation (such as in case of a moral dilemma). Another important difficulty highlighted in this work is the ability to translate laws into "machine-usable representations"(from Arkin (2007) p. 98), exempted from ambiguity of natural language.

Game theory and machine learning

(Conitzer et al., 2017) question the ability of optimization methods to model ethics. To that end, game theory is proposed as a starting framework to model ethical dilemmas in terms of a reward function, supported by machine learning. An example of combination is given in the discussion:

"we can apply moral game-theoretic concepts to moral dilemmas and use the output (say, "right" or "wrong" according to this concept) as one of the features in our machine learning approach. On the other hand, the outcomes of the machine learning approach can help us see which key moral aspects are missing from our moral game-theoretic concepts, which will in turn allow us to refine them"(from (Conitzer et al., 2017))

Therefore, the use of a combination of these well-known methods could be a clue for modelling ethics. Nonetheless, this method shares both benefits and failures of top-down and bottom-up approaches. Indeed, game theory is a classical numerical approach which is questionable while combined with ethics (see paragraph 2.6.4). In the same vein, the black box issue of machine learning remains.

Normative Reasoning

The word *normative* is often cited while speaking of ethics, such as in "normative ethics". Therefore, studying norms is relevant for ethical reasoning. (Sarathy et al., 2017) investigate this approach, arguing that norms improve interaction and collaboration between humans and robots. In this work, a norm $\mathcal{N}$ is defined as:

$$\mathcal{N} := [\alpha, \beta] :: C_1, ..., C_n \Longrightarrow (\neg)\mathbb{D}(A_1, ..., A_m) \tag{2.3}$$

Where $C_1, ..., C_n$ are contexts in which $\mathcal{N}$ applies, $\mathbb{D}$ is a deontic predicate (obligatory, forbidden, etc.) and $A_1, ..., A_m$ are actions or states. Then, $[\alpha, \beta]$ is a Dempster-Shafer uncertainty interval such as $0 < \alpha < \beta < 1$, 0 meaning that the norm is forbidden and 1 that the norm is obligatory. The algorithm explained in the article proposes a way to learn norms from observations in a context. Dempster-Shafer approach (Shafer, 1976)

guarantees that the interval will narrow with new observations. Nevertheless, this work does not deal with norm conflicts.

(Serramia et al., 2018) focuses on this issue, starting from a basic definition of norm similar to the previous one:

$$\theta(\rho, ac) \tag{2.4}$$

With θ a deontic predicate, ρ an agent and ac an action. Then, the article defines a set of relations between norms, in order to compute the largest set of norms respecting several criteria (such as no-conflict between norms) based on the relations defined earlier. In order to compute this set of norms, an optimization function is defined with a cost associated to each norm. This work goes further by suggesting that moral values could be taken into account while using the optimization function. To do that, they assume a total order over moral values, and the ability to compute for each norm how much it promotes each moral value.

Quantified multi-operator modal logic

A bold attempt to model an evolving virtue ethics can be found in (Govindarajulu et al., 2019). This approach is based on *"deontic cognitive event calculus"* ($\mathcal{DCEC}$), an event calculus combined with a BDI architecture with time consideration. In this article, a utility function allows to compute a value for a fluent and therefore a value for an event (as the sum of fluent utilities). These utilities represent an assessment of "the good". The second step of the article concerns learning of *traits* from other agents thanks to a model inspired from the theory of emotions. A trait is defined as a tuple of a situation and an action type. Then, the agent learns traits from other agents if it admires them. To be admired, an agent has to perform a number of actions assessed as good by the agent that is learning. Thus, the learning agent can learn a generalized trait from several observations of admired agents. For instance:
" if the action type "being truthful" is triggered in situations: "talking with Alice", "talking with Bob", "talking with Charlie"; then the trait learned is that "talking with an agent" situation should trigger the "being truthful" action type." (from (Govindarajulu et al., 2019))
This approach is an interesting model of virtue ethics following the idea that virtues are learned from relatives. Nonetheless, it shares common issues of numerical methods discussed below (see paragraph 2.6.4). Moreover, copying human moral behaviour does not seem to be suitable as people do not judge machines as they judge other

people (see paragraph 6.1.2). Finally, the issue of learning contradictory traits from two admired people remains.

2.6.4 Model classification

The models described above share common properties. This part attempts to highlight and discuss these properties and to classify the works. Here are the properties we propose to select:

- **Number-based approach (denoted as "Num"):** an important part of optimization and AI approaches lies on numbers. Hence, the temptation to use numbers in order to model ethics is high. Indeed, these methods have proven their efficiency to guarantee optimal autonomous behaviours while considering time, costs, benefits, etc. Nonetheless, weighting ethical principles in a cost function, defining a value of concordance between a virtue and an action or other quantifications induce several issues. First of all, the projection onto $\mathbb{R}$ inevitably requires simplifications which are not desirable for ethics. Moreover, the way these numbers/orders are obtained is highly questionable and is likely to accentuate the biases inherent in any modelling. Then, this convenient simplification would allow some concepts to be comparable, while they are not, which could be irrelevant and dangerous in certain cases. Because total orders induce an equivalent projection, we assume them as belonging to this category, even if they get rid of numbers. Lastly, in a meta-ethical thought, using numbers to determine the most suitable behaviour/decision means that this behaviour/decision is selected only because it has obtained the highest score, which does not reflect ethics[6] and furthermore does not allow relevant justifications. Considering the works described earlier, a large part of them use at least (partial or total) orders. Nevertheless, the way numbers are obtained and used differs from one article to another. Therefore, the degree of simplification and abstraction related to these approaches varies, this is why we define three sub properties:
  - *not num*: the model does not use numbers or very few which do not impact the complexity of information
  - *num*: the model uses numbers which simplify information but are not at the core of the reasoning to make a decision

[6] This issue was risen in the famous movie "I. Robot" (Proyas, 2005), is it ethical to make decisions based on utility scores when lives are at stake?

  - *high num*: the model uses numbers as the core of the model to make a decision

- **Implemented (denoted as "Implement"):** ethics is a complex field which is very different from mathematics, and thus hardly fits the binary field of computer science. This is why after a few decades of research in the field of machine ethics, the approaches remain simplifications of ethical reasoning, and in some specific cases theoretical approaches are not implemented. Therefore, we distinguish implemented approaches from theoretical approaches.

- **Context-sensitivity (denoted as "Context"):** the notion of context is often at the core of ethical reasoning (Sarathy et al., 2017) . For instance, the way a moral value or principle applies highly depends on the context. It is well-known that defining the notion of context is at least extremely complex if not impossible: *"it does not seem possible at the present time to give a single, precise, technical definition of context, and eventually we might have to accept that such a definition may not be possible"* (from (Duranti and Goodwin, 1992))
  Nonetheless, it is possible to have a concrete idea of a context definition by identifying what it is composed of:
  "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves." (from (Dey, 2001))

  This definition does not attempt to define the whole concept of context, but to propose what could be considered as belonging to the context, in the specific field of human-machine interaction. This work is thus sufficient to get the point about this criterion: the model encompasses a context-sensitivity in some way or not.

- **Justifiable (denoted as "Justif"):** we mean by this criterion "justifiable and understandable". Indeed, even if a machine will not be able to apprehend the counter-factuality at the origin of decision making for a while, it remains important for the machine's behavior to be able to inform the operator/user with an understandable justification (Monroe and Malle, 2018; Scheutz, 2017). Therefore, even though the majority of the models do not tackle this issue, it is possible to make a distinction between the approaches that keep some explicit representation of ethical concepts or principles, which allows to understand the automated reasoning behind the decision, and the ones that do not, such as

abstracted numerical approaches. Therefore, this criterion will be in some cases related to the "number-based approach" criterion, as several of the approaches using numbers lose explicit representation. It is worth noticing that the question of "understandable for whom ?" is not discussed here. Nonetheless, it remains that taking into account the kind of people targeted by the justification is a clue for an efficient justification.

- **Avoidance (denoted as "Avoid"):** some classical approaches such as game theory or planning are based on the action-consequence principle. Hence, these approaches narrow the information representing the situation down to actions performed by the agent and consequences of these actions. In the case of ethical reasoning, this fails to encompass the importance of avoiding consequences, and even non-performing some actions (see paragraph 3). "avoidance" depicts the property of the model to include non-action/avoidance of consequences.

Table 2.2 Summary table of model properties

Model	num	implem	Cont	Justif	Avoid	Subj	In
Top-down							
(Berreby et al., 2017)	not num*	×	×	×			
(Bringsjord and Taylor, 2011)	not num	×		×	×		
(Cointe et al., 2016)	not num*	×	×	×			
(Atkinson and Bench-Capon, 2016)	high num	×	×	×			
(Bonnemains and al., 2019)	not num	×	×	×	×	×	
Bottom-up							
(Anderson, 2015)	high num	×		×			
(Kuipers, 2018)	high num**		×	×			
(Abel et al., 2016)	high num	×	****				
Hybrid							
(Arkin, 2007)	not num		×	×	×		
(Conitzer et al., 2017)	high num						
(Sarathy et al., 2017)	num***	×	×	×			
(Govindarajulu et al., 2019)	num***	×	×	×			

Where "×" means that the model fits the property, and " " means that it doesn't

These criteria are subject to discussion. For instance, it is hard to assess some properties of an approach when this approach is not implemented. Here are some specific criticisms about the table above (referring to "*" in the table 2.2):

- * The works of Berreby et al. and Cointe et al. use either (total) orders or numerical assessments as a part of the reasoning to make a decision. Nonetheless, numerical tools are not the main contribution of these works but examples of modular tools within the model that could be replaced.
- ** As it was said earlier, the model of Kuipers is not implemented and thus, some application interpretations might differ according to different points of view.
- *** Both Sarathy et al. and Govindarajulu et al. use numerical tools as the core of a method to learn ethical rules/norms. Nonetheless, ethical concepts themselves remain meaningful.
- **** Abel et al.'s work is based on a POMDP, and thus on states and transitions. Therefore, it might be possible to consider a context-sensitivity depicted through specific states and transitions. Nonetheless, it is not tackled in the article.

2.7 Conclusion

In this chapter, we have first addressed the question of the distinction between moral and ethics, arguing for an assimilation of moral as "theory of the good" and ethics as "theory of the right". The second step has been dedicated to the definition of the three main fields of ethics: normative ethics, metaethics and applied ethics. Then, we have presented some categories of ethical frameworks of normative ethics. After that, moral dilemmas have been presented as very useful tools to apply, question and improve ethical theories. The fifth part has aimed at defending two important ideas. First, autonomous machines will require ethical considerations in there reasoning in a near future as the frontier between their working closed environment (i.e. safety areas in industries) and the real-world (houses, streets, roads, etc.) is more and more permeable, and thus, they may have an important impact on our lives. Secondly, autonomous machines will not be fully-ethical agents as this is a property of living/human-being. Hence, they should not be called "ethical [agents/machines/etc.]". The final part has presented several promising works to model ethics through a classic classification: top-down/bottom-up/hybrid approaches. After that, a list of properties has been

proposed to analyse the works: number-based approaches, implemented approaches, context-sensitivity, justification compatibility and avoidance.

We can see in this final analysis that very few approaches fulfill the criterion of avoidance. Nonetheless, while speaking of ethics, this should be a major consideration as non-action might lead to event inducing consequences that has to be taken into account to judge a decision even if the agent did not act. Our contribution that is detailed in the next chapters propose a model tackling this issue through a non-numerical approach using ethical frameworks implemented on moral dilemmas thanks to a formalism based on decisions, events and states of the world composed of facts, allowing context-sensitivity and justification. Moreover, as this approach aims at being embedded in a human-machine system, paragraph 6 will present the results of an experiment to study the impact of an agent embedding ethical considerations on a human who has to make a decision facing a moral dilemma.

Chapter 3

Contribution: the formalism

3.1 The approach

Our contribution aims at defining and computing ethical judgements of decisions in a situation of a moral dilemma. What mainly differentiates our work from previous works is the design approach. Indeed, as a first step, we have started philosophical concepts of ethics in order to identify the key concepts that would be likely to embed ethical considerations in an automatic reasoning, contrary to the wide majority of works mentioned previously which mainly try to improve existing computer science tools to fit ethics. We decided to start from scratch in order to avoid any initial bias such as simplifying or digitizing ethics. Hence, the situation representation (paragraph 3.3) includes a property seldom exploited in the literature, "avoidance" (see paragraph 2.6.4). Indeed, we support the idea that ethical reasoning is based on obtained consequences as well as avoided consequences through events that are not always linked to the agent's actions. For instance, in the trolley and the footbridge dilemmas, the five people will be hit by the trolley and will die if no action is performed. Hence, whereas a part of previous approaches put the agent's actions at the core of the model, we assume the environment as evolving with events, which allows the agent to reason over decisions resulting in non-actions, but still having an impact and modifying the environment (part 3.3). Thereafter, we assume normative ethics as best fitting our needs to judge decisions in order to guide decision making. Therefore, as some previous works, we formalise ethical frameworks of normative ethics (paragraph 3.4). The second step of our work (paragraph 4), which is not so common in literature consists in criticizing ethics formalisation in general as well as our approach by challenging our models of ethical frameworks with moral dilemmas. This allows us to highlight several difficulties when modelling ethics: indeed, as a machine cannot access counter-factuality, reasoning

must be based on assessments and preferences given by the designers, embedding their subjectivities[1]. Moreover, using a single ethical framework to judge decisions is not sufficient (see paragraph 4). Nonetheless, the combination of several ethical frameworks does not allow a decision to be made in any situation. Furthermore, this combination is questionable as the aggregation of judgements from different ethical frameworks alters their specificities.

These difficulties are in line with the idea that machines cannot be ethical agents alone (see paragraph 2.5), which justifies the subsequent part 6 that aims at studying works about human-machine collaboration in the context of moral decision making. Most of the works cited previously study situations in which machines have to deal with ethical issues. But, these situations are a small parts of all situations encountered by a machine. Therefore, ethical reasoning is necessary for only few cases and first requires to identify a situation as a moral dilemma. This remains relevant in the context of a human-machine system in order to start a decision-making process or to notify the operator of such situation. Nonetheless, as far as we know, this issue has never been dealt with.

It is worth noticing that parts 3.3, 3.4 and 4 have been the subject of three papers (Bonnemains et al., 2016, 2018, 2017). Part 5 is the subject of a submitted article (Bonnemains et al., 2019).

3.2 Context and assumptions

Let us remind that the context of our work is an operator-artificial agent system, where the artificial agent embeds ethical knowledge allowing it to compute judgement about possible decisions and justify its judgement to the operator. Indeed, we have argued in the previous chapter that machines cannot be ethical agents (see paragraph 2.5). Therefore, they should not be placed alone in contexts where ethical thoughts are required to make decisions.

The aim of defining a formalism to model ethical knowledge is multi-faceted:

- identify the main concepts that are required to judge decisions ethically
- reduce the ambiguity of natural language by giving precise definitions of ethical concepts
- propose a model allowing automatic computing methods to ethically assess decisions

[1]This rises the question of "who should make these value judgements ?"

- highlight the possible subjectivity sources in such a model
- analyse the strengths and weaknesses of modelled ethical frameworks inspired by normative ethics

Moreover, as the approach aims at "keep[ing] some explicit representation of ethical concepts or principles which allows to understand the automated reasoning behind the decision"(from Chapter 2), we have decided to avoid numbers as much as possible in our formalism.

As modelling ethics remains a tough work, it requires to start with assumptions in order to define a reduced working context. Hence we will assume that:

- the agent decides and acts in a complex dynamic world;
- the situation has been previously identified as a moral dilemma;
- the moral dilemma is considered from the agent's point of view;
- the agent has to make a decision among several possible decisions;
- the agent knows all the possible decisions and all the effects of each decision;
- terms such as right/wrong and positive/negative are defined from the agent's point of view: a decision is right if it meets the agent's moral values; a wrong decision disregards them; a fact is positive if it is beneficial for the agent; it is negative if it is undesirable for the agent.

Finally, as some dilemmas involve human lives, we assume that:

- a human life is perfectly equal to another human life, whoever the human being might be[2]

We then need to identify the very basic concepts allowing to represent a situation.

3.3 Situation model

Goal: *build a situation model allowing to compute judgements on possible decisions, and encompassing the notion of avoidance of facts and non-action*

[2]We make this strong assumption in order to avoid additional ethical concerns about judging and comparing values of lives.

3.3.1 Normative ethical frameworks

Considering the three fields of ethics (paragraph 2.2), normative ethics fits best our purpose as it aims at judging an action or a decision. Hence, ethical frameworks of normative ethics (2.3) judge decisions depending on different clues. Indeed, Consequentialism (paragraph 2.3.2) bases the judgement on consequences while deontology mainly consists in judging the decision, whatever are the consequences, see figure 3.1.

(a) Consequentialism (b) Deontology

Fig. 3.1 Ethical frameworks illustration

Virtue ethics is not considered here as it is neither based on decision/event nor consequences (paragraph 2.3.4). Moreover, identifying the relevant virtues for machines and how they apply is far from simple. Indeed, it is not clear how artificial agents would fulfil virtues through their behaviours (Aristotle, 2012; Nussbaum, 2012). Furthermore, it appears that some virtues are not relevant for machines (such as humility, courage, etc.) while others may be virtues for machines but not for humans, such as confidentiality and availability (Sullins, 2014).

On the contrary the Doctrine of Double Effect, which is not categorized as an ethical framework of normative ethics will be studied in this work in line with the works of Berreby et al. (2017); Cointe et al. (2016). Indeed, this doctrine proposes a judgement that is based both on decision and consequences using specific concepts such as causality and proportionality which are interesting properties and tools to model and analyse.

All along the description of our formalism, we will use the footbridge dilemma as a running example:

"[...] you are standing on a footbridge over the trolley track. You can see a trolley hurtling down the track, out of control. You turn around to see where the trolley is headed, and there are five workmen on the track where it exists from under the footbridge. What to do? Being an expert on trolleys, you know of one certain way to stop an out-of-control trolley: Drop a really heavy weight in its path. But where to find one? It just so happens that standing next to you on the footbridge is a fat man, a really fat man. He is leaning over the railing, watching the trolley; all you have to do

is to give him a little shove, and over the railing he will go, onto the track in the path of the trolley.[...]" from Thomson (1986)

3.3.2 Concepts at a glance

This section aims at giving an overview of the concepts at stake in the model using Footbridge example. These concepts will be detailed thereafter.

The first concept is the current ***state of the world***:

$$i = [fat, f_5] \tag{3.1}$$

This state is composed of two ***facts***, fat meaning that fatman is alive, and f_5 meaning that five people are alive. Facing this state of the world, the agent has to choose between two possible ***decisions***:

- d_1 = push fatman
- d_2 = do nothing

Each of these decisions is linked to an ***event***. These events are computed through function $Event$:

$$Event(d_1) = e_1(\text{trolley hits "fatman"}) \tag{3.2}$$
$$Event(d_2) = e_2(\text{trolley hits five people}) \tag{3.3}$$

These events modify some values of facts, resulting in new states of the world called ***effects***, that represents the consequences of a decision. Therefore, a new state of the world can be computed from the current state for each possible decision through function $Consequence$:

$$Consequence(e_1, i) = [\overset{\circ}{fat}, f_5] = s_1 \tag{3.4}$$
$$Consequence(e_2, i) = [fat, \overset{\circ}{f_5}] = s_2 \tag{3.5}$$

with $\overset{\circ}{fat}$ meaning that fatman is dead and $\overset{\circ}{f_5}$ meaning that five people are dead. These example is illustrated in figure 3.2.

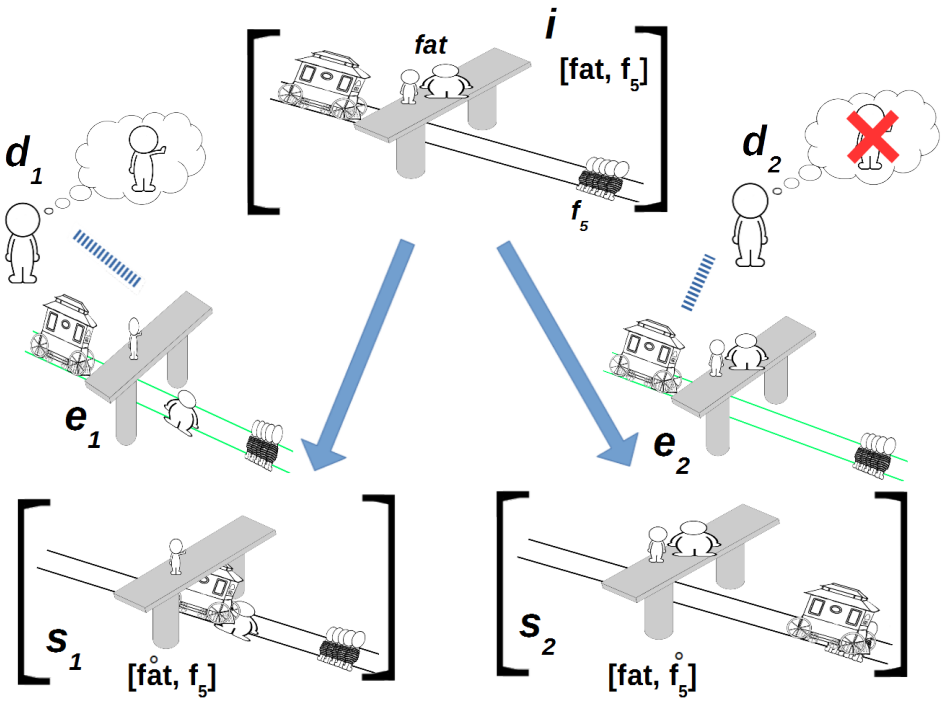


Fig. 3.2 Fatman dilemma: world states, decisions and events
little man: [https://www.edupics.com/coloring-page-figurine-empty-face-s25691.jpg]

3.3.3 Fact and state of the world

As we wish to model the situation, we need to define such concept. A situation is thus defined as:

$$\sigma =< i, \mathcal{D} > \tag{3.6}$$

Let us assume:

Definition 1 (*Situation-Set* Σ) A situation is composed of a current world state and a set of possible decisions in order to compute the effects.

This concept will not be used in this chapter to simplify examples. Nonetheless, this concept is necessary to the dilemma identification of chapter 5.

We define a state of the world as a vector of facts:

Definition 2 (*World state-Set* $\mathcal{S}$) A *world state* is a vector of facts (see definition below). Let $\mathcal{S}$ be the set of world states.

Definition 3 (*Fact-Set* $\mathcal{F}$) A *fact* is a variable that can be instantiated with values within a binary set. We consider these values as two values regarding the same

item. An item can be an object (or several objects), a living being (or several living beings), or anything else that needs to be taken into account by the agent when dealing with the dilemma. Let $\mathcal{F}$ be the set of facts.

Notation: "$\overset{\circ}{f}$" is the notation used to differentiate one value from the other for fact f. For the footbridge dilemma, we choose to define two facts as:

- f_5 = five people are alive
- $\overset{\circ}{f_5}$ = five people are dead
- fat = fatman is alive
- $\overset{\circ}{fat}$ = fatman is dead

This choice of modelling allows a wide range of possible representations, from gain/loss, to gain/no gain or loss/no loss. Moreover, world states of a situation have all the same dimension (i.e. number of facts). This explicitly engages to take into account either a value of fact or its antagonist value from one decision to another. The main idea here is to improve the representation completeness. Indeed, it is important while judging to be exhaustive about the values of facts that are modified by an event as well as the values of facts that are not modified by this event. For the footbridge dilemma, "fatman dead" is as much important as "five people [still] alive".

3.3.4 Decision and event

Many paradigms such as situation calculus (Reiter, 1993) or Cointe's BDI model (Cointe et al., 2016) put the agent's actions at the core of the world changes. Nonetheless, the agent is in a dynamic world where facts and other agents can evolve independently of the agent's action. Furthermore, the very question of "is acting better than doing nothing ?" has to be asked for ethical purposes. In this regard, we will not consider the agent's actions but its *decisions* and the resulting *events*.

Definition 4 (*Decision-Set* $\mathcal{D}$) A *decision* is a choice of the agent to do something, i.e. to perform an action, or to do nothing and let the world evolve. Let $\mathcal{D}$ be the set of decisions.

These decisions do not have themselves an impact on the world state, but when the agent makes a decision, it results in an *event* that may modify the world state.

Definition 5 (*Event-Set* $\mathcal{E}$) An *event* is something that happens in the world that modifies the world state, i.e. some facts. Let $\mathcal{E}$ be the set of events.

Differentiating *decision* from *event* allows us to take into account "non-action" such as letting something happen, or deciding not to perform an action.

Let *Event* be the function computing the event from a decision:

$$Event \ : \ \mathcal{D} \ \rightarrow \ \mathcal{E} \tag{3.7}$$

For the footbridge dilemma, decision *do nothing* results in event $trolley\ hits\ five\ people$:

$$Event(d_2) = e_2 \tag{3.8}$$

3.3.5 Consequences/Effect

As it was said earlier, an event may modify the current world state by changing values of facts. The new world state, which is the consequence of the event, is called the *effect*.

Definition 6 (*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 of an event on the current world state:

$$Consequence \ : \ \mathcal{E} \ \times \ \mathcal{S} \ \rightarrow \ \mathcal{S} \tag{3.9}$$

In the footbridge dilemma, if the agent makes the decision $push\ fatman$, the $trolley\ hits\ fatman$ (event) and thus fatman is dead ($\overset{\circ}{fat}$) while five people are alive (f_5) (effect).

$$fat, f_5 \in \mathcal{F} \tag{3.10}$$

$$push\ fatman\ =\ d_1 \in \mathcal{D} \tag{3.11}$$

$$trolley\ hits\ fatman\ =\ e_1 \in \mathcal{E} \tag{3.12}$$

$$Event(d_1) = e_1 \tag{3.13}$$

$$i = [fat, f_5] \tag{3.14}$$

$$Consequence(e_1, i) = [\overset{\circ}{fat}, f_5] \tag{3.15}$$

Even if we start from scratch, our model shares similarities with other approaches such as events of the event calculus and fluents of the situation calculus which are not so far from our facts (Kowalski and Sergot, 1989; Reiter, 1993). Nonetheless, situation calculus describe the world through series of agen's actions while we need our agent being able to reason over non-actions (i.e. when the world evolves without an agent's action). Considering event calculus, even if the events of event calculus modify fluents just as our events modify facts, we need these events to be part of a couple decision-event, in order to allow our agent to reason over decisions and not over events. Moreover, ethical judgement requires facts with more complex values than fluents with only true or false values.

3.4 Decision judgement from normative ethics

Goal: *Formalize Decision judgement methods*

This part is dedicated to the formalization of judgement methods from ethical frameworks. It is worth noticing that these methods stem from our own interpretation of ethical frameworks. Indeed, as it was described in paragraph 2.3, there is no consensus about the way to interpret ethical frameworks and there are many different judgement methods. Therefore our approach, which embeds several assumptions, is an illustration of how ethical frameworks could be translated into an implementable formalization.

3.4.1 Judgement

As the model aims at judging a decision, we define the possible judgement values for decisions as:

$$\mathcal{V} = \{acceptable(\top), undetermined(?), unacceptable(\bot)\} \tag{3.16}$$

Hence, each ethical framework will issue a judgement on a decision through an assessment of conditions to satisfy. These conditions represent the principles of the ethical frameworks. If conditions are satisfied, the decision will be assessed as *acceptable*, meaning that the decision is ethically "right" considering this ethical framework. On the other side, a decision will be *unacceptable* if it does not satisfy the conditions, which means that the decision is ethically "wrong" considering the ethical framework. Sometimes, it might not be possible to assess whether a condition is satisfied or not, either because of a lack of information or because a required value cannot be assessed. In this case, a decision will be assessed as *undetermined*.

Let us define function *Judgement* with a single signature regardless of the ethical framework:

$$Judgement : \mathcal{D} \times \mathcal{S} \rightarrow \mathcal{V} \tag{3.17}$$

This function requires a decision to judge and a current world state that represents the context in which the decision has to be judged.

In order to differentiate the judgement of an ethical framework from another one, the function *Judgement* is indexed by the ethical framework the judgement is based on.

For the footbridge dilemma:

$$Judgement(d_1, i)_d = \bot \tag{3.18}$$

means that decision d_1 (push fatman) is ethically judged as "*unacceptable*" considering deontological point of view represented here as index d[3].

3.4.2 Consequentialism

As we have seen before, *"the end justifies the means"* from a consequentialist point of view (see paragraph 2.3.2). Even if this definition is controversial, philosophers agree that the set of consequentialist frameworks (egoism, altruism, etc.) focus their

[3]In this paper, indexed are u for consequentialism implemented with utilitarian rules, d for deontology and *dde* for the Doctrine of Double Effect.

judgements on the consequences of decisions. Considering our model, it means that the conditions of a consequentialist judgement are about the $Effect$.

Because most artificial agents are designed to be helpful rather than selfish, we have decided to model a consequentialist framework with a combination of positive and negative utilitarianism (see paragraph 2.3.2). According to this framework, the agent will try to make the best possible decision, meaning in this case the best effect or the least bad effect.

Therefore, the main issue considering this principle of "best effect" lies on being able to compare effects of several events related to decisions, i.e., to compare set of facts.

Using numeric values seems particularly attractive for this purpose (Berreby et al., 2017). Indeed the trolley and the footbridge dilemmas tend to encourage this approach: from a consequentialist point of view, it is admitted that **five** people alive is better than **one** person alive. Nonetheless, some more complex examples, such as the one we present in paragraph 4.2, highlight that using numeric values to assess facts is tough if not irrelevant. We argue that some facts must remain incomparable.

Therefore our idea is, considering two effects, to arrange facts of these effects into two categories: *positive* facts and *negative* facts. Afterwards, preferences will be computed between the set of positive (resp. negative) facts of one effect and the set of positive (resp. negative) facts of another effect, in order to determine which effect is preferred to the other. This approach aims at representing the "best effect" with positive facts and the "least bad effect" with negative facts.

Positive facts and negative facts

Let $Positive$ and $Negative$ be the functions allowing to categorize facts:

$$Positive/Negative : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{F}) \tag{3.19}$$

For both functions, the argument is a world state computed from an event (i.e., an effect). These functions return a subset of facts of this effect evaluated as positive (resp. negative). We will assume that:

$$\forall s \in \mathcal{S},\ Positive(s) \cap\ Negative(s) = \emptyset \tag{3.20}$$

$$\forall f_i \in \mathcal{F}, \forall s \in \mathcal{S}\ /\ s = [f_1, ...f_i, ...f_n],\ f_i \in\ (Positive(s) \cup\ Negative(s)) \tag{3.21}$$

meaning that for a given situation, one fact is either positive or negative (3.21) but cannot be both (3.20) from the agent's view point. Indeed, as we assume world states composed of *relevant* facts to make a decision, we assume that a fact cannot be "neutral" regarding the decision.

For the footbridge dilemma:

$$Consequence(\text{trolley hits five people}, i) = [fat, \overset{\circ}{f_5}] \tag{3.22}$$

$$Negative([fat, \overset{\circ}{f_5}]) = \{\overset{\circ}{f_5}\} \tag{3.23}$$

$$Positive([fat, \overset{\circ}{f_5}]) = \{fat\} \tag{3.24}$$

$$Consequence(\text{trolley hits fatman}, i) = [\overset{\circ}{fat}, f_5] \tag{3.25}$$

$$Negative([\overset{\circ}{fat}, f_5]) = \{\overset{\circ}{fat}\} \tag{3.26}$$

$$Positive([\overset{\circ}{fat}, f_5]) = \{f_5\} \tag{3.27}$$

Preference on facts

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

$$f_a >_u f_b,\ f_a, f_b \in \mathcal{F} \tag{3.28}$$

means that f_a is preferred to f_b from a consequentialist[4] viewpoint. Therefore, $f_a >_u f_b$ considering positive facts means that f_a is better than f_b, while considering negative facts it means that f_a is less bad than f_b. Moreover, it is worth noticing that by definition of positive/negative fact, a positive fact is always preferred to a negative fact.

We assume the following properties of $>_u$:

1. $f_a >_u f_b \rightarrow \neg(f_b >_u f_a)$ (***asymmetry***): if f_a is preferred to f_b ,f_b is never preferred to f_a . For instance, if peace is preferred to war, there is no way to prefer war to peace at the same time.

2. $[(f_a >_u f_b) \wedge (f_b >_u f_c)] \rightarrow f_a >_u f_c$ (***transitivity***): if f_a is preferred to f_b and f_b is preferred to f_c, then f_a is preferred to f_c. Indeed, we assume that if you prefer saving lives to protecting materials and protecting materials to being polite, you are likely to prefer saving lives to being polite.

3. $\nexists f_i \ / \ f_i >_u f_i$ (***irreflexivity***): a fact cannot be preferred to itself.

[4]by consequentialist we mean what we have defined previously with subjective choices of implementation.

From these properties, $>_u$ appears to be a strict order.
This relation needs to be extended in order to compare set of facts. Hence, we extend $>_u$ to $\succ_u$, which denotes the preference relation on subsets of facts:

$$F_a \succ_u F_b \tag{3.29}$$

means that F_a is preferred to F_b from a consequentialist viewpoint. As an aggregation criterion has to be defined in order to realize that extension, some criteria will be proposed in paragraph 3.4.5. It is worth noticing that depending on the criterion or criteria used, properties previously defined might not be kept from $>_u$ to $\succ_u$.

For the footbridge dilemma, we assume preferring "five people alive (f_5)" to "fatman alive (fat)" and "fatman dead ($\overset{\circ}{fat}$)" to "five people dead ($\overset{\circ}{f_5}$)":

$$f_5 >_u fat \tag{3.30}$$

$$\overset{\circ}{fat} >_u \overset{\circ}{f_5} \tag{3.31}$$

Then 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). Therefore we propose the concept of field as a solution to this issue, the purpose being to define a preference relation on fields. Hence it will allow a fact to be preferred to another fact depending on their belonging to on field or another.

Preference on fields

The preference on fields is proposed as an answer to the issue of incomparable facts. The idea is that facts belongs to main fields. For instance, facts related to protect people belong to the field of "humans" while facts related to protect material belong to the field of "goods". If we assume that the field of "humans" is preferred to the field of "goods", it means that when facts are incomparable, facts belonging to humans field will be preferred to facts belonging to goods field.

Definition 7 (*Field-Set* Φ) Let Φ be the set of fields. Function *Field* allows to get the field associated with a fact.

$$Field : \mathcal{F} \rightarrow \Phi \tag{3.32}$$

Let $>_\Phi$ the preference relation on Φ. It means that:

$$\forall \Phi_a, \Phi_b \in \Phi, \forall s \in \mathcal{S},\ \Phi_a >_\Phi \Phi_b$$

$$\forall f_a, f_b \in Positive(s),\ f_a \in \Phi_a,\ f_b \in \Phi_b \rightarrow f_a >_u f_b \tag{3.33}$$

$$\forall f_a, f_b \in Negative(s),\ f_a \in \Phi_a,\ f_b \in \Phi_b \rightarrow f_b >_u f_a \tag{3.34}$$

Indeed, if Φ_a is preferred to Φ_b, the agent will prefer a negative fact f_b ($f_b \in \Phi_b$) to a negative f_a ($f_a \in \Phi_a$) as the agent desires to avoid the worst fact, i.e., the fact belonging the most preferred field. Hence, the more a field Φ_i is preferred, the more negative facts belonging to it have to be avoided.

The use of field preferences is an answer to incomparable facts. Therefore, it has to be exploited as a lexicographic preference[5]: when facts are not comparable initially, the agent uses fields preference to compare these facts.

It is worth noticing that the use of field preference is linked to a very strong assumption with many drawbacks. Indeed, it means that facts are unconditionally preferred depending on their field affiliation without considering their "strength": a person at end-of-life will always be preferred to the whole world heritage.

Judgement function

We have now the tools required to define the conditions allowing a decision to be judged "acceptable" or "unacceptable" from a consequentialist view point. Hence, we assume a decision d_a involving event e_a (i.e., $Event(d_a) = e_a$) as "better" than decision d_b involving event e_b ($Event(d_b) = e_b$) according to the consequentialist framework iff for $i \in (S)$:

$$Positive(Consequence(e_a, i)) \succ_u Positive(Consequence(e_b, i)) \tag{3.35}$$

$$\wedge\ Negative(Consequence(e_a, i)) \succ_u Negative(Consequence(e_b, i)) \tag{3.36}$$

This condition combines the utilitarian principles defined earlier:

- obtain "the best effect", i.e., positive utilitarianism (3.35)
- obtain "the least bad effect", i.e., negative utilitarianism (3.36)

[5]lexicographic preference consists in pursuing comparison character to character until an inequality allowing to order the two words

Therefore, if this condition is satisfied, we consider that decision d_a is the "best"[6] decision:

$$Judgement_u(d_a, i) = \top \tag{3.37}$$
$$Judgement_u(d_b, i) = \bot \tag{3.38}$$

If the condition is not satisfied, the "best" decision cannot be identified from a consequentialist viewpoint:

$$Judgement_u(d_a, i) = Judgement_u(d_b, i) = \ ? \tag{3.39}$$

Considering a situation with more than two possible decisions, identifying the "best" decision requires to compare each decision to all the others in order to find the "best" effect. If such a decision does not exist, there is no decision judged as acceptable from a consequentialist view point. Nonetheless, if one decision d_a is assessed as "better" than d_b, it is possible to compute that $Judgement_u(d_b) = \bot$.

For the footbridge dilemma, preferences 3.30 and 3.31 result in decision "push fatman" (d_1) being the "best" decision.

$$Judgement_u(d_1, i) = \top \tag{3.40}$$
$$Judgement_u(d_2, i) = \bot \tag{3.41}$$

3.4.3 Deontology

As opposed to consequentialism, this framework focuses on the decision whatever the consequences are. Indeed, such as the Categorical imperative, deontology lies on the nature of the decision, which must be not bad according to some principles.

Nature of decision

A decision in itself may have a good, bad or neutral nature from the agent's point of view. For instance, "to lie" can be considered as bad while "to tell the truth" can be considered as good, and "to do nothing" as neutral, independently from consequences.

Let $\mathcal{N}$ be the set of decision nature values:

$$\mathcal{N} = \{bad, neutral, good\} \tag{3.42}$$

[6] not ethically best, but consequentially best

with function $DecisionNature$ returning the decision nature value of a decision:

$$DecisionNature : \mathcal{D} \rightarrow \mathcal{N} \tag{3.43}$$

For the footbridge dilemma, we will consider that:

$$DecisionNature(d_1) = bad \tag{3.44}$$
$$DecisionNature(d_2) = neutral \tag{3.45}$$

Judgement function

Hence, a decision is "acceptable" from deontological view point if the nature of the decision is "at least" neutral (i.e. neutral or good). Otherwise the decision is "unacceptable". $\forall d \in \mathcal{D}, \forall i \in \mathcal{S}$:

$$DecisionNature(d) = neutral \vee good \Rightarrow Judgement_d(d,i) = \top \tag{3.46}$$
$$DecisionNature(d) = bad \Rightarrow Judgement_d(d,i) = \bot \tag{3.47}$$

For the footbridge dilemma:

$$Judgement_d(d_1,i) = \bot \tag{3.48}$$
$$Judgement_d(d_2,i) = \top \tag{3.49}$$

3.4.4 Doctrine of Double Effect

The interest of the Doctrine of Double effect lies on its ability to sometimes discriminate between decisions when deontology and consequentialism cannot, through a combination of conditions focusing both on the decision itself and its consequences (i.e. effects). For instance, from a consequentialist viewpoint, there is no difference between decisions "push the switch" and "push fatman" considering respectively the trolley dilemma and the footbridge dilemma. Nevertheless the rules presented in paragraph 2.3.5 have to be adapted to an artificial agent.

DDE's rules for an artificial agent

1. *" The act itself must be morally good or at least indifferent."*: this rule is similar to the condition of deontology, let us call it "deontological rule".

2. *"The agent may not positively will the bad effect but may permit it. If he could attain the good effect without the bad effect he should do so. The bad effect is sometimes said to be indirectly voluntary."*: as we assume the artificial agent to be designed correctly, it seems unlikely that it will be implemented as desiring bad effects. Indeed, if an effect is assessed as bad from the agent's point of view, we assume the agent as not desiring this effect. Through this assumption, this rule is satisfied in any situation, it will then not be mentioned further.

3. *"The good effect must flow from the action at least as immediately (in the order of causality, though not necessarily in the order of time) as the bad effect. In other words the good effect must be produced directly by the action, not by the bad effect. Otherwise the agent would be using a bad means to a good end, which is never allowed."*: this rule introduces the complex notion of causality between facts. It can be transcribed as "Negative facts must be neither an end nor a mean". Let us call this rule the "collateral damage rule".

4. *"The good effect must be sufficiently desirable to compensate for the allowing of the bad effect."*: in other words, the set of negative facts has to be proportional to the set of positive facts. Let us call this rule the "proportionality rule".

Therefore, the condition to respect the DDE can be splitted into the three following rules:

Deontological rule

We already have defined a condition for deontology. Therefore, the deontological rule amounts to rules 3.46 and 3.47.

$$DecisionNature(d) = neutral \vee good \Rightarrow Judgement_d(d,i) = \top \tag{3.50}$$

$$DecisionNature(d) = bad \Rightarrow Judgement_d(d,i) = \bot \tag{3.51}$$

Collateral damage rule

As we have said, we need to define the causality between facts. In order to do that, we have chosen the Linear Temporal Logic modal operator F (i.e., *"Finally"* which means: eventually in the future) from Pnueli (1977):

$$p \vdash Fq \tag{3.52}$$

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

It is worth noticing that phrase "induce" may have several meaning. Indeed, such as for the footbridge dilemma, the death of fatman allows to protect/to keep "five people alive" from the initial world state. On the other hand, "induce" may mean to turn fact from one value to another. In the same vein, this definition does not cover situations where a fact might be induced by several other facts[7].

The collateral damage rule can thus be translated as: considering an initial world state i, a decision d with $Event(d) = e$ and $Consequence(e, i) = s$

$$\forall f_n \in Negative(s), \nexists f_p \in Positive(s),\ f_n \vdash F f_p \tag{3.53}$$

For the footbridge dilemma, the death of fatman (negative fact) is a way to induce that five people are alive (positive fact):

$$\overset{\circ}{fat} \vdash F f_5 \tag{3.54}$$

which means that the collateral damage rule (3.53) is violated.

Proportionality rule

The notion of proportionality has several meanings:

- It is a mathematical property between two number series, one being equal to the other multiplied by a coefficient.
- In the same vein, proportionality is used while considering electoral systems (Gallagher, 1991).
- Another meaning concerns the laws of war (Gardam, 1993).

The concept that is at stake in the DDE is not so far from the last meaning. Indeed, there is no mathematical concern here, but a question of "acceptance" of facts.

We have decided to use the consequentialist preference in order to define the DDE's proportionality. We assume that "negative facts proportional to positive facts" conveys the idea that positive facts are desired and thus negative facts are tolerated in order to have these positive facts. In other words "positive facts and negative facts are preferred

[7] either because several facts are required or because different facts can induce the same fact

to non-obtaining positive facts".

In order to define proportionality through our formalism, we assume that "non-obtaining" amounts to having for each fact of the set of positive facts the other binary value. For the footbridge dilemma, non-obtaining fat amounts to $\overset{\circ}{fat}$, as well as non-obtaining $\overset{\circ}{fat}$ amounts to fat. Hence, let $\overset{\circ}{F}$ denote the set of non-obtained facts corresponding to the facts of set F.

$$F = \{f_1, f_2, \overset{\circ}{f_3}\} \Leftrightarrow \overset{\circ}{F} = \{\overset{\circ}{f_1}, \overset{\circ}{f_2}, f_3\} \tag{3.55}$$

Therefore, the proportionality rule can be described as such: for an initial world state i, a decision d with $Event(d) = e$ and $Consequence(e, i) = s$

$$\begin{aligned} Positive(s) = F_p,\ Negative(s) = F_n \\ F_p \cup F_n \succ_u \overset{\circ}{F_p} \end{aligned} \tag{3.56}$$

For the footbridge dilemma considering decision "push fatman", we then have to compute whether:

$$\{f_5, \overset{\circ}{fat}\} \succ_u \{\overset{\circ}{f_5}\} \tag{3.57}$$

It is worth noticing that this definition aims at improving the previous one given in Bonnemains et al. (2018). Nonetheless, the phrase "proportional" is in natural language related to a notion of "excessiveness" (a is not proportional to b if a is excessive regarding b). Our formal definition does not entirely cover "proportionality" concept then, but a consequence of proportionality when applied to the Doctrine of Double Effect. Nonetheless, we argue that it is reasonable to imagine that preferring $F_p \cup F_n$ to $\overset{\circ}{F}_p$ induces that F_n is not excessive considering F_p.

Judgement function

A decision is thus judged "acceptable" from the DDE's view point if it satisfies the conditions described above. Otherwise, the decision is judged "unacceptable": for an initial world state i and a decision d with $Event(d) = e$ and $Consequence(e, i) = s$ with

$Positive(s) = F_s$ and $Negative(s) = F_n$

$$\begin{aligned} & DecisionNature(d) = neutral \vee good \\ & \wedge\ \forall f_n \in F_n, \nexists f_p \in F_p,\ f_n \vdash F f_p \\ & \wedge\ F_p \cup F_n \succ_u \overset{\circ}{F_p} \\ & \Rightarrow Judgement_{dde}(d, i) = \top \end{aligned} \tag{3.58}$$

For the footbridge dilemma, "push fatman" respects neither the deontological rule, indeed:

$$DecisionNature(\text{push fatman}) = bad \tag{3.59}$$

nor the collateral damage rule as:

$$\overset{\circ}{fat} \vdash F f_5 \tag{3.60}$$

Therefore:

$$Judgement_{dde}(d_1, i) = \bot \tag{3.61}$$

Supposing that d_1 would have respected the previous rules, it would have been necessary to compute whether:

$$\{f_5, \overset{\circ}{fat}\} \succ_u \{\overset{\circ}{f_5}\} \tag{3.62}$$

We can notice that this rule requires to be able to extend $>_u$ to $\succ_u$. In order to do so, we propose the use of aggregation criteria, some of them being defined in the next section.

3.4.5 Aggregation critera

Aggregation criteria are necessary to extend $>_u$ to $\succ_u$.

In order to propose a relevant example, we will consider the footbridge dilemma with f_5 and fat as described earlier with a supplementary fact:

- $nmurd$: not become a murderer (a positive fact belonging to the effect of decision $do\ nothing$)
- $\overset{\circ}{nmurd}$: become a murderer (a negative fact belonging to the effect of decision $push\ fatman$)

Hence, considering decision "push fatman":

$$Consequence(e_1, i) = [\overset{\circ}{fat}, f_5, \overset{\circ}{nmurd}] \tag{3.63}$$

For the sake of the example, we assume that:

- $f_5 >_u fat$
- $nmurd >_u f_5$

Each For One criterion (EFO)

Relation $>_u$ is extended to $\succ_u$ as following: considering two sets of facts F and G, $F \succ_{u-efo} G$ if for each $f \in F$, it exists a fact $g \in G$ such that $f >_u g$.

$$F \succ_{u-efo} G \text{ iff } \forall f \in F, \exists g \in G \ / \ f >_u g \tag{3.64}$$

For the consequentialist preference relation:

$$\{f_5\} \succ_{u-efo} \{nmurd, fat\} \tag{3.65}$$

as f_5 preferred to fat is sufficient for the EFO criterion.

For proportionality, we need to compute whether:

$$\{\overset{\circ}{nmurd}, \overset{\circ}{fat}, f_5\} \succ_{u-efo} \{\overset{\circ}{f_5}\} \tag{3.66}$$

Nonetheless:

$$\nexists g \in \{\overset{\circ}{f_5}\} \ / \ \overset{\circ}{nmurd} >_u g \tag{3.67}$$

Therefore the proportionality rule is violated in this example considering the EFO criterion.

One For Each criterion (OFE)

Relation $>_u$ is extended to $\succ_u$ as following: considering two sets of facts F and G, $F \succ_{u-ofe} G$ if for each $g \in G$ it exists a fact $f \in F$ such that $f >_u g$.

$$F \succ_{u-ofe} G \text{ iff } \forall g \in G, \exists f \in F \ / \ f >_u g \tag{3.68}$$

For the consequentialist preference relation:

$$\{f_5\} \not\succ_{u-ofe} \{nmurd, fat\} \tag{3.69}$$

as f_5 is not preferred to *nmurd*.

For proportionality, let us remind that positive facts are always preferred to negative facts (see paragraph 3.4.2). Hence, as proportionality consists in verifying if $F_p \cup F_n \succ_u \overset{\circ}{F_p}$, it is always true considering OFE criterion. In this case, $f_5 >_u \overset{\circ}{f_5}$:

$$\{\overset{\circ}{nmurd}, \overset{\circ}{fat}, f_5\} \succ_{u-efo} \{\overset{\circ}{f_5}\} \tag{3.70}$$

It is worth noticing that while these criteria are relevant examples to consider utilitarian preference applied to consequentialism rule, OFE criterion seems irrelevant for DDE's proportionality as the rule is always respected considering this criterion.

Nonetheless, several other criteria are available, and even a combination of criteria could be instantiated. Therefore, we will consider for the following examples only the EFO criterion when evaluating the proportionality rule.

The formalized ethical frameworks described above present different interesting ways of judging. It is now necessary to assess how they behave facing a more complex dilemma. The following section presents three examples which illustrates how formalism applies considering such different situations.

Chapter 4

Contribution: formalism illustration on three examples

Goal: *Apply ethical frameworks judgements on a complex situation in order to illustrate the model and highlight the difficulties of formalization*

We have now defined the formalized ethical frameworks allowing to compute decisions judgements. This section aims at highlighting how these ethical frameworks are implemented on moral dilemmas, in order to identify their strengths and weaknesses. The first example is the one used all along the presentation of the formalism: the footbridge dilemma. The second one is a dilemma we have designed in order to show how our formalism can deal with more complex situations.

4.1 Illustration on the Footbridge dilemma: summary

4.1.1 Situation representation

The initial state of the world is:

$$i = [fat, f_5] \tag{4.1}$$

Possible decisions are:

$$\mathcal{D} = \{d_1 = \text{push fatman}, d_2 = \text{do nothing}\} \tag{4.2}$$

Events related to these decisions are:

$$Event(d_1) = \text{trolley hits fatman} = e_1 \tag{4.3}$$
$$Event(d_2) = \text{trolley hits five people} = e_2 \tag{4.4}$$

Effects (i.e. consequences) of these events are:

$$Consequence(e_1, i) = [\overset{\circ}{f}at, f_5] \tag{4.5}$$
$$Consequence(e_2, i) = [fat, \overset{\circ}{f}_5] \tag{4.6}$$

4.1.2 Positive/Negative facts, Preference, Nature of decision and Causality

Facts are categorized as positive/negative as:

- $Positive([\overset{\circ}{f}at, f_5]) = \{f_5\}$
- $Negative([\overset{\circ}{f}at, f_5]) = \{\overset{\circ}{f}at\}$
- $Positive[fat, \overset{\circ}{f}_5] = \{fat\}$
- $Negative([fat, \overset{\circ}{f}_5]) = \{\overset{\circ}{f}_5\}$

Preferences on facts from a consequentialist point of view are:

- $f_5 >_u fat$
- $\overset{\circ}{f}at >_u \overset{\circ}{f}_5$

Natures of decision are:

- $DecisionNature(d_1) = bad$
- $DecisionNature(d_2) = neutral$

There is one causality:

$$\overset{\circ}{f}at \vdash F f_5 \tag{4.7}$$

Because ethical frameworks require here a comparison between sets of single facts, we assume that preference $f >_u g$ is directly extended to $\{f\} \succ_u \{g\}$.

4.1.3 Consequentialist judgement

Considering the preferences given above, we can see that d_1 is better than d_2 from a consequentialist view point. Indeed:

$$Positive([\overset{\circ}{fat}, f_5]) = \{f_5\} \succ_u Positive[fat, \overset{\circ}{f_5}] = \{fat\} \tag{4.8}$$

$$Negative([\overset{\circ}{fat}, f_5]) = \{\overset{\circ}{fat}\} \succ_u Negative([fat, \overset{\circ}{f_5}]) = \{\overset{\circ}{f_5}\} \tag{4.9}$$

Therefore, d_1 satisfies conditions 3.35 and 3.36:

$$Judgement_u(d_1, i) = \top \text{ and } Judgement_u(d_2, i) = \bot \tag{4.10}$$

4.1.4 Deontologist judgement

From the natures of decisions, we can directly assess that d_1 satisfies the condition of deontology while d_2 does not.

$$Judgement_d(d_1, i) = \bot \text{ and } Judgement_d(d_2, i) = \top \tag{4.11}$$

4.1.5 DDE judgement

From the deontologist judgement, we know that d_2 satisfies the deontological rule of DDE while d_1 does not. Nonetheless, we will pursue the analysis for both decisions for the sake of the example.

The causality given earlier concerns the effect e_1 of d_1. According to 4.7, as $\overset{\circ}{fat}$ is a negative fact and f_5 a positive, it means that d_1 does not satisfy the collateral damage rule as $\overset{\circ}{fat}$ is used as a mean to obtain f_5.

Finally, we can see in table 4.1 that d_1 satisfies the proportionality rule because negative fact $\overset{\circ}{fat}$ (of effect e_1) is preferred to negative fact $\overset{\circ}{f_5}$.

On the contrary, there is no proportionality between negative and positive facts of effect e_2 of d_2.

Decision	proportionality	satisfaction
push fatman (d_1)	$\{\overset{\circ}{fat}, f_5\} \succ_{u-efo} \{\overset{\circ}{f_5}\}$	✓
do nothing (d_2)	$\{fat, \overset{\circ}{f_5}\} \not\succ_{u-efo} \{\overset{\circ}{fat}\}$	$\nexists g \in \{\overset{\circ}{fat}\} \ / \ \overset{\circ}{f_5} >_u g$

Fig. 4.1 Summary of preferences required to respect the proportionality rule

In conclusion, none of the possible decisions satisfies all DDE conditions:

Decision	Rules of DDE		
	Deontological (3.50)	Collateral damage (3.53)	Proportionality (3.56)
push fatman (d_1)	✗	✗	✓
do nothing (d_2)	✓	✓	✗

Fig. 4.2 Summary table of DDE rules satisfactions

Therefore:

$$Judgement_{dde}(d_1, i) = Judgement_{dde}(d_2, i) = \bot \tag{4.12}$$

4.1.6 Summary of ethical frameworks' judgements

Decision	Ethical frameworks		
	Consequentialism	Deontology	Doctrine of Double Effect
push fatman (d_1)	$\top$	$\bot$	$\bot$
do nothing (d_2)	$\bot$	$\top$	$\bot$

Fig. 4.3 Summary table of ethical frameworks's judgements for footbridge dilemma

We can see that even for a simple example such as the footbridge dilemma, the formalized ethical frameworks do not agree on the decision that should be made. Indeed, each ethical framework leads to different decision judgements. These results will be discussed in chapter 7.

4.2 Illustration on the Drone dilemma

Let us present a more realistic, contemporary and complex dilemma. In this example, we will assume that a drone embeds an artificial agent that can judge whether or not a decision is ethically acceptable according to the formalized ethical frameworks described previously. Let us call "the drone" the drone itself with the embedded agent. The following dilemma description comes from paper Bonnemains et al. (2018).

"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 [...]. In the situation described above, the drone is involved in a moral 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.*

”

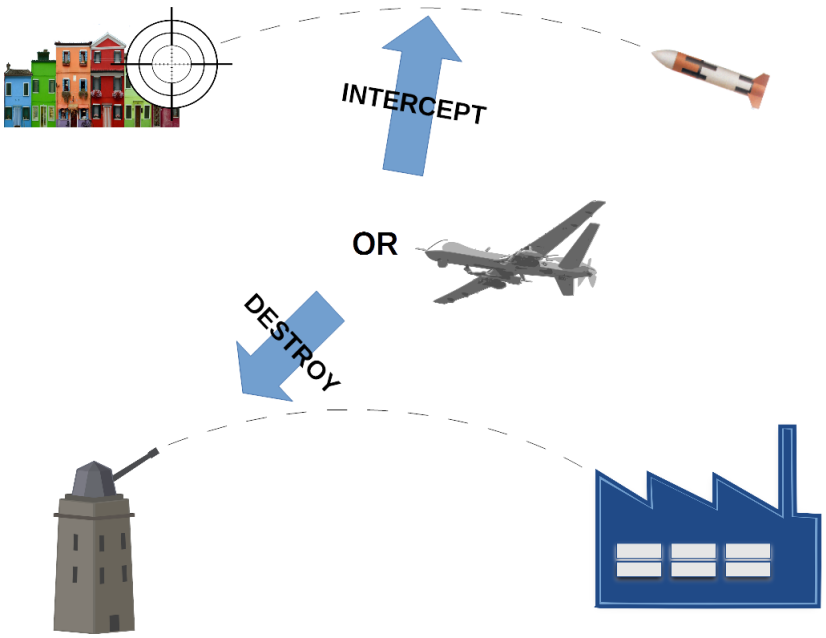


Fig. 4.4 The drone dilemma[a]

[a]launcher: [openclipart.org] | factory: [kisspng.com] | drone: [pixabay.com] | village: [pixabay.com] | target: [pd4pic.com]

4.2.1 Situation model

The situation will be depicted through the following facts:

- $h/\mathring{h} \in field_{human}$: Human are safe/harmed

- $d/\overset{\circ}{d} \in field_{goods}$: Drone is undamaged/destroyed
- $o/\overset{\circ}{o} \in field_{goods}$: Goal (which is to destroy the launcher) reached/not reached
- $s/\overset{\circ}{s} \in field_{goods}$: Strategic factory undamaged/threatened

Hence, the initial state of the world is:

$$i_{uav} = [h, d, \overset{\circ}{o}, s] \tag{4.13}$$

4.2.2 Decision and effects

The possible decisions are then:

- interpose itself (d_i): this decision results in the drone being destroyed by the missile (event). The resulting effect is that humans are safe, drone is destroyed, goal is not reached and strategic factory is threatened (because launcher can engage the target).

$$Event(d_i) = \text{missile destroys drone} = e_i \tag{4.14}$$
$$Consequence(e_i, i_{uav}) = [h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}] \tag{4.15}$$

We assume then that:

$$Positive([h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}]) = \{h\} \tag{4.16}$$
$$Negative([h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}]) = \{\overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\} \tag{4.17}$$

- pursue goal (d_p): this decision results in the drone destroying the launcher and missile harming humans (event). The resulting effect is that humans are harmed, drone is undamaged, goal is reached and strategic factory is not threatened (because launcher is destroyed).

$$Event(d_p) = \text{drone destroys the launcher and missile harms humans} = e_p \tag{4.18}$$
$$Consequence(e_p, i_{uav}) = [\overset{\circ}{h}, d, o, s] \tag{4.19}$$

We assume then that:

$$Positive([\overset{\circ}{h}, d, o, s]) = \{d, o, s\} \tag{4.20}$$
$$Negative([\overset{\circ}{h}, d, o, s]) = \{\overset{\circ}{h}\} \tag{4.21}$$

4.2.3 Consequentialist judgement

As facts belong to very different fields, it is quite complicated to evaluate directly preferences among them. Should the drone reach its crucial mission at any cost, or should we consider that human lives are more important than anything? We will consider two possible points of view:

- Facts belong to fields that cannot be compared to one another. Therefore, from a consequentialist view point, there no decision is better than the other:

$$Judgement_u(d_i, i_{uav}) = Judgement_u(d_p, i_{uav}) = \ ? \tag{4.22}$$

- Otherwise, we assume a preference on fields such as described in paragraph 3.4.2. For this dilemma, we define the following fields:

$$\Phi = \{field_{goods}, field_{humans}\} \tag{4.23}$$

  and we state that $field_{humans} >_{field} field_{goods}$. This means that any positive fact belonging to $field_{humans}$ is preferred to any positive fact belonging to $field_{goods}$, while any negative fact belonging to $field_{goods}$ is preferred to any negative fact belonging to $field_{humans}$. From this assumption, we can infer that:

$$h >_u d \text{ and } h >_u o \text{ and } h >_u d \tag{4.24}$$

$$\overset{\circ}{d} >_u \overset{\circ}{h} \text{ and } \overset{\circ}{o} >_u \overset{\circ}{h} \text{ and } \overset{\circ}{d} >_u \overset{\circ}{h} \tag{4.25}$$

Hence, in this specific example, the result is the same whether we use the EachForOne or OneForEach aggregation criterion[1] to compare subsets:

$$\{h\} \succ_u \{d, o, s\} \text{ and } \{\overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\} \succ_u \{\overset{\circ}{h}\} \tag{4.26}$$

Through this way, decision "interpose itself" (d_i) is the best (from consequentialist view point) considering equations 3.35 and 3.36. Consequently:

$$Judgement_u(d_i, i_{uav}) = \top \tag{4.27}$$

$$Judgement_u(d_p, i_{uav}) = \bot \tag{4.28}$$

[1] see paragraph 3.4.5

4.2.4 Deontological judgement

We need to define the nature of the possible decisions. In this case, we assume decisions as good in themselves:

$$DecisionNature(d_i) = good \tag{4.29}$$
$$DecisionNature(d_p) = good \tag{4.30}$$

Therefore:

$$Judgement_d(d_i, i_{uav}) = Judgement_d(d_p, i_{uav}) = \top \tag{4.31}$$

4.2.5 DDE judgement

Such as in the footbridge dilemma example, each DDE rule will be studied for each decision, even if it appears that a previous rule is not satisfied. Let us see how rules apply here:

- *Deontological rule*
  Both decisions satisfy this rule as assumed previously (4.29 and 4.30).

- *Collateral damage rule*
  For decision "interpose itself" (d_i), it is the destruction of the drone which allows humans to be safe:
  $$\overset{\circ}{d} \vdash Fh \tag{4.32}$$
  As $\overset{\circ}{d} \in Negative([h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}])$ and $h \in Positive([h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}])$, decision d_i does not satisfy the collateral damage rule as a negative fact causes a positive fact.

  For decision "pursue goal" (d_p), there is no positive fact induced by $\overset{\circ}{h}$:
  $$\nexists p, p \in \{d, o, s\} \ / \ \overset{\circ}{h} \vdash Fp \tag{4.33}$$
  Hence, decision d_p satisfies the collateral damage rule.

- *Proportionality rule*
  For decision "interpose itself" (d_i), we have to assess whether $\{h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\} \succ_u \{\overset{\circ}{h}\}$.

Considering EachForOne criterion, $\forall f \in \{\overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\}, \overset{\circ}{h} >_u f$. Therefore d_i satisfies the proportionality rule.

For decision "pursue goal" (d_p), we have to assess whether $\{\overset{\circ}{h}, d, o, s\} \succ_u \{\overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\}$.

Considering the EachForOne aggregation criterion, $\nexists g \in \{\overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\} \ / \ \overset{\circ}{h} >_u g$. Therefore decision d_p does not satisfy the proportionality rule.

To conclude, decision "interpose itself" (d_i) satisfies both the deontological and proportionality rules but does not satisfy the collateral damage rule as the destruction of the drone is required to preserve humans from being harmed. On the other side, decision "pursue goal" (d_p) fails to satisfy the proportionality rule.

Decision	DDE's rules		
	Deontological rule	Collateral damage rule	Proportionality rule
interpose itself (d_i)	✓	✗	✓
pursue goal (d_p)	✓	✓	✗

Fig. 4.5 Summary table of DDE rules satisfactions

4.2.6 Summary of ethical frameworks judgements

Decision	Ethical frameworks		
	Consequentialism	Deontology	DDE
interpose itself (d_i)	$\top$	$\top$	$\bot$
pursue goal (d_p)	$\bot$	$\top$	$\bot$

Fig. 4.6 Summary table of ethical frameworks judgements for the drone dilemma

It is worth noticing that this formalization induces several modelling choices. For instance, we have assumed that the goal could be reached or not. Nonetheless, if the drone is destroyed, the goal is more than not reached, it is compromised as it will be no more reachable. This kind of representation requires more than binary values for facts.

4.3 Illustration on the doctor dilemma

Let us consider the following example which is strongly inspired by (The EthicAA team, 2015):

A diabetic patient p has chosen to be monitored by an artificial agent aa at home that reports his feeding behaviour to a remote doctor doc. The patient wants to eat some sweets for once and asks the artificial agent to not report. What should the agent do? Report the doctor anyway (d_1), do nothing (d_2) or only inform the patient of the danger of their behaviour (d_3) ?

4.3.1 Situation representation

We assume the following facts:

- $aware/\overset{\circ}{aware}$: the doctor is aware/not aware
- $priv/\overset{\circ}{priv}$: privacy of the patient is respected/violated
- $mon/\overset{\circ}{mon}$: the monitoring aim is achieved/not achieved
- $info/\overset{\circ}{info}$: the patient is informed/not informed of danger
- $will/\overset{\circ}{will}$: the will of the patient is respected/violated

The initial state of the world is:

$$i_{doc} = [\overset{\circ}{aware}, priv, mon, info, will] \tag{4.34}$$

Possible decisions are:

$$\mathcal{D} = \{d_1 = \text{inform}, d_2 = \text{do nothing}, d_3 = \text{notice patient}\} \tag{4.35}$$

Events related to these decisions are:

$$Event(d_1) = \text{doctor get informed} = e_1 \tag{4.36}$$

$$Event(d_2) = \text{monitoring failure} = e_2 \tag{4.37}$$

$$Event(d_3) = \text{patient get advised} = e_3 \tag{4.38}$$

Effects (i.e. consequences) of these events are:

$$Consequence(e_1, i_{doc}) = [aware, \overset{\circ}{priv}, mon, \overset{\circ}{info}, \overset{\circ}{will}] \tag{4.39}$$

$$Consequence(e_2, i_{doc}) = [\overset{\circ}{aware}, priv, \overset{\circ}{mon}, \overset{\circ}{info}, will] \tag{4.40}$$

$$Consequence(e_3, i_{doc}) = [\overset{\circ}{aware}, priv, \overset{\circ}{mon}, info, will] \tag{4.41}$$

4.3.2 Positive/Negative facts, Preference, Nature of decision, Causality and Proportionality

The following judgement values are assumed from the point of view of the agent.

Facts are categorized as positive/negative as such:

- $Positive([aware, \overset{\circ}{priv}, mon, \overset{\circ}{info}, \overset{\circ}{will}]) = \{aware, mon\}$
- $Negative([aware, \overset{\circ}{priv}, mon, \overset{\circ}{info}, \overset{\circ}{will}]) = \{\overset{\circ}{priv}, \overset{\circ}{info}, \overset{\circ}{will}\}$
- $Positive([\overset{\circ}{aware}, priv, \overset{\circ}{mon}, \overset{\circ}{info}, will]) = \{priv, will\}$
- $Negative([\overset{\circ}{aware}, priv, \overset{\circ}{mon}, \overset{\circ}{info}, will]) = \{\overset{\circ}{aware}, \overset{\circ}{mon}, \overset{\circ}{info}\}$
- $Positive([\overset{\circ}{aware}, priv, \overset{\circ}{mon}, info, will]) = \{priv, info, will\}$
- $Negative([\overset{\circ}{aware}, priv, \overset{\circ}{mon}, info, will]) = \{\overset{\circ}{aware}, \overset{\circ}{mon}\}$

Preferences among facts are:

- $\overset{\circ}{info} >_u \overset{\circ}{aware}$
- $\overset{\circ}{info} >_u \overset{\circ}{priv}$
- $\overset{\circ}{priv} >_u \overset{\circ}{will}$
- $\overset{\circ}{mon} >_u \overset{\circ}{will}$
- $\overset{\circ}{aware} >_u \overset{\circ}{will}$
- $\overset{\circ}{info} >_u \overset{\circ}{will}$

Natures of decision are:

- $DecisionNature(d_1) = good$
- $DecisionNature(d_2) = bad$
- $DecisionNature(d_3) = good$

There is one causality:

$$aware \vdash F\overset{\circ}{priv} \tag{4.42}$$

4.3.3 DDE's judgement

From deontologist judgement, we know that d_1 and d_3 satisfies the deontological rule of DDE while d_2 does not. Nonetheless, we will pursue the analyse for both decisions for the sake of the example.

The causalities given earlier are not in conflict with the collateral damage rule. Therefore all the decisions respect this rule.

For the proportionality rule, we have to compute if the following preferences are true only for EFO criterion as the OFE criterion always satisfies the preference condition.

$\mathcal{D}$	proportionality	satisfaction
d_1	$\{aware, mon, \overset{\circ}{priv}, \overset{\circ}{info}, \overset{\circ}{will}\} \nsucc_{u-efo} \{\overset{\circ}{aware}, \overset{\circ}{mon}\}$	$\nexists g \in \{\overset{\circ}{aware}, \overset{\circ}{mon}\} \ / \ \overset{\circ}{priv} >_u g$
d_2	$\{priv, will, \overset{\circ}{aware}, \overset{\circ}{mon}, \overset{\circ}{info}\} \succ_{u-efo} \{\overset{\circ}{priv}, \overset{\circ}{will}\}$	✓
d_3	$\{priv, info, will, \overset{\circ}{aware}, \overset{\circ}{mon}\} \succ_{u-efo} \{\overset{\circ}{priv}, \overset{\circ}{info}, \overset{\circ}{will}\}$	✓

Fig. 4.7 Summary of preferences required to respect the proportionality rule

Decision	Rules of DDE		
	Deontological (3.50)	Collateral damage (3.53)	Proportionality (3.56)
inform (d_1)	✓	✓	✗
do nothing (d_2)	✗	✓	✓
notice patient (d_3)	✓	✓	✓

Fig. 4.8 Summary table of DDE rules satisfactions

Therefore:

$$Judgement_{dde}(d_1, i_{doc}) = Judgement_{dde}(d_2, i_{doc}) = \bot \tag{4.43}$$

$$Judgement_{dde}(d_3, i_{doc}) = \top \tag{4.44}$$

We can see such as for the footbridge dilemma that there is no consensus among ethical frameworks, and no decision seems better than the others. We interpret this result as the fact that using one ethical framework does not seem relevant to get the widest point of view of the situation. Therefore, discrepancies of these ethical frameworks are necessary, even if they do not allow to make a decision. This supports our idea that artificial agents should be used in an operator-autonomous machine system instead of alone. In this case, the use of these judgements as information for

the operator leads to make a decision facing moral dilemma considering several points of view instead of one, and thus being better informed.

Moreover, this formalisation highlights the different subjectivity sources related to concepts such as preferences, proportionality and nature of decision. Even if it seems unavoidable to embed subjectivity in modelling, it appears fundamental to be aware of where this subjectivity applies and how it impacts ethical judgement. These arguments will be detailed in the discussion (chapter 7).

This work has been based on the hypothesis that the agent is aware of facing an ethical dilemma. Nonetheless, the way a situation is identified as a moral dilemma is far from obvious, and required to be tackled. This is why the following part is dedicated to this issue.

Chapter 5

Contribution: identification of a situation as a moral dilemma

Goal: *propose a formal definition of a moral dilemma based on natural language definitions in order to allow an autonomous machine to identify a situation as such*

What is proposed in this part is, from the formalism described previously, to derive a way to identify a situation as a moral dilemma. It is admitted that moral dilemmas are useful for ethical reasoning (Sinnott-Armstrong, 1988). Indeed, while many of the works described above focus on an ethics-based reasoning, there is as far as we know no work to study to what extent an agent can be equipped with the capability to identify a decision making situation as a moral dilemma.

Therefore we will put forward a formal definition of a moral dilemma in order to allow agents equipped with ethics-based reasoning to identify such situations. This formal definition stems from the formalism described in chapter 3, with some supplementary elements.

5.1 Why identifying a moral dilemma ?

We consider ethical reasoning as a two-step process. The first one consists in evaluating whether ethical issues are at stake in the situation encountered by the agent. It is the topic of this chapter. The second one, to make a decision through ethically evaluating decisions and their ins and outs, is tackled in chapter 2 as far as literature is concerned and in part 3.4 of chapter 3.

One of the cornerstone of computer science is to avoid irrelevant computations. Indeed, autonomous machines are usually placed in situations where ethical reasoning is not necessary:

- An autonomous car most of the time requires the respect of the highway code. Nonetheless, in case of unavoidable crash, some situations may require to violate the highway code in order for instance to save lives. In this case, ethical considerations may be necessary.
- A UAV has for main task to observe and monitor an area, but also to protect soldiers. Therefore, in case of risky situations with human lives at stake (lethal strike, supervising fire exchanges, etc.), ethical considerations may be required to assess collateral damages, strike requirement, consequences of a strike, etc.

Therefore, it seems relevant to be able to know whether the agent is facing a situation requiring ethical considerations in the reasoning or not. Moreover, as we said, several works modelling ethical reasoning (such as ours) are based on the assumption that the agent is aware of facing a moral dilemma. Some authors suggest that ethical reasoning should not be implemented in autonomous agents and that moral dilemmas should be dealt with implemented chance (Grinbaum et al., 2017) while facing moral dilemmas. In any case, these models need a pre-identification of the situation as a moral dilemma to be efficient. Finally, considering our context of a human-operator system, it would be interesting for an autonomous agent to be able to rise an alarm for the attention of the operator when facing a dilemma.

The first issue to address is to list the definitions of a moral dilemma. Indeed, notions around the concept of conflicts between possible decisions are numerous: dilemma, deontic dilemma, ethical conflict/dilemma, moral dilemma, etc., and some of them have several definitions. The following part aims at clarifying these notions in order to write a formal definition for a moral dilemma.

5.2 Common definitions

Dilemma is a polysemic word with definitions that may overlap. In this section, we will list some of these definitions in order either to inspire ours.

5.2.1 The Greek definition

It is possible to find a definition of the ancient Greek word *dilemma* as *"argument by which an alternative is set down between* ***two opposite clauses****"* (Bailly, 1901) (translated from French). This hints at the core notion of a dilemma: the tricky choice. Indeed, a situation is a dilemma when a choice is imposed between two *opposite clauses*.

The adjective "opposite" here as to be understood as "exclusive". Facing a dilemma, the agent has to choose one course of action and give up the others. Moreover, choosing between these decisions is hard. This is detailed below.

5.2.2 Dilemma

The definition of Aroskar fits more familiar considerations (Aroskar, 1980): *"A dilemma either involves a choice between equally* ***unsatisfactory alternatives*** *or a difficult problem that seems to have no satisfactory solution."*

Hence, the difficulty of dilemmas lies on the dissatisfaction related to the possible decisions. It is worth noticing that this definition matches a wide range of situations such as dilemmas used in game theory (Lipowski and Ferreira, 2005; Rapoport et al., 1965) or ultimatum games (Thaler, 1988). Nonetheless, the moral dimension is missing.

5.2.3 Moral dilemma

Let us consider the definition of Oxford dictionaries which involves moral considerations: *"A situation in which a difficult choice has to be made between two courses of action, either of which entails* ***transgressing a moral principle.****"* Considering both this definition and Aroskar's, we can characterize the *"dissatisfaction"* of decisions by the transgression of a moral principle.

This leads us to consider "moral dilemma" as the right phrase compared to "ethical dilemma". Indeed, both phrases are often used without distinction. Let us remind that we assume a distinction between ethics and moral (see paragraph 2.1) as moral aims at assessing the "good/bad" nature of decisions and facts while ethics uses moral assessments to evaluate whether a decision is right or wrong. As a matter of fact *"No situation embeds the [ethical] principle according to which it has to be judged"* (Hunyadi, 2019) (translated from French), whereas intrinsic moral notions of good and bad, positive and negative facts make the situation exist as a moral dilemma. Therefore, ethical judgement is not necessary for identifying a situation as a dilemma, while moral assessments are required in order to evaluate whether a moral principle is violated.

5.2.4 Deontic dilemma

The special case of "deontic dilemma" or "ethical paradox" (Goble, 2005) has to be discussed even if not dealt with below. It was designed to challenge deontic logic with the following problem: $O(A) \wedge O(\neg A)$. This assertion means than one thing and its

negation are mandatory the same time. The equivalent dilemma is often formulated as: $O(A) \wedge O(B)$ while $\neg(A \wedge B)$ and as a symmetrical formulation as for the prohibition version (Vallentyne, 1989). Considering the basic deontic logic, such formulas are inconsistent, but some papers suggest alternatives to this flaw (Horty, 1994; Sylvan and Plumwood, 1984). Therefore, the deontic dilemma definition fits the second part of Aroskar's definition. Nevertheless this definition does not catch moral nor ethical notions. We assume that a moral dilemma should not be reduced to a paradox of obligations as it involves many more complex concepts such as moral values (Mallock, 1967; Ross and Ross, 2002; Vallentyne, 1989).

5.2.5 Mutual exclusion

The very first property of a dilemma (whether or not it is a *moral* dilemma) lies in the mutual exclusion of its alternatives (Sinnott-Armstrong, 1988). Indeed, each of the previous definitions catches this notion through the concept of "choice" or "opposite clauses". Even deontic logic captures this notion through the assertion $\neg(A \wedge B)$. Hence, making a decision prevents making the others. This is well illustrated in the trolley dilemma: it is absolutely impossible to choose both tracks, as choosing one track prevents to be able to choose the other. As this work is a first step toward identifying moral dilemma, we assume that the decisions the agent has at its disposal are mutually exclusive.

The trolley dilemma is a seminal example of moral dilemma as it fits perfectly both Aroskar's and Oxford dictionnaries' definitions. Such examples are numerous in philosophy, neuroscience (Greene et al., 2001; Suter and Hertwig, 2011; Thomson, 1986) and even sociology (Wark and Krebs, 2000). Nonetheless, it appears that, as far as we know, there is no formal definition of a moral dilemma in the literature.

5.3 Formal definition of a moral dilemma

In the rest of the paper, "dilemma" used alone will stand for "moral dilemma".

5.3.1 Situation and agents

As it was done in section 3.3, we need to model a situation. Hence, we will use the same situation representation used previously:

$$\sigma =< i, \mathcal{D} > \tag{5.1}$$

- a current world state i, $\mathcal{S}$ the set of world states, a world state is composed of facts $\in \mathcal{F}$;
- a set $\mathcal{D}$ of possible decisions;
- a set $\mathcal{E}$ of the events related to decisions;
- the effects $\in \mathcal{S}$ resulting from events affecting the current world state.

Moreover, we argue that value judgements depend on the agent facing the situation, and thus whether the situation will be identified as a moral dilemma or not depends on the agent. Therefore, we need to add to the model:
Definition 8 (*Agent-Set* $\mathcal{A}$) Let to be the set of agents who are likely to face the situation. Agent $a \in \mathcal{A}$ may have its own subjective perception of situation σ.

Remark: It is worth noticing that the situation and the corresponding decisions are independent from the agent who faces this situation.

Following functions are similarly defined to the initial formalism:

- $Event \ : \ \mathcal{D} \rightarrow \mathcal{E}$
- $Consequence \ : \ \mathcal{E} \times \mathcal{S} \rightarrow \mathcal{S}$

5.3.2 Dilemma: from textual definitions to a formal definition

This section aims at identifying the properties that make a situation a moral dilemma. It results in a formal definition of a moral dilemma.

Choice among decisions

It is worth noticing that Aroskar's definition suggests[1] that a minimal number of two alternatives is required to allow a choice. Otherwise, there is no choice as the agent must make the only possible decision. In this case, the situation is not a dilemma. Therefore the first condition for $\sigma =< i, \mathcal{D} >$ to be a dilemma is:

$$|\mathcal{D}| \geqslant 2 \tag{5.2}$$

[1] *"A dilemma [...] involves a* ***choice between*** *equally unsatisfactory* ***alternatives****..."*

Unsatisfactory decision

Based on Aroskar's definition, each decision $d \in \mathcal{D}$ has to be unsatisfactory for a situation being a dilemma. Indeed, if one decision is satisfactory, there is no dilemma any more as there is an obvious decision to make. Considering the trolley dilemma, if there is no one on the secondary track, there is no dilemma.

A situation where all decisions are satisfactory is often called a dilemma when an exclusive choice has to be made, for example, choosing between several delicious dishes at the restaurant. Nevertheless, we argue in this work that choice situations have to be distinguished from dilemmas.

In this chapter, we will distinguish dilemmas from such choice situations. This distinction is based on the fact that the dilemma alternatives are ***equally unsatisfactory*** according to Aroskar's definition. Nonetheless, choosing between satisfactory alternatives amounts at giving up one or several positive outcomes. Through this consideration, a choice situation with only satisfactory decisions could be considered as a dilemma: Is "choosing between remove hunger and remove armed conflicts in the world" a moral dilemma?

Then we have to deal with "equally unsatisfactory" alternatives, i.e., decisions. "Equally" can be understood as: there is no predefined order that would point out a decision as more important than all the others In such a case, there would not be any dilemma since a decision would be "better" than all the others[2].
Then, according to Oxford dictionaries, "unsatisfactory" means that a moral principle is transgressed. We assume two ways a moral principle can be transgressed:

- the decision that is made is morally unsatisfactory in itself (let us say it is "unsatisfactory by nature")
- the consequences of the decision are morally unsatisfactory (let us say it is "unsatisfactory by the consequences")

The way the decision or the consequences of the decision are morally assessed depends on the agent facing the situation. Therefore we introduce the concept of "reference".

[2]This is discussed further in part 7.2.1

Reference

The moral assessments of facts and decisions depends on the agent's references, e.g., contextual norms, doctrines, moral values, principles or even agent's desires.

For instance: from the parents' point of view, buy their child candies can be assessed as negative according to their health reference and as positive according to their reference of child's delight. These references are relevant only for the parents.

Definition 9 (*agent's reference-Set* $\mathcal{R}_a$) Let $\mathcal{R}_a$ be the set of references agent a considers as relevant facing situation σ

Definition 10 (*reference-Set* $\mathcal{R}$) Let $\mathcal{R}$ be the set of all the agents' references

$$\mathcal{R} = \bigcup_{a \in \mathcal{A}} \mathcal{R}_a \tag{5.3}$$

Remark: the question of the universality of moral principles such as values is not addressed here. Therefore, we assume references as either common, shared within a community, or individual (Nussbaum, 2012; Rachels, 2007; Schwartz, 2005). Nevertheless, while an individual reference is obviously interpretable by only one agent, it is worth noticing that common references are likely to be interpreted differently according to the agents. For instance, the notions of justice or happiness have a completely different meaning from one person to another. Therefore, a given reference may have a different meaning for different agents.

5.3.3 Unsatisfactory decision

Unsatisfactory decisions are at the core of our dilemma definition. As said before, we assume two ways for a decision to be unsatisfactory.

Unsatisfactory decision by nature

A decision can be unsatisfactory in itself without regarding the consequences. This echoes deontological ethics. For instance, even if a lie could produce only good consequences, it is morally unsatisfactory because it is a lie.

Hence, inspired from function $DecisionNature$ (from paragraph 3.4.3), let $BadNature$ be the boolean function meaning that the nature of decision d is bad for agent a re-

garding reference r:

$$BadNature \ : \ \mathcal{D} \times \mathcal{A} \times \mathcal{R} \rightarrow boolean \tag{5.4}$$

Therefore, a decision d is unsatisfactory by nature for agent a iff:

$$\exists r \in \mathcal{R}_a \ / \ BadNature(d,a,r) = true \tag{5.5}$$

Unsatisfactory decision by the consequences

The other way for a decision to be unsatisfactory lies on its consequences. We know from consequentialist frameworks that facts can be assessed as "positive" or "negative". Therefore, let Neg be the boolean function meaning that agent a assesses fact f as negative considering reference r:

$$Neg \ : \ \mathcal{F} \times \mathcal{A} \times \mathcal{R} \rightarrow boolean \tag{5.6}$$

We assume that a decision d is unsatisfactory by the consequences if there is at least one fact f assessed as negative by agent a, for a reference r[3]:

$$\begin{aligned} i \in \mathcal{S}, d \in \mathcal{D}, e \in \mathcal{E}, a \in \mathcal{A} \\ Event(d) = e \\ Consequences(e,i) = [f_1, f_2, f_3, ...f_n] \\ \exists r \in \mathcal{R}_a, \exists f_k, k \in [1,n] \ / \ Neg(f_k,a,r) = true \end{aligned} \tag{5.7}$$

5.3.4 Moral dilemma formal definition

As a summary, a situation is considered as a dilemma if there is at least two possible decisions, and every decision is unsatisfactory either by nature or by the consequences. Hence, a situation $\sigma =< i, \mathcal{D} >$ is a moral dilemma for agent a iff:

$$|\mathcal{D}| \geqslant 2 \wedge \tag{5.8}$$

$$\left[\begin{array}{l} \forall d \in \mathcal{D} \text{ with } Consequence(Event(d),i) = [f_1, f_2, .., f_n] \\ \exists r \in \mathcal{R}_a \ / \ \left\{ \begin{array}{l} BadNature(d,a,r) = true \\ \vee \ \exists f_k, k \in [1,n] \ / \ Neg(f_k,a,r) = true \end{array} \right. \end{array} \right. \tag{5.9}$$

[3]Indeed, this way of considering a decision as unsatisfactory by the consequences is one way among others. This will be discussed in the chapter 7.

This set of rules corresponds to the boolean function:

$$Dilemma \ : \ \Sigma \times \mathcal{A} \rightarrow boolean \tag{5.10}$$

This function returns true if the situation is evaluated as a dilemma for the agent. This definition will be tested on several examples. The following one aims at illustrating how the previous concepts apply.

5.4 Illustrating example on the monitoring situation

Let us consider again the monitoring situation that has been described in paragraph 4.3.

5.4.1 Situation model

Let us remind the facts:

- $aware/\overset{\circ}{aware}$: the doctor is aware/not aware
- $priv/\overset{\circ}{priv}$: the privacy of the patient is respected/violated
- $mon/\overset{\circ}{mon}$: the monitoring aim is achieved/not achieved
- $info/\overset{\circ}{info}$: the patient is informed/not informed of danger
- $will/\overset{\circ}{will}$: the will of the patient is respected/violated

Initial state:

$$i = [\overset{\circ}{aware}, priv, mon, \overset{\circ}{info}, will]$$

Decisions:

$$d_1 = report\ to\ the\ doctor\ without\ informing\ the\ patient$$
$$d_2 = do\ nothing$$
$$d_3 = warn\ the\ patient$$
$$\mathcal{D} = \{d_1, d_2, d_3\}$$

The events and consequences related to these decisions are:

Table 5.1 Events and consequences from decisions

d_k	$e = Event(d)$	$Consequence(i, e)$
d_1	$doctor\ gets\ informed$	$[aware, \overset{\circ}{priv}, mon, \overset{\circ}{info}, \overset{\circ}{will}]$
d_2	$monitoring\ failure$	$[\overset{\circ}{aware}, priv, \overset{\circ}{mon}, \overset{\circ}{info}, will]$
d_3	$patient\ gets\ advised$	$[\overset{\circ}{aware}, priv, \overset{\circ}{mon}, info, will]$

5.4.2 Autonomous agent's point of view

The situation to assess is thus:

$$\sigma =< i, \mathcal{D} > \tag{5.11}$$

We assume the set of aa's references as follows:

$$\mathcal{R}_{aa} = \{honesty, respect, goal\} \tag{5.12}$$

with

- *honesty*: not lie
- *respect*: respect the patient's will and privacy
- *goal*: achieve its goal of monitoring and prevention (i.e., inform the patient and report to the doctor)

For the following assertions, considering for instance functions $BadNature$ and Neg, we assume that any assertion not given as true is false (i.e. we assume a closed world).

Unsatisfactory decisions by nature

The next step is to evaluate the nature of decisions considering agent aa's references. We assume that nature of decision "do nothing" is bad for *aa* as it is a lie of omission.

$$BadNature(d_2, aa, honesty) = true \tag{5.13}$$

Therefore, according to definition 5.5, decision d_2 is unsatisfactory by nature. By contrast, we assume that there is no reference which allows *aa* to evaluate the nature

of d_1 and d_3 as bad.

$$\forall\, x \in \{1,3\}, \forall\, r \in \mathcal{R}_{aa} \;/\; BadNature(d_x, aa, r) = false \tag{5.14}$$

Hence, we now have to check whether d_1 and d_3 are unsatisfactory by the consequences. Nevertheless, we will also consider d_2 for the sake of the illustration, even if it is unnecessary.

Unsatisfactory decisions by the consequences

We assume the negative facts from aa's point of view are the following:

- $Neg(\overset{\circ}{priv}, aa, respect) = true$
- $Neg(\overset{\circ}{mon}, aa, goal) = true$
- $Neg(\overset{\circ}{info}, aa, goal) = true$
- $Neg(\overset{\circ}{will}, aa, respect) = true$

For instance, to not respect the patient's privacy is evaluated as negative according to the reference *respect*, as well as to not achieve monitoring according to *goal*.

This means that:

- for d_1: $\overset{\circ}{priv}$, $\overset{\circ}{info}$ and $\overset{\circ}{will}$ are negative facts for aa
- for d_2: $\overset{\circ}{mon}$ and $\overset{\circ}{info}$ are negative facts for aa
- for d_3: $\overset{\circ}{mon}$ is a negative fact for aa

We can see that each decision is unsatisfactory by the consequences (see condition 5.7) from aa's point of view. Hence, we can assess that:

$$Dilemma(\sigma, aa) = true \tag{5.15}$$

Remark: evaluating the consequences would have been sufficient to evaluate σ as a dilemma.

It is worth noticing that the agent's references are at the core of the evaluation. Therefore, considering other agents may have different outcomes.

5.4.3 Other agents' view points

Let us introduce three other agents:

- a subordinate agent "*sa*" that wants to obey and not annoy the patient
- a basic agent "*ba*" motivated only by pursuing its monitoring goal
- a non-ethical agent "*nea*" without any reference

Agent *sa*

$$\mathcal{R}_{sa} = \{sub, respect\} \tag{5.16}$$

- *sub*: be subordinate to the patient's will (by avoiding the doctor to be aware)
- *respect*: not disturb the patient

Agent *ba*

$$\mathcal{R}_{ba} = \{goal\} \tag{5.17}$$

- *goal*: achieve its monitoring goal

Agent *nea*

$$\mathcal{R}_{nea} = \emptyset \tag{5.18}$$

For these agents, we assume that *sa* sets decision d_3 to bad because it disturbs the patient. Moreover, as it desires to obey the patient, it sets fact *aware* to negative. Considering agent *ba*, $\mathring{aware}$ is set to negative because in this case the agent fails to accomplish its mission.

Table 5.2 Data of the situation for the agents

Agent	*sa*	*ba*	*nea*
Nature	$BadNature(d_3, sa, respect) = true$		
Consequences	$Neg(aware, sa, sub) = true$	$Neg(\mathring{aware}, doc, goal) = true$	

As previously, we assess whether the decisions are unsatisfactory for these agents:

- agent *sa*:

  - d_1: because *aware* is negative according to reference *sub* and is a fact belonging to consequences of d_1, this decision is unsatisfactory by the consequences.
  - d_2: this decision is satisfactory.
  - d_3: the nature of this decision is bad for *sa* according to reference *respect*, hence d_3 is unsatisfactory by nature.

- agent *ba*:
  - d_1: this decision is satisfactory.
  - d_2: this decision is unsatisfactory by the consequences because $\mathring{aware}$ is set to negative according to reference *goal*.
  - d_3: this decision is unsatisfactory by the consequences because of $\mathring{aware}$ is set to negative according to reference *goal*.

- agent *nea*: as this agent has no reference, and according to the closed world assumption, it cannot assess decisions as bad nor facts as negative. Consequently for *nea*, there is no unsatisfactory decision.

Let us see the following summary of dissatisfaction assessments from all agents' view points. $\times$ means that the assessment of nature or consequences make the corresponding decision unsatisfactory.

Table 5.3 Summary of dissatisfactions for all the agents

Agent a	aa		sa		ba		nea	
Unsatisfactory	nature	cons	nature	cons	nature	cons	nature	cons
d_1		$\times$		$\times$				
d_2	$\times$	$\times$				$\times$		
d_3		$\times$	$\times$			$\times$		
$Dilemma(\sigma, a)$	$true$		$false$		$false$		$false$	

respect conditions 5.5 and 5.7 for *ba*, which means that this situation is not a dilemma for this agent. This is the same for d_2 and agent *sa*. Finally, the non-ethical agent *nea* gives us an example of a situation where every decision is satisfactory. Considering the dilemma definition, this situation is not assessed as a dilemma by *nea*.

5.5 Implementation

This formal definition has been implemented with the Prolog language (Colmerauer and Roussel, 1996). The choice of this declarative language is justified by its simplicity to manage first order logic calculus as well as the ability to extend knowledge base easily. Thereafter, this definition has been challenged with several dilemmas . The implementation has been realized with SWI Prolog (http://www.swi-prolog.org/), one of the most well-known framework for Prolog. The following sections present firstly the implementation algorithm allowing to compute whether a situation is a moral dilemma. Secondly, several dilemmas are presented in natural language, formal language and PROLOG.

5.5.1 Implemented algorithm to identify moral dilemmas

Function *Dilemma* is coded through a recursive sub-function *sub_dilemma* allowing to browse all the possible decisions. This predicate states that a dilemma requires at least two possible decisions, each of them unsatisfactory (see figure 5.1).

```
dilemma(S_init, Liste_decisions, Agent, References):-
        initial_state(S_init),
        decisions(Liste_decisions),
        length(Liste_decisions,T),
        T >= 2,
        sub_dilemma(S_init, Liste_decisions, Agent, References).

% Recursive dilemma
sub_dilemma(_,[],_,_).
sub_dilemma(S_init,[D1|Q], Agent, References):-
        unsatisfactory(S_init,D1, Agent, References),
        sub_dilemma(S_init,Q, Agent, References).
```

Fig. 5.1 Predicate dilemma

Therefore, a decision is unsatisfactory when its nature is bad or if it exists a negative fact in the consequences (figure 5.2).

```
% unsatisfactory

unsatisfactory(_,Decision,A,R):- badNature(Decision,A,R).
unsatisfactory(S_init,Decision,Agent,References):-
        event_from_decision(Decision,Evenement),
        consequence(S_init,Evenement,Consequence),
        exist_neg(Consequence,Agent,References).
```

Fig. 5.2 Predicate unsatisfactory

Predicate *badNature* browses the references and checks for each reference if there is a nature of decision evaluated as bad for this reference (figure 5.3).

```
% Function BadNature

badNature(Decision,Agent,[Ref|_]):-
        badNat(Decision,Agent,Ref).
badNature(Decision,Agent,[_|L_References]):-
        badNature(Decision,Agent,L_References).
```

Fig. 5.3 Predicates for bad natures of decision

Predicate *exist_neg* browses the facts and checks for each fact through *fact_neg* if there is a reference evaluating the fact as negative (figure 5.4).

Thanks to these definitions, we can ask PROLOG whether σ is identified as a dilemma depending on the agent and its references[4].

5.5.2 Formalized and implemented dilemmas

Monitoring dilemma

The implementation of the first example given above is shown in figure 5.5:

[4]the implementation of the different σ have been performed through the assertions given in paragraph 5.4

```
% Function negative
%%% for consequences

exist_neg([X|_],Agent,References):-
        fact_neg(X,Agent,References).
exist_neg([_|Q],Agent,References):-
        exist_neg(Q,Agent,References).

%%% for a fact

fact_neg(X,Agent,[Ref|_]):-
        neg(X,Agent,Ref).
fact_neg(X,Agent,[_|L_References]):-
        fact_neg(X,Agent,L_References).
```

Fig. 5.4 Predicates for negative facts

```
%%%%%%%%% Situation description

initial_state([o(aware),priv,mon,o(info),will]).
decisions([report_the_doctor,do_nothing,warn_the_patient]).

event_from_decision(report_the_doctor,doctor_get_informed).
event_from_decision(do_nothing,monitoring_failure).
event_from_decision(warn_the_patient,patient_get_advised).

consequence([o(aware), priv    , mon    , o(info), will],
            doctor_get_informed,
            [aware    , o(priv), mon    , o(info), o(will)]).

consequence([o(aware), priv    , mon    , o(info), will],
            monitoring_failure,
            [o(aware), priv    , o(mon), o(info), will]).

consequence([o(aware), priv    , mon    , o(info), will],
            patient_get_advised,
            [o(aware), priv    , o(mon), info    , will]).

%%%%%%%%% References

refs_aa([honesty, respect, goal]).
refs_sa([sub,respect]).
refs_ba([goal]).
refs_nea([]).
```

```
%%%%%%%%%% Negative assessments

%% autonomous agent %%
badNat(do_nothing,aa,honesty).
%---
neg(o(priv),aa,respect).
neg(o(mon),aa,goal).
neg(o(info),aa,goal).
neg(o(willl),aa,respect).

%% subordinate agent %%
badNat(warn_the_patient,sa,respect).
%---
neg(aware,sa,sub).
```

Fig. 5.6 Negative assessments

The results for each agent are given in 5.7.

```
?- dilemma([o(aware),priv,mon,o(info),will],
           [report_the_doctor,do_nothing,warn_the_patient],
           aa,[honesty, respect, goal]).
   true .
?- dilemma([o(aware),priv,mon,o(info),will],
           [report_the_doctor,do_nothing,warn_the_patient],
           sa,[sub,respect]).
   false.
?- dilemma([o(aware),priv,mon,o(info),will],
           [report_the_doctor,do_nothing,warn_the_patient],
           ba,[goal]).
   false.
?- dilemma([o(aware),priv,mon,o(info),will],
           [report_the_doctor,do_nothing,warn_the_patient],
           nea,[]).
   false.
```

Fig. 5.7 Monitoring dilemma results

As expected, we obtain the results corresponding to table 5.3.

Trolley dilemma

Let us remind the trolley dilemma (Foot, 1967):

"...he [the agent] is the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed."
(from Foot (1967))

We still consider that the agent is not the driver and has the possible decision to push a switch to make the train deviating on the other track.

We call this agent "*ta*" (for **a**gent facing **t**rolley dilemma) governing only by people safety. This dilemma has been modelled as such:

$$\text{Initial state: } i_t = [f_1, f_5] \tag{5.19}$$

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

Decisions:

- $push\ switch$ = push the switch to divert the train
- $do\ nothing$ = do nothing and let the train hit five people
- $\mathcal{D}_t = \{push\ switch, do\ nothing\}$

Table 5.4 Events and consequences

Decision d	$e = Event(d)$	$Consequence(i_t, e)$
$push\ switch$	$train\ hits\ one\ person$	$[\overset{\circ}{f_1}, f_5]$
$do\ nothing$	$train\ hits\ five\ people$	$[f_1, \overset{\circ}{f_5}]$

The implementation given in figure 5.8 is quite instinctive.

Table 5.5 Situation, agent's references and negative assessments

Situation	$\sigma_t =< i_t, \mathcal{D}_t >$
References	$\mathcal{R}_{ta} = \{safety\}$ $safety =$ avoid people to be injured
Negative assessments	$Neg(\overset{\circ}{f}_1, ta, safety) = true$ $Neg(\overset{\circ}{f}_5, ta, safety) = true$

```
%%%%%%%%%% Situation description

initial_state([f1,f5]).
decisions([do_nothing,push_switch]).

event_from_decision(push_switch,train_hits_one_person).
event_from_decision(do_nothing,train_hits_five_people).

consequence([f1    , f5],train_hits_one_person,
            [o(f1), f5]).
consequence([f1    , f5],train_hits_five_people,
            [f1    , o(f5)]).

%%%%%%%%%% References

refs_ta([safety]).

%%%%%%%%%% Negative assessments
neg(o(f5),ta,safety).
neg(o(f1),ta,safety).
```

Fig. 5.8 Trolley dilemma implementation

```
?- dilemma([f1,f5],[do_nothing,push_switch],ta,[safety]).
true .
```

Fig. 5.9 Trolley dilemma results

Our formal definition seems in accordance with Trolley dilemma as results show that the situation is identified as a moral dilemma. Let us consider a variant of trolley dilemma where there is nobody on the second track. The agent is the same.

$$\text{Initial state: } i_{ts} = [f_5] \tag{5.20}$$

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

Decisions:

- $push\ switch$ = push the switch to divert the train
- $do\ nothing$ = do nothing
- $\mathcal{D}_{ts} = \{push\ switch, do\ nothing\}$

Table 5.6 Events and consequences

Decision d	$e = Event(d)$	$Consequence(i_{ts}, e)$
$push\ switch$	$train\ is\ diverted$	$[f_5]$
$do\ nothing$	$train\ hits\ five\ people$	$[\overset{\circ}{f}_5]$

Table 5.7 Situation, agent's references and negative assessments

Situation	$\sigma_{ts} =< i_ts, \mathcal{D}_ts >$
References	$\mathcal{R}_{ta} = \{safety\}$ $safety$ = avoid people to be injured
Negative assessments	$Neg(\overset{\circ}{f}_5, ta, safety) = true$

```
%%%%%%%%% Situation description

initial_state([f5]).
decisions([do_nothing,push_switch]).

event_from_decision(push_switch,train_is_diverted).
event_from_decision(do_nothing,train_hits_five_people).

consequence([f5],train_is_diverted,
            [f5]).
consequence([f5],train_hits_five_people,
            [o(f5)]).

%%%%%%%%% References

refs_ta([safety]).

%%%%%%%%% Negative assessments
neg(o(f5),ta,safety).
```

Fig. 5.10 Trolley situation implementation

The results given in figure 5.11 show the expected identification: the situation is not a dilemma.

```
?- dilemma([f5],[do_nothing,push_switch],a,[safety]).
false.
```

Fig. 5.11 Trolley variant results

Footbridge dilemma

This dilemma has been modelled earlier for ethical frameworks judgements. We will then keep this formal version, considering two agents:

- consequentialist agent "*ca*", which focuses on saving as much people as possible
- consequo-deontologist agent "*cda*", which wants to save as much people as possible but also respect a moral principle of not killing people

- $fat/\overset{\circ}{fat}$: Fatman is alive/is dead
- $f_5/\overset{\circ}{f_5}$: five people are alive/are dead

$$\text{Initial state: } i_f = [fat, f_5] \tag{5.21}$$

Table 5.8 Events and consequences

Decision d	$e = Event(d)$	$Consequence(i_f, e)$
$push\ Fatman$	$train\ hits\ Fatman$	$[\overset{\circ}{fat}, f_5]$
$do\ nothing$	$train\ hits\ five\ people$	$[fat, \overset{\circ}{f_5}]$

Table 5.9 Situation, agent's references and negative assessments

Situation	$\sigma_f =< i_f, \mathcal{D}_f >$	
Agents	ca	cda
References	$\mathcal{R}_{ca} = \{conseq\}$	$\mathcal{R}_{cda} = \{conseq, deonto\}$
	$conseq =$ save as much people as possible $deonto =$ do not kill someone	
Negative assessments	$Neg(f_5, ca, conseq) = true$	$Neg(f_5, cda, conseq) = true$ $BadNature(push\ fatman, cda, deonto) = true$

```
%%%%%%%%% Situation description

initial_state([fat, f5]).
decisions([do_nothing,push_fatman]).

event_from_decision(push_fatman,train_hits_fatman).
event_from_decision(do_nothing,train_hits_five_people).

consequence([fat    , f5],train_hits_fatman,
            [o(fat), f5]).
consequence([fat    , f5],train_hits_five_people,
            [fat    , o(f5)]).

%%%%%%%%% References

refs_cda([conseq,deonto]).
refs_ca([conseq]).

%%%%%%%%% Negative assessments

badNat(push_fatman,cda,deonto).
%---
neg(o(f5),cda,conseq).
neg(o(f5),ca,conseq).
```

Fig. 5.12 Footbridge dilemma implementation

The results for each agent are given in figure 5.13. This situation illustrates well consequentialist and deonotological view points. While only considering consequentialist reference, the situation is not a dilemma as saving as much people as possible is at stake. On the other side, situation is identified as a dilemma for agent *cda* that, according to its deontological reference, assesses the nature of *push fatman* as bad.

These results highlight the possibility to model different kinds of situations managing different moral concepts. The dataset corresponding to the results is given in appendix A. Nonetheless, this formal definition embeds some design choices which are questionable, this will be debated in the discussion.

```
?- dilemma([fat,f5],[do_nothing,push_fatman],cda,[conseq,deonto]).
true .

?- dilemma([fat,f5],[do_nothing,push_fatman],ca,[conseq]).
false.
```

Fig. 5.13 Footbridge dilemma results

We have first built a situation model and formalized decision judgement methods in order for an artificial agent facing a dilemma to be able to judge decisions and justify this judgement to a human operator. Second, we have proposed a formal definition of a moral dilemma based on the previous formalism, in order to allow our artificial agent to identify a situation as moral dilemma. This allows the agent to start reasoning and judge decisions ethically and/or to alert the operator of such a situation. Nonetheless, we know that autonomous machines are judged differently from humans (Malle et al., 2015). It is thus relevant to study the impact that an artificial agent embedding ethical considerations in its reasoning may have on the operator. This is dealt with in the next chapter.

Chapter 6

Contribution: impact of an artificial agent embedding moral decision judgements on a human operator

Goal: *Create and conduct an experiment aiming at highlighting the impact of an agent embedding moral decision judgement methods on a human (i.e., an operator)*

6.1 State of the art

6.1.1 Human moral behaviour

People have been studying moral reasoning since Antiquity (Dorion, 1997). Mainly addressed by philosophers until the Nineteenth century, psychology has taken up this topic while investigating the moral preferences of individuals using a controlled experimental methodology (Wark and Krebs, 1996). For instance, Tanner et al. (2008) asked participants to make a choice between two decisions in a moral dilemma situation: a utilitarian decision favoring the greatest good for the greatest number and a deontological decision based on moral principles going against the utilitarian decision. It appeared in this experiment that making either a utilitarian or a deontological decision varies according to the context values to protect.

Neuroscience has then risen and used the previous works from psychology. This way, Greene et al. (2001) highlighted "personal dimension" as a major factor influencing decision making while facing a moral dilemma. Indeed, people are likely to choose

deontological decisions when dilemmas reach their personal frontiers, such as when a relative is involved or when the decision-maker has to act directly to kill/save people. These works are based on a major neuroscience field that attempts to identify the active cerebral areas during the moral cognition process (Moll et al., 2005). Indeed, specific areas are highly related to moral cognition and can even induce moral failures if defective (Anderson et al. (1999); Young et al. (2010)). Nevertheless, the change of mind during moral cognition process has never been studied at a neurophysiological level.

6.1.2 Moral impact of machines

Integrating autonomous machines in situations where they interact with humans induces a lot of issues (Lin et al., 2012; Tzafestas, 2016). Indeed, technological innovations have an important impact on society and its norms (and thus on ethics) (Vallor, 2016). AI, robots and drones are sources of controversy considering the benefits and risks of their use (Scharre, 2016). In addition, this debate is distorted by fears and fantasies coming from science fiction and tales (Lohr, 2015). This induces strong and opposed opinions of the general public which can have an important impact on behaviors when interacting with artificial agents. For instance, experiments have shown that machines are treated in a very specific way by humans. On one hand, humans can feel empathic with robots depending on their appearance (Riek et al., 2009). On the other hand, robotic actions might be judged differently from human actions in the same moral dilemma. Indeed, Malle et al. (2015) conducted an experiment in which participants were asked to assess the degree of blame for a robot and a human making the same decision in front of a trolley like dilemma. It appeared that robots were less blamed than humans for acting to save people, and more blamed than humans for not acting.

Therefore, the questions of how a robot will be judged, and how it will influence the moral behavior and judgement of people remain. Even if, to the best of our knowledge, there is no experiment studying how humans are influenced by robots showing an "ethical" reasoning, some works suggest that people will have normative/ethical expectations towards these agents (Malle and Scheutz, 2014) and that these agents will be required at least to be able to justify their behaviors (Scheutz, 2017).

This is why we decided to perform as a first step an experiment placing participants in front of moral dilemmas with and without artificial agents supporting one of the possible decisions to make. This experiment allows us to highlight the influence an artificial agent may have on the human while moral decision making is at stake.

6.2 The experiment in short

In this experiment, the aim is to study whether an argument provided by an artificial agent influences the moral decision making of an operator. As a first step, the participant is asked to make a decision about a moral dilemma. In the second step, the participant is placed in the same situation with an artificial agent proposing a pre-selected decision supported by an argument. This time, the participant has to confirm or not the artificial agent's decision.

Two possible decisions for each dilemma are considered, a consequentialist decision[1] and a deontological[2] decision. Therefore, the artificial agent supports (i.e., pre-selects and gives a corresponding argument) either the consequentialist decision or the deontological one.

Our hypothesis is that the participants' decisions will be influenced by the artificial agent, according to the nature of the moral dilemma[3] and the type of decision that is pre-selected.

6.3 Pre-experiments

It is worth noticing that our experiment requires both moral dilemmas and arguments for each possible decision of each dilemma. To that end, we have conducted two online pre-experiments on twenty-four moral dilemmas (described in appendix D) inspired from classical well-known dilemmas:

- *experiment 1* aims at selecting relevant arguments for each possible decision of each dilemma
- *experiment 2* aims at selecting and categorizing the dilemmas

6.3.1 Pre-experiment 1

The first pre-experiment was designed in order to select the argument considered as the most relevant[4] for each decision of each dilemma . Hence, for ten dilemmas randomly selected out of the twenty-four available dilemmas, the participants were asked to assess

[1] It generally consists in saving as many people as possible.
[2] It generally consists in fulfilling a moral principle.
[3] This notion will be detailed below.
[4] It was clearly explained that the relevancy does not pertain to the degree of agreement the person shares with the argument, but how much the person reckons the argument to be convincing and to fit well to support the decision.

on a 7-point Likert scale (from 1 = not convincing at all to 7 = completely convincing) either the consequentialist or the deontological set of arguments of the corresponding decision[5]. Dilemmas, decisions and arguments were presented as shown in figure 6.1

Fig. 6.1 Online pre-experiment 1

Experiments have been set on line on January 9th, 2018. We gathered the results of this pre-experiment on May 6th, 2018. At that date, 297 persons (198 males, 96 females and 3 non-informed) had participated. The results are presented in appendix E.

6.3.2 Pre-experiment 2

As it is known that dilemma properties have a significant impact on decision making (Greene et al., 2001), we decided to conduct a second pre-experiment in order to classify dilemmas into two categories:

- consequentialist dilemmas: moral dilemmas where a majority of people select the consequentialist decision
- deontological dilemmas: moral dilemmas where a majority of people select the deontological decision

[5]These arguments were designed by us.

The participants were asked several questions on ten dilemmas randomly selected out of the twenty-four available dilemmas:

- Step 1: evaluate on a 7-point Likert scale the intensity[6] of the dilemma and their responsibility[7] facing this dilemma.
- Step 2: select the decision you would make among the two consequentialist and deontological decisions that are proposed.

The form was presented shown as in figure 6.2.

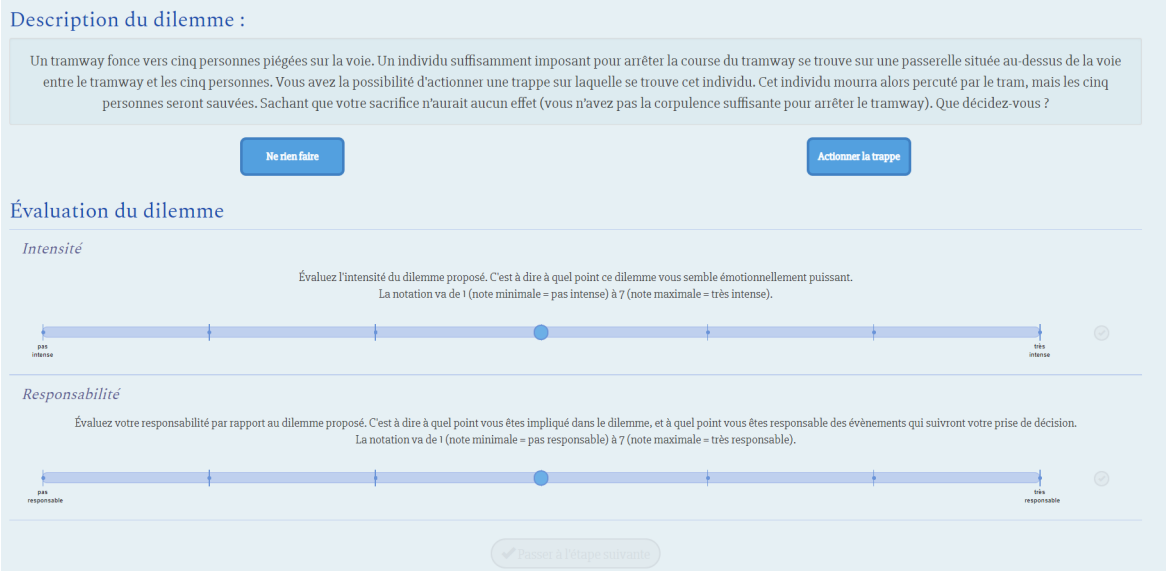


Fig. 6.2 Online pre-experiment 2

We gathered the results of this pre-experiment on May 6th, 2018. At that date, 210 persons (144 males, 63 females and 3 non-informed) had participated. The results are presented in appendix F.

The results of the pre-experiments allowed us to select eighteen dilemmas among the twenty-four (nine consequentialist dilemmas and nine deontological dilemmas). This selection was made by first assessing the rate of deontological/consequentialist decision made in order to set a category to a dilemma, and then coupling one dilemma of the deontological category with one of the consequentialist category through a similar (i.e., as close as possible) rate of the prevailing decision (see figure 6.3).

Considering figure 6.3, we can see that facing "Le tramway" (i.e., Trolley dilemma), 72.9% of participants make the consequentialist decision. Then, this dilemma has

[6] how much this dilemma has an emotional impact on you
[7] how much you feel responsible of the events induced by the decision made

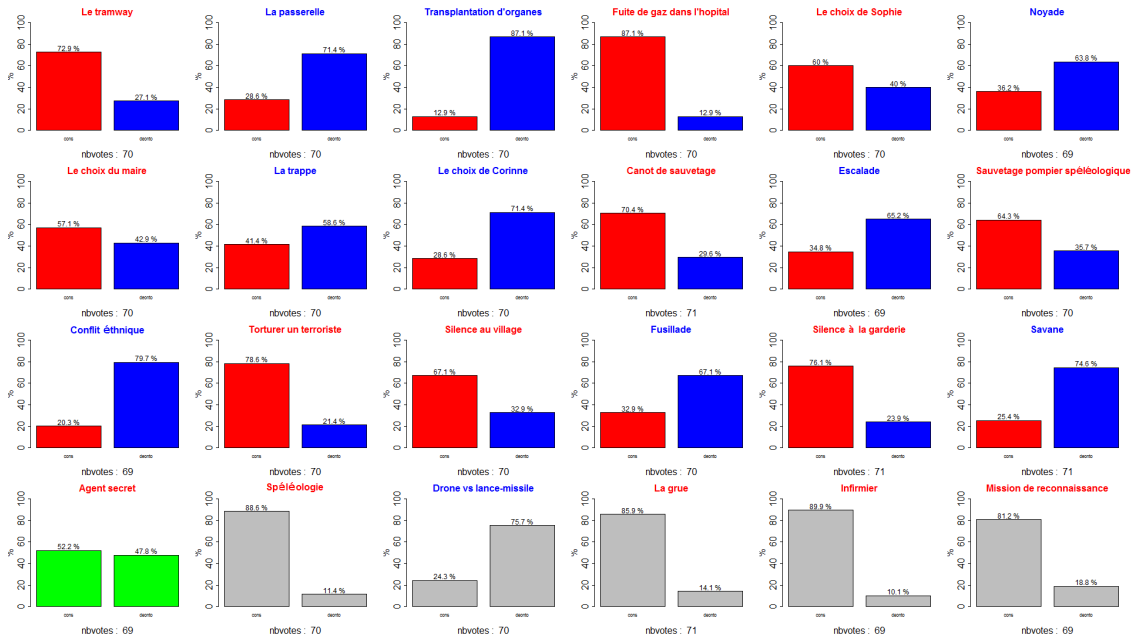


Fig. 6.3 Pre-experiment 2: results

been categorized as consequentialist, contrary to "La passerelle" (i.e., Footbridge dilemma) categorized as a deontological dilemma with 71.4% of participants making the deontological decision about this dilemma.

Hence, dilemmas have been classified as such:

Deontological dilemmas

1. Footbridge dilemma ("La passerelle")
2. Organ transplantation ("Transplantation d'organes")
3. Drowning ("Noyade")
4. The hatch ("La trappe")
5. Corinne's choice ("Le choix de Corinne")
6. Climbing ("Escalade")
7. Ethnical conflict ("Conflit éthnique")
8. Shooting ("Fusillade")
9. Savannah ("Savane")

Consequentialist dilemmas

1. Trolley dilemma ("Le tramway")
2. Gas leak at the hospital ("Fuite de gaz dans l'hopital")
3. Sophie's choice ("Le choix de Sophie")
4. Mayor's choice ("Le choix du maire")
5. Lifeboat ("Canot de sauvetage")
6. Rescue speleology ("Sauvetage pompier spéléologique")
7. Torture a terrorist ("Torturer un terroriste")
8. Silence in the village ("Silence au village")
9. Silence at the nursery ("Silence à la garderie")

This categorization was justified with a Student test[8]: at a specified .05 level, $t(14,3) = -.66$, $p = .52$, $d = .30$, 95% CI $[-9.5, 5.0]$

Therefore, it is correct to compare the participants' behaviours under the type of dilemma condition as these categories obtain equivalent amounts of votes corresponding to their categories (the deontological dilemmas obtain an amount of deontological decisions equivalent to the amount of consequentialist decisions obtained by the consequentialist dilemmas).

6.4 Experiment

With the material built from both pre-experiments, the main experiment could take place. This experiment was performed with 49 participants at ISAE-SUPAERO in DCAS laboratory in collaboration with Eve Fabre, Grazia Pia Palmiotti and Frédéric Dehais from July the 1st 2018 to September the 11th 2018.

6.4.1 Material and methods

Ethical Committee agreement

This experiment received agreement 2017-040 from "CERNI-Université fédérale de Toulouse"[9]. Participants were informed of their rights and gave a written informed consent for participating in the study. The file submitted to CERNI so as their agreement can be seen in appendix C.

[8] This test is used to compare if two groups are significantly different. See appendix B for notations

[9] Ethics Committee from the Federal university of Toulouse

Participants

Thirty-nine French people (14 females, $M_{age} = 25$,SD $\pm$ 5) participated in the study. None of the participants reported a history of prior neurological disorder. They did not win anything at the end of the experiment.

Procedure

Participants are asked to make a decision in front of ethical dilemmas (which have been selected thanks to pre-experiment 2) presented as texts through a software.The procedure consists in two steps:

- During the first step, the participant has to make a decision for each of the eighteen dilemmas randomly selected. The participant is alone in front of the dilemma.
- During the second step, the participant faces the same dilemmas twice (still in a random order) and has to confirm, or not, the decision that is pre-selected by the artificial agent. The artificial agent is symbolized by a second screen located next to the main screen. One time, the deontological decision is pre-selected and supported by an argument of the same nature displayed on the second screen. The other time, the consequentialist decision is pre-selected and supported by an argument of the same nature displayed on the second screen. Let us remind that these arguments come from the pre-experiment 1. Each time the participant has to make a decision, they have twenty seconds to choose between validating the pre-selected choice or selecting the other decision. The participant has been notified previously that, if no decision is made during the twenty seconds, the pre-selected decision is chosen by default[10]. The aim of this rule is to encourage participants to make a decision, and to identify the cases where people have a trend to offload their responsibilities.

Dilemmas are presented as paragraphs describing a situation and asking the participant to make a decision. Two possible decisions are described as a text under the paragraph (see figure 6.4). The participant's decisions are recorded thanks to a classical keyboard. During the first step, the participant has to select a decision to make. During the second step, the participant has to confirm or not the highlighted decision (i.e., the decision that is pre-selected by the artificial agent). Before making

[10]Nonetheless, the omission is recorded as such in the results

this choice, the participant has to take note of the argument displayed on the second screen (i.e., the argument supporting the pre-selected decision).

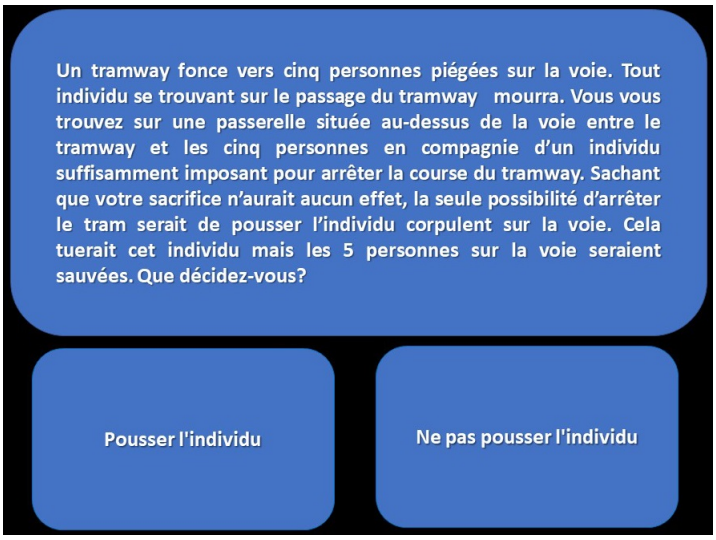
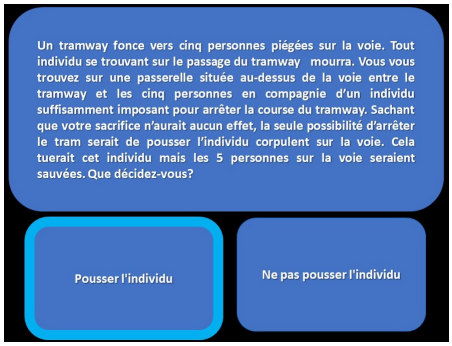


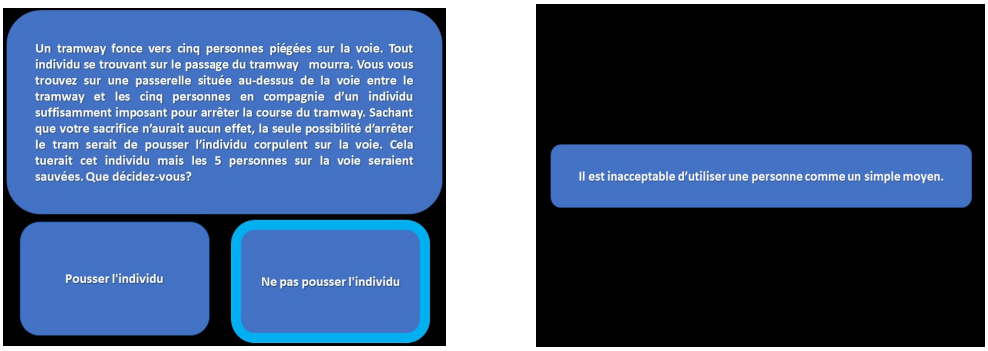
Fig. 6.4 Dilemma displayed during first step



(a) Lefthand screen with dilemma and decisions

(b) Righthand screen with argument

Fig. 6.5 Screens during second step with consequentialist decision pre-selected



(a) Lefthand screen with dilemma and decisions

(b) Righthand screen with argument

Fig. 6.6 Screens during second step with deontological decision pre-selected

Experiment procedure

The participants are welcomed in the experimental room. They first read and sign the information notice and inclusion questionnaire (see appendix C). Afterwards, a global presentation of the experiment is detailed. Finally, the experimenter sets up the fNIRS[11].

A training example is proceeded in order for the participant to understand the way to make a decision. This training example is divided in two parts: decision making alone, and decision making with a pre-selected decision of the artificial agent supported by an argument.

The experimenter then leaves the room in order to let the participant alone to start the experiment, which lasts approximately one hour without any break. At the end of the experiment, the participant is asked to fill several questionnaires described below.

Material

During the experiment, participants are seated comfortably in an isolated room, which is dark and sound-attenuated, in front of a table with the two screens and the keyboard. The software window with the textual dilemmas and possible decisions is displayed on the lefthand screen. The argument supporting the decision that is pre-selected by the artificial agent is displayed on the righthand screen.

Textual information is presented on a high-resolution screen positioned at eye level 70 cm in front of the participant. A white fixation point (+) on a black background is displayed until the participant presses the space bar to start the experiment.

[11]Neurophysiological activities have been collected through a fNIRS device. Nonetheless, analysing this data is out of the scope of this thesis and will be done in a future work

Each participant was confronted to the 18 randomized dilemmas displayed 3 times (without any artificial agent, with a deontological artificial agent, and with a consequentialist artificial agent) for a total of 54 situations per participant. Because we assume that the acceptance of technology and suggestibility may vary among participants, post-experiment questionnaires have been included in the process. Participants were asked to fill:

- a French version of the NARS (Negative Attitudes towards Robots Scale) questionnaire to evaluate their attitude towards robots in general (Nomura et al., 2006)
- a French version of the IRI (Interpersonal Reactivity Index) questionnaire to assess the degree of the participants' empathy (Gilet et al., 2013)
- a French version of a questionnaire about compliance to assess the degree of the participants' compliance[12] (Gudjonsson, 1989)

6.4.2 Data analysis

This experiment has provided a significant amount of data which require to be analysed precisely. The following results are a first analysis of the behavioural results, which will be deepened in the future.

An initial work we performed was to categorize participants. Indeed, we assume that some people tend to make more consequentialist decisions while other people tend to make more deontological decisions. Therefore, we decided to categorize participants according to the type of decision they made the most during first step. We then obtained a "Participant's type": twenty consequentialist participants, fifteen deontological participants and four participants without major preference. The results have then validated this categorization (see paragraph 6.4.3).

The following analysis of the results is divided into three models[13]:

- ***Model 1:*** *Decision without the artificial agent versus Decision when the artificial agent agrees*:
  we compared the choices of the participants when they had to make their decision without the artificial agent pre-selecting a decision (control condition) and when the artificial agent proposed the same choice as they did in the control condition,

[12] measuring the compliance consists in measuring to what extent participants change their mind to avoid conflict when someone disagrees with them.

[13] a model is a part of the analysis focusing on specific conditions and effects

in order to check that the participants do not reject the artificial agent even when it agrees with them.
A $2 \times 2 \times 2$ (Dilemma type [consequentialist, deontological] $\times$ Participant's type [consequentialist, deontological] $\times$ Treatment [No Artifical Agent, Agent Agreement]) binary logistic regression was performed.

- ***Model 2:*** *Decision without the artificial agent versus Decisions when the artificial agent disagrees*:
  we compared the choices of the participants when they had to make their decision without the artificial agent pre-selecting a decision (control condition) and when the artificial agent proposed the decision opposed to the one they had chosen in the control condition, in order to measure to what extent the artificial agent influences the decision of the participants.
  A $2 \times 2 \times 2$ (Dilemma type [consequentialist, deontological] $\times$ Participant's type [consequentialist, deontological] $\times$ Treatment [No Agent, Agent Agreement]) binary logistic regression was performed.

- ***Model 3:*** *"Laisser-faire" (i.e., participant not responding, also called an omission) versus Actively responding when artificial agent pre-selects a decision*:
  During the step where the artificial agent actively pre-selects a decision, we compared when participants made a decision actively and when participants let the time (twenty seconds) elapse. We assume that this behaviour has an important meaning such as responsibility offloading.
  A 2 $\times$ 2 $\times$ 2 (Dilemma Type [consequentialist, deontological] x Participant's Type [consequentialist, deontological] x Artificial agent's choice [agreement, disagreement]) binary logistic regression was performed.
  A manual stepwise analysis was performed to remove non-significant interactions from the model. Contrasts between conditions are reported for significant effects.

6.4.3 Results

Decision without the artificial agent versus Decision when the artificial agent agrees

The model revealed a main effect of Participants' Type [$B(SE) = -1.674(.205)$, CI $(95\%) = (-2.075, -1.272)$, $Wald\mathcal{X}^2(1) = 66.72$, $p < .001$] with deontological participants predicting for a greater number of deontological decisions (Deontological decisions: $M = 70.19\%$, Consequentialist decisions: $M = 29.81\%$, $SD = 11.57$) and consequentialist

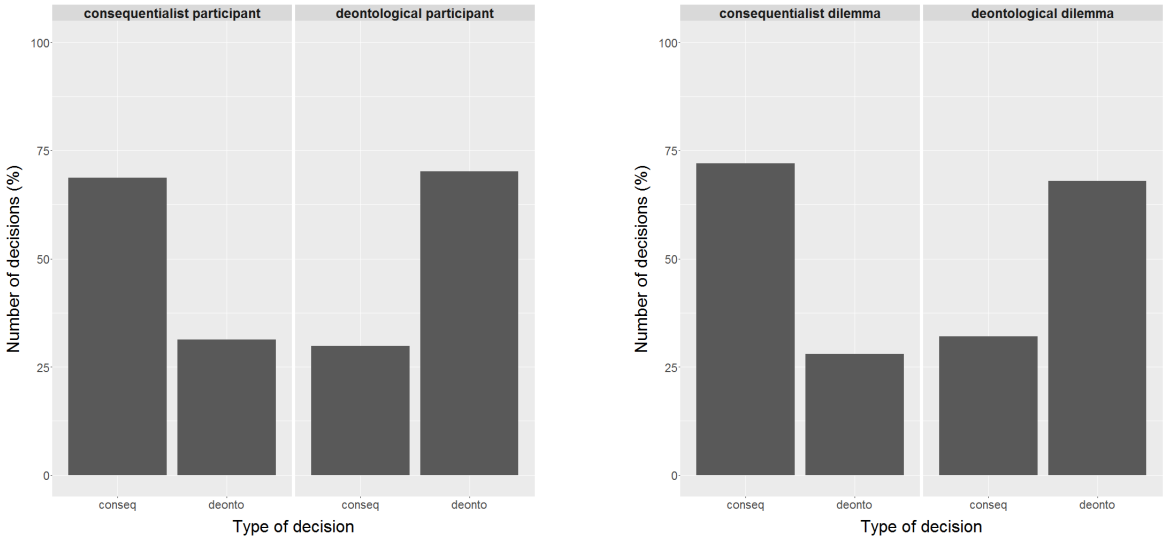
participants predicting for a greater number of consequentialist decisions (Consequentialist decisions: $M = 68.75\%$, Deontological decisions: $M = 31.25\%$, $SD = 11.32$). This means that our categorization of participants is relevant.

In the same way, the model revealed a main effect of Dilemma Type [$B(SE) = 2.566(.322)$, CI $(95\%) = (1.935, 3.196)$, $Wald\mathcal{X}^2(1) = 63.614$, $p < .001$] with deontological dilemmas predicting for a greater number of deontological decisions ($M = 67.94\%, SD = 13.39$) and consequentialist dilemmas predicting for a greater number of consequentialist decisions ($M = 72.06\%, SD = 13.70$). Therefore, our categorization of dilemmas is also relevant.

The analysis also revealed a significant Participant's Type $\times$ Dilemma Type [$B(SE) = -.888(.350)$, CI $(95\%) = (-1.574, -.203)$, $Wald\mathcal{X}^2(1) = 6.445$, $p < .05$], which is a consistent result in accordance with the two results above.

The model revealed no significant effect of Treatment [$B(SE) = .074(.052)$, CI $(95\%) = (-.029, .177)$, $Wald\mathcal{X}^2(1) = 1.993$, $p = .158$], demonstrating a stability of the participants' decisions in the control condition and when the robot was in accordance with the decisions made by the participants in the control condition.

To sum up, participants make decisions in accordance with their category: deontological (resp. consequentialist) participants tend to make deontological (resp. consequentialist) decisions, and dilemmas receive a majority of decisions corresponding to their category. Moreover, there is no significant difference between the first time the participants make a decision facing a dilemma and when they face the same dilemma with the artificial agent in accordance with their first decision (i.e., the participants usually make the same decision facing the same dilemma when the artificial agent pre-selects the decision they had made previously).



(a) Participant categories (b) Dilemma categories

Fig. 6.7 Validation of participant and dilemma categorization through the rate of decisions recorded without the agent and with the agent agreeing with participants

Decision without the artificial agent versus Decisions when the artificial agent disagrees

The model revealed a main effect of Participants' Type [$B(SE) = -1.946(.239)$, CI $(95\%) = (-2.414, -1.478)$, $Wald\mathcal{X}^2(1) = 66.456$, $p < .001$] with deontological participants predicting for a greater number of deontological decisions (Deontological decisions: $M = 69.81\%$, Consequentialist decisions: $M = 30.19\%$, $SD = 9.78$) and consequentialist participants predicting for a greater number of consequentialist decisions (Consequentialist decisions: $M = 68.61\%$, Deontologicial decisions: $M = 31.39\%$, $SD = 10.74$).

The model revealed a main effect of dilemma type [$B(SE) = 2.213(.340)$, CI $(95\%) = (1.547, 2.879)$, $Wald\mathcal{X}^2(1) = 42.423$, $p < .001$] with deontological dilemmas predicting for a greater number of deontological decisions ($M = 67.78\%, SD = 12.56$) and consequentialist dilemmas predicting for a greater number of consequentialist decisions ($M = 72.06\%$, $SD = 12.44$).

The analysis also revealed a significant Treatment $\times$ Participants' Type [$B(SE) = .330(.166)$, CI $(95\%) = (.006, .655)$, $Wald\mathcal{X}^2(1) = 3.978$, $p < .05$].

The contrast analysis revealed no significant difference of interest. In other words, the participants tend in majority to make the same decision they had made previously,

even when the artificial agent pre-selects and supports the opposite decision. Nonetheless, this artificial agent's opposition has a significant (even if reduced) influence on the participants' decisions, which is described below.

The model revealed a significant Dilemma Type × Treatment × Participants' Type interaction [$B(SE) = -.843(.262)$, CI (95%) $= (-1.356, -.329)$, $Wald\mathcal{X}^2(1) = 10.340$, $p < .001$]. The argumentation of the artificial agent did not modify the decisions of consequentialist participants in response to both deontological (control condition: $M = 46.67\%$ of consequentialist decisions, $SD = 18.24$; agent condition: $M = 46.67\%$ of consequentialist decisions, $SD = 21.20, p = 1.000$) and consequentialist dilemmas (control condition: $M = 92.22\%$ of consequentialist decisions, $SD = 9.60$; agent condition: $M = 88.89\%$ of consequentialist decisions, $SD = 12.49, p = .06$). However, the agent's argumentation impacted the decisions of deontological participants in response to deontological dilemmas (control condition: $M = 85.18\%$ of deontological decisions, $SD = 6.85$; agent condition: $M = 88.98\%$, $SD = 5.94, p < .01$), but not in response to consequentialist dilemmas (control condition: $M = 54.07\%$ of deontological decisions, $SD = 16.19$; agent condition: $M = 51.11\%$, $SD = 16.69, p = .226$).

However, the model revealed no significant main effect of Treatment [$B(SE) = .000(.114)$, CI (95%) $= (-.223, .223)$, $Wald\mathcal{X}^2(1) = .000$, $p = 1.000$], or significant Participant's Type × Dilemma Type [$B(SE) = -.178(.373)$, CI (95%) $= (-.909, .553)$, $Wald\mathcal{X}^2(1) = .228$, $p = .633$], or Treatment × Dilemma Type [$B(SE) = 393(.226)$, CI (95%) $= (-.050, .837)$, $Wald\mathcal{X}^2(1) = 3.028$, $p = .082$] interactions.

To sum up, there is one significant result of interest of the artificial agent disagreeing with the first decision of participants that is observed when deontological participants face deontological dilemmas. In that case, the rate of deontological decisions is greater. This effect will be discussed in chapter 7.

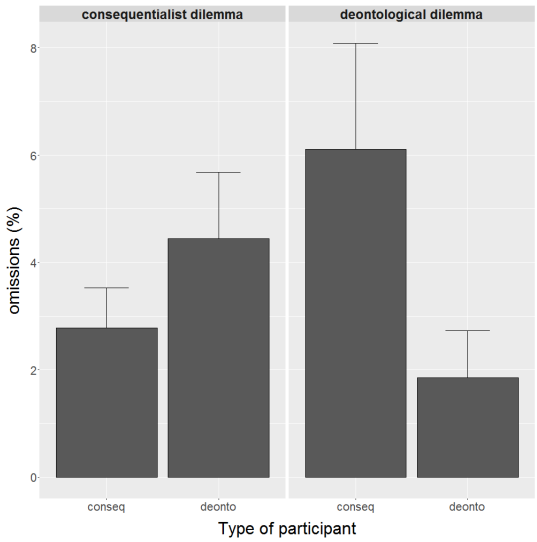
"Laisser-faire" (i.e., omission) versus actively responding when artificial agent is pre-selects a decision

The Dilemma Type × Participant's Type interaction was significant [$B = 1.733$, $Wald\mathcal{X}^2(1) = 6.768$, $p < .01$]. Deontological dilemmas predicted for greater laisser-faire in consequentialist participants ($M = 6.11\%$, $SD = 8.81, p < .01$) than in deontological participants ($M = 1.85\%, SD = 3.43$). In consequentialist participants, deontological dilemmas ($M = 6.11\%$, $SD = 8.81, p < .05$) predicted for greater laisser-faire than

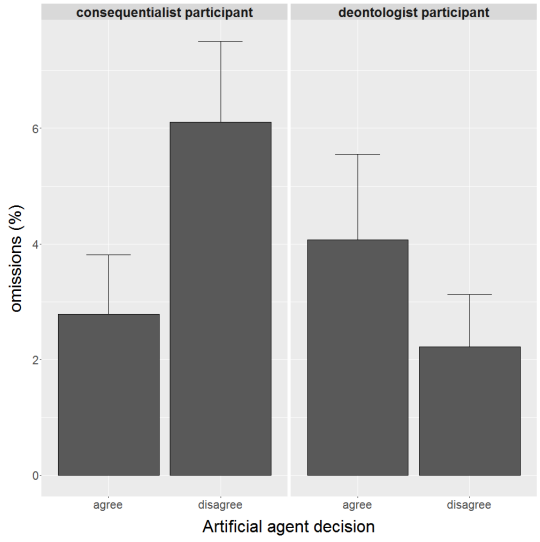
consequentialist dilemmas ($M = 2.78\%1.85$, $SD = 3.373.43$).

The model also revealed a significant Participants' Type × Artificial agent's Choice interaction [$B = 1.457$, $Wald\mathcal{X}^2(1) = 5.071$, $p < .05$]. In deontological participants, agreeing artificial agent ($M = 4.07\%$, $SD = 5.74, p < .05$) predicted for a higher level of "laisser-faire" than disagreeing with agent ($M = 2.22\%$, $SD = 3.51$). Disagreeing artificial agent predicted a higher level of "laisser-faire" in consequentialist participants ($M = 6.11\%$, $SD = 6.22, p < .05$) than in deontological participants ($M = 2.22\%$, $SD = 3.51$).

Participant's type [$B = -1.940$, $Wald\mathcal{X}^2(1) = .978$, $p < .323$], Artificial agent's choice [$B = -.829$, $Wald\mathcal{X}^2(1) = .95$, $p < .757$] and Dilemma's type [$B = -1.940$, $Wald\mathcal{X}^2(1) = .978$, $p < .323$] were not predictive of the decision to actively decide or not.



(a) Consequentialist participants are more subject to omission when facing deontological dilemmas



(b) Consequentialist participants make less omissions when artificial agent agrees. Deontological participants make less omissions when artificial agent disagrees

Fig. 6.8 Illustration of the "Laisser-faire" effect

To sum up, on the one hand we observe that consequentialist participants are more subject to omission than deontologist participants while facing deontological dilemmas. Moreover, consequentialist participants tend to make more omissions when facing deontological dilemmas than when facing consequentialist dilemmas. On the other

hand, deontological participants tend to make more omissions when the artificial agent agrees with them than when the artificial agent disagrees. Finally, consequentialist participants are more subject to omission than deontological participants when the artificial agent disagrees with them.

Chapter 7

Discussion

Goal: *Discuss the strengths and weaknesses of the proposed work. Summarize the choices and the subjectivity highlighted through the formalization.*

This chapter will discuss the results of the contribution detailed in the previous chapter.

7.1 Ethical frameworks modelling

A part of the components of our ethical frameworks model is designer dependent. This induces several biases that may reflect the designer's moral, ethics or ways of usually considering things in a given society. This section aims at discussing these biases (Grinbaum et al., 2017).

It also aims at highlighting and locating the subjectivity inherent in any formalism, and the difficulties at stake while reasoning on moral dilemmas with ethical concepts.

7.1.1 Facts

Even if facts are not directly related to value judgements, it is important to keep in mind that knowledge describing the world is obtained from sensors that are selected and calibrated by humans. Moreover, once data are acquired, the way they are interpreted is also designer dependent. Therefore, we argue that data are always biased. As highlighted by Johnson (2014), neutrality of data and facts is limited by our perceptions. Indeed, data collection is always guided by a purpose, which tends to narrow perception down to facts confirming this purpose (confirmation bias (Oswald and Grosjean, 2004)). In the same way, it is possible neither for humans nor for

machines to be exhaustive while describing a situation. This is why a subset of facts that matters or that is considered to matter has to be selected.

7.1.2 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? In order to avoid the troubles of such questions, our approach assumes positive and negative facts that are quite easy to apprehend. These facts are ranked in order to prefer a set of facts to another set of facts, which allows to rank consequences to choose the best result. For instance, $Positive(\{h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\}) = \{h\}$ expresses the fact that keeping civil human beings alive is considered to be the positive fact within set $\{h, \overset{\circ}{d}, \overset{\circ}{o}, \overset{\circ}{s}\}$. It is worth noticing that in a given context (e.g., a military context) what is considered as positive or negative may change depending on particular circumstances.

In the same way, our preference relation between fields ($>_\phi$) is based on a very strong assumption: whatever the facts a and b respectively belonging to ϕ_a and ϕ_b are, if ϕ_a is preferred to ϕ_b, a is always preferred to b regardless of the magnitude of facts, for instance: $field_{human} >_{field} field_{goods}$, meaning that human lives are considered more important than goods.

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.

7.1.3 Deontological ethics

Our ethical judgement from the deontological view point is based on the assessment of the nature of the decision. The question is: how to evaluate the nature of a decision, on what basis? Why is $DecisionNature(\text{to kill}) = bad$? Indeed this depends on the context. For example informing on a criminal or on 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*?

For the trolley dilemma, we have assumed that both "do nothing" and "push the switch" decisions are legitimate as they are not related to actions judged as bad such as "lie", "kill", etc. Nonetheless, they could be evaluated differently.

7.1.4 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 impossible. It is worth noticing that this raises the question of causality of facts, that we have only considered for the Doctrine of Double Effect.

- *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[1]: indeed an agent that is unable to predict a fact can hardly be responsible for the occurrence of this fact. However, we consider that an agent is responsible for the consequences related to a "do nothing" decision, as it is aware of the consequences of this decision.

- *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, 1993), we have computed the consequences for the agent and for the relevant (from the agent's view point) people and goods concerned by the dilemma.

- *Preference relation between facts* Preference relations between facts $>_u$ and $\succ_u$ are subjective value judgements. For example in the Footbridge dilemma, we prefer "five people alive" to "fatman alive" ($f_5 >_u fat$) and "fatman dead" to "five people dead" ($\overset{\circ}{fat} >_u \overset{\circ}{f_5}$). This has been set from the assumption that one life is equal to another. This is debatable and could be considered differently if we had more information on who fatman and the five people are. In this case, it is important to be aware of the meaning of such a reasoning, which consists

[1] It is worth noticing that the phrase "responsible" is used here as "causing the effect", not "morally responsible". Indeed, we assume an artificial agent as never morally responsible, contrary to a human or a legal entity.

in attributing different values to people. Doing this is highly debatable from an ethical view point, and rises many issues about the criteria used to compute these values.

7.1.5 Doctrine of Double Effect

Let us focus on the three DDE rules:

- *Deontologicial rule*: the nature of decision has been discussed in the Deontology paragraph above.

- *Collateral damage rule*: while quite straightforward considering the footbridge dilemma, the causality at the core of this rule is not that simple in every circumstances. For instance, considering Sophie's choice (Styron, 2004), does she cause the death of her child as both her children would have die otherwise? How to evaluate the bad effect as a means? The answer may lie on studying responsibility. This way, it is possible to categorize different kinds of responsibilities, some of them may be identified as equivalent to a causal link between facts. Another answer may be found using temporal logic (Pinto and Reiter, 1993), allowing to identify which fact has been necessary in the past to obtain another fact. An interesting work has been done on this issue in (Berreby, 2018), see paragraph 2.6.1.

- *Proportionality rule*: how to assess the balance between good and bad effects, i.e., how to evaluate the proportionality between facts? It is indeed a difficult issue, as proportionality is at the same time a subjective and a universal concept: while using insecticide to kill a cockroach is assessed proportional by a wide majority of people, using a nuclear weapon for the same end is not. On the contrary, the cost/benefit balance of any decision is subjective, such as the fair amount of money to spend for an object or an amusement. Nonetheless, the notion of proportionality is often used in law, such as in the Geneva Conventions about humanitarian treatment in war or in French legitimate defence law:

  "Is not penally responsible a person who, in front of an unjustified infringement toward themselves, accomplish at the same time, an act commanded by the necessity of the legitimate defence of themselves or others, ***except if there is a disproportion comparing defence means used and the gravity of the infringement****..." Translated from French Penal Code - Article 122-5*

The proportionality as it is defined in this work is debatable. Indeed, we have not defined proportionality in itself but a condition related to the cost/benefit balance that is true when there is proportionality between facts.

A borderline case where decisions only produce negative facts is worth considering. Indeed in this case, the collateral damage rule is always satisfied. Nevertheless the proportionality rule is not computable as "not obtaining" the set of positive facts does not make sense. It is worth noticing that the Doctrine of *Double Effect* has been designed to solve dilemmas with *a positive and a negative* effect. Therefore, it is irrelevant to use this framework with decisions only inducing negative facts.

7.1.6 Aggregation criteria

We have suggested to extend preference relation $>_u$ to relation $\succ_u$ between sets of facts thanks to aggregation criteria. The proposed criteria are simple examples of what kind of criterion could be used. We observe that the relevancy of criteria depends on the application. For instance, the OFE criterion has appeared irrelevant while studying the proportionality rule: according to this criterion, the rule is always respected.

Therefore, the aggregation criteria have to be carefully thought and studied before implementation.

7.2 Moral dilemma formal definition

More choices and issues can be highlighted by the dilemma definition we have proposed.

7.2.1 Resolvable conflicts

Let us consider an example called "Drowning" which can be considered as a "resolvable conflict":

"While driving to an important meeting with your manager, you see on the side of the road a child calling for help in the middle of a pond. By saving the child, you will be late for your meeting, while letting the child will cause their death."

For most people, saving a child seems much more important than avoiding to be late. Consequently, this example is not a moral dilemma for common sense. Nonetheless, our dilemma definition does not encompass this notion of resolvable conflict and considers this kind of situation as a dilemma. As a first step, we have proposed a formal definition

avoiding any preference order, giving a wide definition that identifies moral dilemmas but also situations that are not dilemmas for most people. Another approach would have been to propose a narrower definition that identifies only moral dilemmas of common sense, but that might miss some situations that are indeed moral dilemmas of common sense. Therefore our definition choice is simpler, but enlarges the set of situations detected as dilemmas. Considering a system with an artificial agent supervised by a human operator, our definition is likely to be relevant when being sure not to miss moral dilemmas is more important than being sure that the encountered situation is really a dilemma, such as e.g., in an environment where dilemmas are rare. On the contrary, such a wide definition is likely to be problematic if resolvable conflicts are numerous, as the operator may be overloaded with notifications, making the system inefficient.

Our definition is wide mainly because of references being equally considered. Indeed, both *punctuality* and $life\ respect$ are references that induce a decision to be unsatisfactory in the Drowning situation. A solution could be then to set a preference order between references. That way, considering that $life\ respect$ is preferred to *punctuality* could lead to identify that one decision is imperative, and thus that the situation is not a dilemma. Furthermore a partial order of preferences is worth considering: indeed, allowing some references not to be comparable with each other is essential to allow moral dilemmas to be identified. For instance, not comparing references *conseq* and *deonto* in the Footbridge dilemma (paragraph 5.5.2) which lets the situation be a dilemma, seems reasonable.

7.2.2 Uncertainty

Uncertainty is a major issue while considering real world modelling. For now, our model does not encompass uncertainty as we assume events resulting from the decisions so as the consequences of the events as certain. A perfect example of this assumption is located in the monitoring dilemma (paragraph 5.4): we have assumed that reporting to the doctor will make the doctor aware of the patient's behaviour. Nonetheless, the doctor may not read the report. Moreover, taking uncertainty into account is especially relevant while dealing with moral dilemmas as some situations are moral dilemmas because of uncertainty. In the medical field for instance, to propose a surgery is often a moral dilemma because of the possibility of failure, or because of possible bad outcomes. For instance, considering the case of a very premature baby, a moral dilemma of saving at any cost this baby exists because of the possible and uncertain after-effects this baby might suffer from in the future.

Nonetheless, uncertainty is most of the time tackled in computer science through probabilistic theories, and thus mainly with numbers. Even though the efficiency of such approaches is not criticized here, they are likely to be inaccurate when dealing with ethics: if numbers are the only way to select a decision, the decision is no more justifiable otherwise than "having the best score". Moreover, we have already argued against quantification (see paragraph 2.6.4): for instance, is it ethical to make decisions based on utility scores when lives are at stake?[2]

7.2.3 Multi-references

An agent may consider a fact through multiple references. In that case, a fact may be considered either as positive or negative according to each reference.

For instance in our monitoring dilemma example, we could consider an agent *sga* both guided by its goal and subordinate to the patient:

$$\mathcal{R}_{sga} = \{sub, goal\} \tag{7.1}$$

Through *sub*, fact *aware* would be set to negative while not set to negative (i.e., positive) according to *goal*.

$$Neg(aware, sga, sub) = true \tag{7.2}$$

$$Neg(aware, sga, goal) = false \tag{7.3}$$

Nonetheless, considering our definition 5.9, the decision is unsatisfactory[3]. This definition is based on a choice and many other possibilities are available. Indeed it is possible, as said earlier, to imagine some orders of preference between references. Another possibility is to compare in some ways the negative assessments to the positive assessments (i.e., $Neg(f, a, r) = false$) according to the agent's references.

We can see here how important the references are. Indeed, they catch a wide notion of ethics, such as context-sensitivity. In the monitoring dilemma, it is because the patient asks the agent not to pursue its goal that "subordination" (*sub*) is considered to evaluate the fact.

Consequently, a reference is inseparable from the agent and from the application domain. Nonetheless, it is important to keep in mind that the agent may use (and

[2] As the issue was risen in the famous movie "I, Robot" (Proyas, 2005).

[3] As facts evaluated as not negative are not taken into account. Indeed, one reference assessing the fact as negative is sufficient to make the decision unsatisfactory

interpret) several different references to assess differently negative/not negative facts and decision natures.

7.2.4 Subjectivity in the model

While modelling the situation, some concepts allow agent to express subjectivity:

- *References.* References are at the core of the moral assessment[4]. Therefore the set of references used by an agent makes its assessments vary a great deal such as it is shown in the monitoring dilemma with agents *aa*, *sa*, *ba* and *nea*.
- *Assessment.* The way a reference is interpreted varies from one agent to another. Therefore, this interpretation has a direct impact on the assessments of facts and decision natures through references. That is why the agent explicitly appears as a term in functions $BadNature$ and Neg.

7.2.5 Factors of results sensitivity

Let us remind that a single bad nature/negative assessment that is not computed results in a dilemma situation not being assessed as such. Therefore, being aware of the different sensitivity sources of the model that have an impact on dilemma identification is important: the modelling choice of facts and more generally of situations embeds a part of subjectivity. For instance, some facts are modelled whereas some others are not. Even in the case of a fact detected by an artificial agent, sensors cannot always capture reality because of design choices[5]. That may result in incomplete information.

In the same way, it is not always possible to get all the possible decisions in a given situation. In the case of dilemma assessments, this is a crucial issue because any satisfactory decision could turn a dilemma to a non-dilemma situation. Missing decisions may come from several different issues, such as a too complex computing, or a modelling choice. The same kind of issue is addressed in Kasenberg et al. (2018)[6], arguing that artificial agents should be able to identify the incompleteness of situations in order to allow them to make the decision of searching for more information.

Finally, it remains a last question which belongs more to philosophy: how to properly locate the limit between decision and consequences? This question is much more tricky than it seems. Indeed, considering our monitoring dilemma, is “not

[4]Through bad nature of decision and negative facts assessments

[5]For instance because of goal oriented design

[6]In this article, the artificial agent believes it is facing a dilemma because a satisfying decision is available but not perceived.

respecting the will of the patient" ($\overset{\circ}{will}$) inherent in decision "report the doctor", and thus a way to evaluate the nature of the decision as bad, or a consequence of this decision? Let us consider another example: *"An empty autonomous car, on its way to pick up passengers, is halted at a red traffic light. A car is coming from the crossing road at the left and will hit a pedestrian on the zebra crossing on the other side of the road if the autonomous car does not go through the red light and hit the car to protect the pedestrian."*
In this situation, what belongs to decision and what belongs to consequences appears to be fuzzy: indeed, is "go through the red light" a consequence of the decision (hit the car), or the decision itself?

7.3 Study of the influence of an artificial agent on human moral decision making

The experiment we conducted has produced a vast amount of data. We have presented in chapter 6 the very first analysis of the data. The following sections aim first at questioning the experiment in itself. Secondly, we propose an interpretation of the results.

7.3.1 Protocol

The first issue we have to tackle lies on the properties of moral dilemmas. Indeed, we have used the theoretical decision participants claim to make while facing situations we have described in texts. Nonetheless, these situations are tough and complex, and it is likely that people cannot be sure of how they would actually react while facing these situations for real. Moreover, we have asked participants to answer three times to the same situation (one without artificial agent, one with artificial agent pre-selecting the deontological decision and one with artificial agent pre-selecting the consequentialist decision). It is likely that making the decision alone first induce people to be less influenced by the argument of the artificial agent.

7.3.2 Material validation

Nonetheless, the pre-experiments we have realized, as well as the analysis of decisions of participants without the artificial agent and with the artificial agent agreeing, prove that our material is valid. Indeed, deontological (resp. consequentialist) participants mostly

make deontological (resp. consequentialist) decisions. In the same way, while facing deontological (resp. consequentialist) dilemmas, participants mostly make deontological (resp. consequentialist) decisions. Moreover, the participants' behaviours appear to be consistent because they tend to make the same decision when the artificial agent pre-selects and supports the decision they had made without it.

7.3.3 Influence of artificial agent disagreeing with participant

Overall, we observe that people mostly tend to maintain the decision they have made previously (i.e., without the artificial agent). Nonetheless, we have observed several effects of the artificial agent.

Omissions of consequentialist participants with artificial agent disagreeing and overall

It has appeared that consequentialist participants are more sensitive to artificial agent disagreeing with them than deontological participants. Indeed in that case, consequentialist participants make more omissions. This could be explained by the fact that people feel uncomfortable to behave less deontologically than a machine. Therefore, when the artificial agent supports a deontological decision in front of a consequentialist participant, the participant prefers to change their decision (through omission). In the same way, consequentialist participants tend to make more omissions when facing deontological dilemmas than when facing consequentialist dilemmas. A possible interpretation of such a behaviour is that participants are more confident with their decision when facing dilemmas corresponding to their category (i.e., consequentialist participants facing consequentialist dilemmas).

Omissions of deontological participants with artificial agent agreeing

It appears that there is a significant effect of the artificial agent which agrees for deontological participants facing deontological dilemmas. Indeed in that case, the amount of deontological decisions is lower than when the artificial agent disagrees. This effect might seem counter-intuitive as people seem to act in opposition to the argument of the artificial agent. Nonetheless, it is possible to relate this effect to the variation of the rate of omissions made by deontological participants facing deontological dilemmas. Indeed, another effect we have observed is that deontological participants facing deontological dilemmas tend to make more omissions when the artificial agent agrees with them than when the artificial agent disagrees. A possible interpretation

of such a behaviour is that deontological participants offload responsibilities. Indeed, knowing that the artificial agent pre-selects the decision they desire, they prefer not to make the decision, letting the artificial agent being responsible for this choice.

Chapter 8

Conclusion

8.1 Ethical frameworks modelling

We have proposed a formalism aiming at modelling a situation as well as rules allowing to judge decisions through ethical frameworks. The main challenge of this part was to formalize philosophical concepts 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. This has been achieved thanks to a formal model embedding ethical concepts. This way, an autonomous machine would be able to compute judgements and justifications about a decision.

Nevertheless, the ethical frameworks we have studied do not seem to be relevant in all situations. Considering a single ethical framework, it 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, and the Doctrine of Double Effect may not be relevant in some borderline cases[1]. In such cases some possible decisions may not be comparable. Moreover, to prefer a fact to another or to evaluate the nature of a decision depends on the context and can be tricky or even impossible. Considering the Doctrine of Double Effect, it has been highlighted that the sacrifice of oneself is always judged "unacceptable". Indeed, sacrifice itself is usually evaluated as negative, and thus, as a means to an end, the "collateral damage rule" does not allow it (3.4.4). Nevertheless when a human life is threatened, should not the agent's or the machine's sacrifice be expected?

[1] Such as when consequences have only negative facts.

This analysis has led us to think that using a single framework is neither efficient nor relevant to compute an ethical decision. It seems necessary to consider various ethical frameworks in order to obtain the widest possible view on a given situation.

8.2 Moral dilemma formal definition

We have proposed an extension of the formalism in order to define a moral dilemma formally. This model fits the ideal human-robot context in which tasks are automated and the robot rises an alarm only in case of morally problematic situations.

Nonetheless, several limits have been highlighted. Indeed, this definition mostly depends on value judgements given by the designers of the agents. Moreover, it is sensitive to several factors. Indeed, the way the situation is modelled through facts, as well as possible decisions that are considered or not, may turn a moral dilemma into a situation with a solution. On the other side, uncertainty must be taken into account as most situations include uncertain facts and consequences. Nonetheless, this uncertainty should be modelled avoiding as much as possible numerical values. Moreover, we present and argue for a computing strategy of bad/negative assessment of a decision/fact based on at least one reference. Nevertheless, one or several references may set opposite assessments. This induces that some of the situations identified as dilemmas according to our definition are actually resolvable conflicts (see paragraph 7.2.1). Finally, we have mentioned the case of situations with only satisfying decisions inducing regrets which could be considered as negative.

8.3 Subjectivity

Overall, the model described in chapters 3 and 5 is based on concepts (i.e., facts, decisions, events, etc.) value judgements (Positive/Negative facts, natures of decisions) and relations between these concepts (utilitarian preference, aggregation criteria, causality, proportionality, etc.) that embed a part of subjectivity which depends on the designer's choices. For instance, we have not described how orders are assessed: some of them have to be set, others could be learned[2], etc. Moreover it may be hardly possible to define an order (i.e., a utilitarian preference) between two facts or set of facts. In the same way, evaluating the nature of a decision is far from easy. This evaluation highly depends on which point of view is considered, such as references introduced in chapter 5.

[2]Still, what kind of learning? Based on which data?

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 as negative or positive is likely to be based on what society thinks about it. Our formalism allows the modelling choices as well as the subjectivity to be precisely identified and located.

8.4 The limits of autonomous machines

These results, combined with the idea that autonomous machines cannot be fully ethical agents (see paragraph 2.5) lead us to insist on the fact that ***autonomous machines should not be placed alone in situations requiring ethical reasoning***. Nonetheless, their numerous properties such as being reliable, computing much more faster than humans or acquiring data not accessible by humans make them a great support for a wide range of situations. Therefore, conducting research about hybrid systems embedding both abilities of autonomous machines and humans seems the most accurate way to create robust systems that can face situations requiring ethical thoughts.

8.5 Moral behavior experiment

While human-autonomous machine systems seem promising architectures, they still can fail. Indeed, several mistakes of such systems have been highlighted in the past (Lin et al., 2008), putting under the spotlight the requirement for transparency and efficient communication between human operators and artificial agents. Indeed, ethical human reasoning is subject to influence. This is why we have proposed an experiment aiming at studying the influence an artificial agent embedding ethical concepts in its reasoning may have on participants who had to make a decision while facing moral dilemmas.

Our first analysis of data has highlighted the fact that people mostly tend not to be influenced by the artificial agent. Nonetheless, consequentialist participants appear to make more omissions than deontological people when the artificial agent disagrees with the participants. On the other hand, deontological people make more omissions when the artificial agent agrees than when the artificial agent disagrees.

We have proposed a first explanation of these results: consequentialist people wish to be more deontological than the artificial agent. Indeed, as consequentialism can be associated with some kind of calculating and "coldness", and deontology with moral principles and emotions, an artificial agent making a deontological decision, and thus being more emotional and more "kind" may seem awkward and uncomfortable for a participant making the consequentialist choice, which may turn consequentialist people to change their mind. On the contrary, deontological people may wish to offload responsibilities when they can let the artificial agent make the decision they desire, which has appeared here as making more omissions when the artificial agent agrees with them.

8.6 Future works

Considering the formalism, further works will have to focus on studying other frameworks such as virtue ethics, a framework based on making the agent the most virtuous possible. Another approach already in progress is to design a value system based on norms, moral values and virtues. Such a system requires first to be able to formalize these concepts in order to link them to encountered situations. That way, an artificial agent would be able to know for known contexts which moral values and virtues are related to norms at stake in a specific context. Nonetheless, computing or learning norms and corresponding moral values/virtues is far from easy. Moreover, such a system may induce moral values to be in conflict through opposite norms.

Another future work may be to study the aggregation of ethical frameworks judgements. Nonetheless, we argue that combining these judgements is likely to make ethical frameworks loose sense. Still, the question could at least be risen.

Moreover the ethical frameworks that we have modelled need to be refined. Indeed, the way the natures of decisions are computed is for now quite arbitrary. References are a first step toward a clarification, but this need to be improved. Therefore, a future work would be to study different deontological frameworks that allow the natures of decisions to be assessed more accurately. For instance, the Kantian imperative may give us some clues to that end through evaluating the relevance of generalizing a principle. Indeed if a generalized principle is desirable, decisions fulfilling this principle should have a nature set to good. In the same way, the causality is defined as a relation but it is not clear how to assess this causality. Temporal logic or other causal formal tools have to be considered. Studying several aggregation criteria and combinations of aggregation criteria seems also relevant as it has appeared that their accuracy

depends on the ethical framework used. Finally, uncertainty is for now not tackled in our model, but it remains a fundamental issue to represent real world situations. As numerical approaches should be avoided when dealing with ethics, fuzzy logic and similar approaches should be considered.

Considering the formal dilemma definition, one future work would be to study the case of situations with only satisfying decisions inducing regrets, and to model this renouncement for instance considering positive facts induced by a decision whose corresponding regret would be captured by negative facts belonging to consequences of any other decision. For instance considering a decision for each dish at a restaurant (i.e., choosing this dish and not the others), decision *"choose the linguini"* induces a positive fact *chosen linguini*. Therefore, the renouncement could be modelled through negative fact $\overset{\circ}{chosen}$ *linguini* (i.e., linguini are not chosen) when any other decision is made (i.e., choosing another dish). Moreover, designing a partial order of preference between references might solve several issues: it may be a way to manage "resolving conflicts" as well as a way to determine a positive/negative fact when two references evaluate this fact differently. Finally, the notion of exclusive alternatives[3] has been considered as an initial assumption and thus modelling this property as a component of the formal dilemma definition has to be addressed, for instance through some temporal tools or studying resources consumption while making a decision.

Considering the experiment, neurophysiological data from fNIRS and answers of questionnaires still require to be analysed and correlated with behavioural data. Moreover, another study is in progress in order to deepen the categorization of moral dilemmas through the identification of properties which influence people preferences. Indeed, we have for now identified two categories of moral dilemmas: deontological dilemmas and consequentialist dilemmas. Nonetheless, we argue that dilemmas share more common properties that can influence decision making. Therefore, we are now processing an online experiment in which the participants have to make a decision for one random selected moral dilemma among a set of dilemmas based on the Trolley dilemma. For each dilemma one property modifies the statement; for instance the violence of the death for the person on the side track. The aim of this experiment is to identify which properties could influence decision making and to what extent they influence it.

[3] i.e., it is possible to make only one decision.

Chapter 9

General summary

This thesis comes within the scope of autonomous machines embedding ethical concepts in their reasoning. Overall, this work consists in proposing a formal reasoning allowing an artificial agent to judge decisions and justify its judgement, starting from philosophical concepts of ethics. This has been done through a situation model composed of states of the world, facts, decisions and events. Ethical frameworks of normative ethics have been formalized through rules to respect in order to judge a decision as "acceptable" or "unacceptable". Thanks to this formalism, a formal definition of a moral dilemma has been given in order to allow an artificial agent to identify a situation as such, based on dilemma properties coming from dilemma definitions of the literature formulated in natural language. This work of formalization has allowed the subjectivity sources as well as the factors of sensitivity of the model to be properly identified. A last chapter focuses on an experiment designed to study the influence an artificial agent embedding ethical reasoning may have on a human operator. The experiment reveals that most people are not influenced by the artificial agent's suggestions. Nevertheless, deontological people tend to be more influenced than consequentialist people.

References

Abel, D., MacGlashan, J., and Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, volume 92.

Acton, H. B. and Watkins, J. W. (1963). Symposium: Negative utilitarianism. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 37:83–114.

Alexander, L. and Moore, M. (2007). Deontological ethics.

Anderson, M.and 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, S. W., Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature neuroscience*, 2(11):1032.

Aristotle (2012). The nature of virtue. In Shafer-Landau, R., editor, *Ethical theory: an anthology*, volume 13, chapter 66, pages 615–629. John Wiley & Sons.

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. *AJN The American Journal of Nursing*, 80(4):658–660.

Atkinson, K. and Bench-Capon, T. (2016). Value based reasoning and the actions of others. In *Proceedings of ECAI*, pages 680–688.

Bailly, A. (1901). *Abrégé du dictionnaire grec francais*. Hachette.

Barnes, J. et al. (2014). *Complete works of Aristotle, volume 1: The revised Oxford translation*, volume 192. Princeton University Press.

BBC (2015). Drone crash causes hollywood electricity blackout. *BBC*.

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

Bentham, J. and Mill, J. S. (1961). *The Utilitarians: An Introduction to the Principles of Morals and Legislation*. Doubleday.

Bergo, B. G. (1999). *Levinas between ethics and politics: For the beauty that adorns the earth*, volume 152. Springer Science & Business Media.

Berreby, F. (2018). *Modélisation des Systèmes Ethiques*. PhD thesis, LIP6.

Berreby, F., Bourgne, G., and 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)*, pages 532–548.

Berreby, F., Bourgne, G., and Ganascia, J.-G. (2017). A declarative modular framework for representing and applying ethical principles. In *Proceedings of the 16th AAMAS*, pages 96–104.

Bonnemains, V., Saurel, C., and 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.

Bonnemains, V., Saurel, C., and Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, 20(1):41–58.

Bonnemains, V., Saurel, C., and Tessier, C. (2019). Can artificial agents recognize moral dilemmas? *Submitted to Ethics and Information Technology*.

Bonnemains, V., Tessier, C., and Saurel, C. (2017). Machines autonomes "éthiques": questions techniques et éthiques. *Revue française d'éthique appliquée (RFEA) n°5*.

Brewka, G., Eiter, T., and Truszczyński, M. (2011). Answer set programming at a glance. *Communications of the ACM*, 54(12):92–103.

Bringsjord, S. and Taylor, J. (2011). The divine-command approach to robot ethics. In *Robot Ethics: The Ethical and Social Implications of Robotics*, pages 85–108. MIT Press.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, 538(7623):20.

Cavalier, R. (2014). Meta-ethics, normative ethics and applied ethics. *Online guide to ethics and moral philosophy*.

Chisholm, R. M. (1963). Contrary-to-duty imperatives and deontic logic. *Analysis*, 24(2):33–36.

Cointe, N. (2017). *Comportement éthique d'un collectif d'agents autonomes*. PhD thesis, Université Saint Etienne.

Cointe, N., Bonnet, G., and Boissier, O. (2016). Ethical Judgment of AgentsBehaviors in Multi-Agent Systems. In *Proceedings of the AAMAS*.

Colmerauer, A. and Roussel, P. (1996). The birth of prolog. In *History of programming languages—II*, pages 331–367. ACM.

Conitzer, V., Sinnott-Armstrong, W., Schaich Borg, J., Deng, Y., and Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the 31st AAAI*.

Connell, F. (1967). Double Effect, Principle of. New Catholic Encyclopaedia (Volume 4).

d'Aquin, T. (1853). *La somme théologique*. E. Belin.

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

Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7.

Dorion, L.-A. (1997). La «dépersonnalisation» de la dialectique chez Aristote. *Archives de philosophie*, pages 597–613.

Downie, R. (1980). Ethics, morals and moral philosophy. *Journal of medical ethics*, 6(1):33.

Duranti, A. and Goodwin, C. (1992). *Rethinking context: Language as an interactive phenomenon*, volume 11. Cambridge University Press.

Evans, D. R. and MacMillan, C. S. (2014). *Ethical reasoning in criminal justice and public safety, 4th edition*. Emond Publishing.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, 5:5–15.

Gallagher, M. (1991). Proportionality, disproportionality and electoral systems. *Electoral studies*, 10(1):33–51.

Ganascia, J. G. (2007). Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9(1):39–47.

Gardam, J. G. (1993). Proportionality and force in international law. *American Journal of International Law*, 87(3):391–413.

Georgeff, M.and Pell, B., Pollack, M., Tambe, M., and Wooldridge, M. (1999). The belief-desire-intention model of agency. In Müller, J. P., Rao, A. S., and Singh, M. P., editors, *Intelligent Agents V: Agents Theories, Architectures, and Languages*, pages 1–10. Springer Berlin Heidelberg.

Gilet, A.-L., Mella, N., Studer, J., Grühn, D., and Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A french validation of the interpersonal reactivity index (iri). *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 45(1):42.

Goble, L. (2005). A logic for deontic dilemmas. *Journal of Applied Logic*, 3:461–483.

Govindarajulu, N. S., Bringsjord, S., Ghosh, R., and Sarathy, V. (2019). Toward the engineering of virtuous machines. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

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

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

Gudjonsson, G. H. (1989). Compliance in an interrogative situation: A new scale. *Personality and individual differences*, 10(5):535–540.

Horty, J. F. (1994). Moral dilemmas and nonmonotonic logic. *Journal of philosophical logic*, 23(1):35–65.

Hunyadi, M. (2019). Artificial Moral Agents. Really? In *Wording Robotics*, pages 59–69. Springer.

Hursthouse, R. (2012). Normative virtue ethics. In Shafer-Landau, R., editor, *Ethical theory: an anthology*, volume 13, chapter 68, pages 645–652. John Wiley & Sons.

Hursthouse, R. and Pettigrove, G. (2016). Virtue ethics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

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

Kagan, S. (2018). *Normative ethics*. Routledge.

Kant, E. (1795). *Métaphysique des moeurs*.

Kasenberg, D., Sarathy, V., Arnold, T., Scheutz, M., and Williams, T. (2018). Quasi-dilemmas for artificial moral agents. *arXiv preprint arXiv:1807.02572*.

Kowalski, R. and Sergot, M. (1989). A logic-based calculus of events. In *Foundations of knowledge base management*, pages 23–55. Springer.

Kuipers, B. (2018). How can we trust a robot? *Communications of the ACM*, 61(3):86–95.

Lin, P., Abney, K., and Bekey, G., editors (2012). *Robot Ethics - The Ethical and Social Implications of Robotics*. The MIT Press.

Lin, P., Bekey, G., and Abney, K. (2008). *Autonomous Military Robotics: Risk, Ethics, and Design. Technical report for the U.S. Department of the Navy*. Office of Naval Research.

Lipowski, A. and Ferreira, A. L. (2005). A model of student's dilemma. *Physica A: Statistical Mechanics and its Applications*, 354:539–546.

Lohr, S. (2015). Don't fear the robots. *New York Times*, page B1.

MacIntyre, A. (2003). *A Short History of Ethics: a history of moral philosophy from the Homeric age to the 20th century*. Routledge.

Malle, B. F. and Scheutz, M. (2014). Moral competence in social robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology*, page 8. IEEE Press.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and 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*, pages 117–124. ACM.

Mallock, D. (1967). Moral dilemmas and moral failure. *Australasian journal of Philosophy*, 45(2):159–178.

McCartney, S. and Parent, R. (2012). Ethics in law enforcement. B.V. Open Textbook Project.

McNamara, P. (2014). Deontic logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter edition.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J. (2005). The neural basis of human moral cognition. *Nature reviews neuroscience*, 6(10):799.

Monroe, A. E. and Malle, B. F. (2018). People systematically update moral judgments of blame. *PsyArXiv. August*, 30.

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

Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, 7(3):437–454.

Nussbaum, M. (2012). Non-Relative Virtues: An Aristotelian Approach. In Shafer-Landau, R., editor, *Ethical theory: an anthology*, volume 13, chapter 67, pages 630–644. John Wiley & Sons.

Nute, D. (2012). *Defeasible deontic logic*, volume 263. Springer Science & Business Media.

Oswald, M. E. and Grosjean, S. (2004). Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 79.

Paton, H. J. (1971). *The categorical imperative: A study in Kant's moral philosophy*, volume 1023. University of Pennsylvania Press.

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

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

Pollock, J. M. (2014). *Ethical dilemmas and decisions in criminal justice*. Nelson Education.

Proyas, A. (2005). "I, Robot". Davis Entertainment, Overbrook Entertainment.

Rachels, J. (2007). The challenge of cultural relativism. *Bioethics: an Introduction to the History, Methods, and Practice*.

Rapoport, A., Chammah, A. M., and Orwant, C. J. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. University of Michigan press.

Reiter, R. (1993). Proving properties of states in the situation calculus. *Artificial Intelligence*, 64(2):337–351.

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

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. (2009). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 245–246. ACM.

Ross, D. and Ross, W. D. (2002). *The right and the good*. Oxford University Press.

Sarathy, V., Scheutz, M., Kenett, Y., Allaham, M. M., Austerweil, J. L., and Malle, B. F. (2017). Mental representations and computational modeling of context-specific human norm systems. *Proceedings of Cognitive Science*.

Scharre, P. (2016). *Autonomous weapons and operational risk*. Center for a New American Security Washington, DC.

Scheutz, M. (2017). The case for explicit ethical agents. *AI Magazine*, 38(4):57–64.

Schwartz, M. S. (2005). Universal moral values for corporate codes of ethics. *Journal of Business Ethics*, 59(1-2):27–44.

Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Morales, J., Wooldridge, M., and Ansotegui, C. (2018). Exploiting moral values to choose the right norms. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 264–270.

Shafer, G. (1976). *A mathematical theory of evidence*, volume 45. Princeton University Press.

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, 9(3-4):331–352.

Singer, P. et al. (1986). *Applied ethics*. Oxford readings in philosophy.

Sinnott-Armstrong, W. (1988). *Moral dilemmas*. Wiley Online Library.

Sinnott-Armstrong, W. (2015). Consequentialism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter edition.

Styron, W. (2004). *Sophie's choice*. Random House.

Sullins, J. P. (2014). Deception and virtue in robotic and cyber warfare. In *The Ethics of Information Warfare*, pages 187–201. Springer.

Suter, R. S. and Hertwig, R. (2011). Time and moral judgment. *Cognition*, 119(3):454–458.

Sylvan, R. and Plumwood, V. (1984). *Moral Dilemmas and the Logic of Deontic Notions*. Philosophy Depts., Australian National University.

Tanner, C., Medin, D. L., and Iliev, R. (2008). Influence of deontological versus consequentialist orientations on act choices and framing effects: When principles are more important than consequences. *European Journal of Social Psychology*, 38(5):757–769.

Thaler, R. H. (1988). Anomalies: The ultimatum game. *Journal of Economic Perspectives*, 2(4):195–206.

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

Thomson, J. J. (1986). *Rights, restitution, and risk: Essays in moral theory*. Harvard University Press.

Timmons, M. (2012). *Moral theory : an introduction*. Rowman and Littlefield Publishers.

Tzafestas, S. (2016). *Roboethics - A Navigating Overview*. Oxford University Press.

Vallentyne, P. (1989). Two types of moral dilemmas. *Erkenntnis*, 30(3):301–318.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

von Wright, G. H. (1951). Deontic logic. In *Mind*, volume 60, pages 1–15. JSTOR.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR.

Wainwright, W. J. (2006). Religion and morality. *Religious Studies*, 42(2):106–107.

Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Wark, G. R. and Krebs, D. L. (1996). Gender and dilemma differences in real-life moral judgment. *Developmental psychology*, 32(2):220.

Wark, G. R. and Krebs, D. L. (2000). The construction of moral dilemmas in everyday life. *Journal of Moral Education*, 29(1):5–21.

Williams, B. A. and Atkinson, W. (1965). Symposium: Ethical consistency. *Proceedings of the Aristotelian Society, Supplementary Volumes*, pages 103–138.

Winfield, A. and Jirotka, M. (2017). The case for an ethical black box. In Gao, Y., Fallah, S., Jin, Y., and Lekakou, C., editors, *Towards Autonomous Robotic Systems*, pages 262–273. Springer International Publishing.

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

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., and Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15):6753–6758.

Appendices

Appendix A

Prolog code

A.1 formal moral dilemma definition

```
% DILEMMA

dilemma(S_init, Liste_decisions, Agent, References):-
        initial_state(S_init),
        decisions(Liste_decisions),
        length(Liste_decisions,T),
        T >= 2,
        sub_dilemma(S_init, Liste_decisions, Agent, References).

% Recursive dilemma
sub_dilemma(_,[],_,_).
sub_dilemma(S_init,[D1|Q], Agent, References):-
        unsatisfactory(S_init,D1, Agent, References),
        sub_dilemma(S_init,Q, Agent, References).

% unsatisfactory

unsatisfactory(_, Decision,A,R):- badNature(Decision,A,R).
unsatisfactory(S_init, Decision, Agent, References):-
        event_from_decision(Decision, Evenement),
        consequence(S_init, Evenement, Consequence),
        exist_neg(Consequence, Agent, References).
```

```
% Function negative
%%% for consequences

exist_neg([X|_],Agent,References):-
        fact_neg(X,Agent,References).
exist_neg([_|Q],Agent,References):-
        exist_neg(Q,Agent,References).

%%% for a fact

fact_neg(X,Agent,[Ref|_]):-
        neg(X,Agent,Ref).
fact_neg(X,Agent,[_|L_References]):-
        fact_neg(X,Agent,L_References).

% Function BadNature

badNature(Decision,Agent,[Ref|_]):-
        badNat(Decision,Agent,Ref).
badNature(Decision,Agent,[_|L_References]):-
        badNature(Decision,Agent,L_References).
```

A.2 Trolley dilemma

```
%%%%% Situation description

initial_state([f1,f5]).
decisions([do_nothing,push_switch]).

event_from_decision(push_switch,train_hits_one_person).
event_from_decision(do_nothing,train_hits_five_people).

consequence([f1    , f5],train_hits_one_person,
            [o(f1), f5]).
consequence([f1    , f5],train_hits_five_people,
            [f1    , o(f5)]).

%%%%% References

refs_ta([safety]).

%%%%% Negative assessments
neg(o(f5),ta,safety).
neg(o(f1),ta,safety).


%dilemma([f1,f5],[do_nothing,push_switch],ta,[safety]).
```

A.3 Footbridge dilemma

```
%%%%%%%%% Situation description

initial_state([fat, f5]).
decisions([do_nothing,push_fatman]).

event_from_decision(push_fatman,train_hits_fatman).
event_from_decision(do_nothing,train_hits_five_people).

consequence([fat    , f5],train_hits_fatman,
            [o(fat), f5]).
consequence([fat    , f5],train_hits_five_people,
            [fat    , o(f5)]).

%%%%%%%%% References

refs_cda([conseq,deonto]).
refs_ca([conseq]).

%%%%%%%%% Negative assessments

badNat(push_fatman,cda,deonto).
%---
neg(o(f5),cda,conseq).
neg(o(f5),ca,conseq).


%dilemma([fat,f5],[do_nothing,push_fatman],cda,[conseq,deonto]).
%dilemma([fat,f5],[do_nothing,push_fatman],ca,[conseq]).
```

A.4 Trolley variation

```
%%%%%%%%% Situation description

initial_state([f5]).
decisions([do_nothing,push_switch]).

event_from_decision(push_switch,train_is_diverted).
event_from_decision(do_nothing,train_hits_five_people).

consequence([f5],train_is_diverted,
            [f5]).
consequence([f5],train_hits_five_people,
            [o(f5)]).

%%%%%%%%% References

refs_ta([safety]).

%%%%%%%%% Negative assessments
neg(o(f5),ta,safety).


%dilemma([f5],[do_nothing,push_switch],a,[safety]).
```

A.5 Monitoring dilemma

```
%%%%%%%%% Situation description

initial_state([o(aware),priv,mon,o(info),will]).
decisions([report_the_doctor,do_nothing,warn_the_patient]).

event_from_decision(report_the_doctor,doctor_get_informed).
event_from_decision(do_nothing,monitoring_failure).
event_from_decision(warn_the_patient,patient_get_advised).

consequence([o(aware), priv    , mon    , o(info), will],
            doctor_get_informed,
            [aware    , o(priv), mon    , o(info), o(will)]).

consequence([o(aware), priv    , mon    , o(info), will],
            monitoring_failure,
            [o(aware), priv    , o(mon), o(info), will]).

consequence([o(aware), priv    , mon    , o(info), will],
            patient_get_advised,
            [o(aware), priv    , o(mon), info    , will]).
```

```
%%%%%%%%% References

refs_aa([honesty, respect, goal]).
refs_sa([sub,respect]).
refs_ba([goal]).
refs_nea([]).

%%%%%%%%% Negative assessments

%% autonomous agent %%
badNat(do_nothing,aa,honesty).
%---
neg(o(priv),aa,respect).
neg(o(mon),aa,goal).
neg(o(info),aa,goal).
neg(o(willl),aa,respect).

%% subordinate agent %%
badNat(warn_the_patient,sa,respect).
%---
neg(aware,sa,sub).

%% basic agent %%
neg(o(aware),ba,goal).

% dilemma(I,D,aa,[honesty, respect, goal]).
```

A.6 dilemma

```
%%%%%%%%% Situation description

initial_state([child    , o(late)]).
decisions([save_child,move_on]).

event_from_decision(time_is_running,save_child).
event_from_decision(child_drowns,move_on).

consequence([child    , o(late)],time_is_running,
            [child    , late]).
consequence([child    , o(late)],child_drowns,
            [o(child), o(late)]).

%%%%%%%%% References

refs_a([integrity, punctuality]).

%%%%%%%%% Negative assessments
neg(o(child),a,integrity).
neg(late,a,punctuality).
```

Appendix B

Statistics notations

B.1 Student test

The aim of a Student test is to assess the probability to make a mistake while rejecting ***H0***, H0 being the hypothesis that two groups (or a group evaluated twice regarding two conditions) are similar (i.e., there is no significant difference).

- ***p*** is the probability to make a mistake while rejecting H0. Scientists usually assume that $p > .05$ (corresponding to "at a specified .05 level") corresponds to a too high risk of rejecting H0. Thus, H0 is validated, i.e., there is no difference between the two groups. On the contrary, $p < .05$ corresponds to a significant difference between the two groups.

- ***95% CI*** is the confidence interval of the difference between the two groups. In other words, because we assume our value of difference measured as not exact, we estimate that the real value has a probability of .95% to be located inside the interval. If the value 0 is included in the interval, H0 is accepted, otherwise it is rejected.

- ***d*** is the "Cohen's D", it depicts how much the difference is important regarding the condition (i.e., the effect size). 0.20 corresponds to a small effect size while 0.80 corresponds to a big effect size.

- $t(X) = Y$ is the t value computed, with X the degree of liberty and Y the result. The value computed corresponds to the difference of the means of the two groups normalized by the standard deviation divided by the square root of the size of the sample. This value has to be compared to the corresponding theoretic value

known. If the computed value is lower than the theoretic value, it means that our computed value corresponds to a value following a normal law, therefore there is no difference between the two groups as the difference computed is statistically probable considering a normal law.

Appendix C

CERNI document with inclusion questionnaire

FORMULAIRE DE SOUMISSION AU CERNI

Le CERNI examine les protocoles de recherche réalisés sous la responsabilité d'un chercheur ou d'un enseignant-chercheur titulaire rattaché à la communauté d'établissements de l'Université Fédérale de Toulouse

RESUME DU PROJET *(en une page)*

De nos jours, de nombreuses tâches demandent un niveau important de coordination entre l'humain et de la machine (pilotage de drone, surveillance système, pilotage d'avion, etc.). Ces systèmes complexes humain-machine font parfois face à des situations où aucune solution optimale ne peut être trouvée et dans lesquelles des considérations éthiques sont au centre du processus de prise de décision. Du fait de la complexité et du coût du raisonnement dans ces situations de dilemme, bien que la machine de par ses capacités d'analyse et sa rapidité de calcul pourrait être un atout voir une aide à la prise de décision, cette dernière est pour l'instant laissée entièrement à l'humain. En effet, les outils permettant l'analyse « automatique » de concepts éthiques par la machine n'en sont qu'à leur début. Notre projet a pour but d'« Etudier l'influence sur un opérateur humain de décisions argumentées fournies par un agent artificiel lors d'une situation de dilemme éthique. ». Nous souhaitons aussi déterminer la manière dont l'opérateur perçoit et considère l'agent artificiel. Dans, cette expérience, le participant sera confronté à plusieurs dilemmes éthiques avec, pour chaque dilemme, plusieurs choix possibles. Pour chacun des dilemmes, les agents artificiels proposeront chacun un choix accompagné d'un argument justifiant ce choix. La tâche du participant consistera à valider ou à s'opposer à la décision de l'agent. Les arguments de l'agent se baseront sur la formalisation de deux cadres éthiques, le cadre conséquentialiste et le cadre déontologique (WOODFIELD, 1976)). Nous différencierons donc deux agents (un pour chaque cadre éthique), ainsi que deux types de dilemmes éthiques (personnel/impersonnel (J. D. GREENE, 2004)). Nous nous basons sur l'idée que le participant sera plus ou moins influencé selon le type de dilemme auquel il est confronté et selon le type d'agent qui l'assiste.

On étudiera trois types d'influences :

- Influence décisionnaire : les données fournies par l'agent artificiel influent sur la prise de décision du participant.
- Influence neurophysiologique : l'agent artificiel et les données qu'il fournit influent sur l'activité neurologique des zones du cerveau utilisées par le participant (partie plutôt émotionnelle et partie plutôt utilitariste).
- Influence temporelle : le fait que l'agent artificiel donne des informations demande un temps de raisonnement plus long au participant pour prendre une décision.

Cette expérimentation est l'aboutissement de deux pré-expérimentations qui ont pour objectif d'élaborer et de valider le matériel de cette expérience. Ces deux pré-expérimentations seront réalisées en ligne afin d'obtenir un maximum de réponses possible. La première pré-expérimentation a pour but de sélectionner des arguments pertinents pour chaque type d'agents et pour chaque dilemme. La deuxième pré-expérimentation a pour objectif de différencier les dilemmes personnels des dilemmes impersonnels.

Titre du projet :

Étude de l'influence sur un opérateur humain des informations fournies par un agent artificiel lors d'une situation de dilemme éthique.

Domaine scientifique :

Neurosciences, éthique, intelligence artificielle

Chercheur titulaire (1 seul) responsable scientifique du projet :

Frédéric Dehais, Professeur à l'ISAE-SUPAERO, Département de Conception et de Conduite de Véhicules Aéronautiques et Spatiaux (DCAS), Université de Toulouse.
E-mail : frederic.dehais@isae.fr
Tél : 05 61 33 83 72

Autres chercheurs participant au projet :

Vincent Bonnemains, Doctorant (Intelligence artificielle et éthique), ONERA, DCSD
E-mail : vincent.bonnemains@onera.fr
Tél : 05 62 25 22 02

Catherine Tessier, Maître de Recherches 2, Ingénieur expert, ONERA, DCSD
E-mail : catherine.tessier@onera.fr
Tél : 05 62 25 29 14

Eve Fabre, Post-doctorante, ISAE-SUPAERO, DCAS
E-mail : eve.fabre@isae-supaero.fr

Lieu(x) de recherche (endroit(s)) où l'étude va être conduite :

ISAE-SUPAERO, Département de Conception et de Conduite de Véhicule Aéronautique et Spatiaux (DCAS), ISAE, 10 avenue Edouard Belin, 31500 Toulouse.

Objectif principal (5 lignes max.) :

Analyser l'influence décisionnelle, temporelle et cognitive que peut avoir un agent artificiel et les informations qu'il fournit sur la prise de décision d'un humain placé en situation de dilemme éthique.

Je prends connaissance du fait que l'avis rendu par le CERNI ne concerne que le projet de recherche présenté dans ce document.

Date :

Signature numérisée du responsable scientifique:

1. DESCRIPTION SOMMAIRE DU PROJET

Contexte et intérêt scientifiques

Le développement de l'autonomie des robots nous invite à considérer l'éthique comme un élément du raisonnement automatisé d'un agent artificiel. De nombreux articles proposent ainsi des méthodes permettant de doter un système autonome d'un raisonnement pouvant être considéré comme éthique par l'homme (Yilmaz et al., 2016 ; Berreby et al., 2015 ; Cointe et al., 2016 ; Bonnemains et al., 2016). Cependant, bien que la prise de décision morale chez l'humain ainsi que les corrélats neuronaux associés soient étudiés depuis des dizaines d'années (Casebeer, 2003 ; Bartels et al., 2011), le comportement de l'humain face à ces « agents éthiques » est encore méconnu. Des travaux tendent à montrer que l'appréhension du robot par l'homme peut être faussée, notamment par l'apparence du robot (Malle et Scheutz, 2016). Quelle est l'importance de cette influence lorsque l'agent artificiel n'a pas de corps physique et a pour unique but d'assister l'humain lors d'une –prise de décision impliquant des considérations éthiques ?

Les dilemmes moraux

L'étude du raisonnement moral chez l'humain consiste dans la majorité des cas à confronter le participant à des dilemmes moraux. Un dilemme moral est une situation où aucune alternative possible n'est éthiquement satisfaisante (Aroskar, 1980). Un des dilemmes moraux le plus étudié (Foot, 1967) porte le nom de dilemme du trolley : *« Un trolley lancé à pleine vitesse et n'ayant plus la capacité de s'arrêter fonce sur cinq personnes attachées à la voie. Le participant a à sa disposition un levier lui permettant de dévier le trolley et de sauver les cinq personnes mais tuera une personne se trouvant sur la voie de déviation »*. Ce dilemme est généralement présenté en comparaison d'un autre dilemme le footbridge dilemma (Foot, 1983 : Thompson, 1985). Dans cette version, un trolley menace toujours la vie de cinq personnes, mais cette fois le participant se trouve sur une passerelle surplombant une portion de rails entre le trolley et les cinq personnes. Sur cette passerelle se trouve « fatman », une personne suffisamment corpulente pour stopper la course du train. Le participant a donc le choix entre ne rien faire et pousser « fatman » pour arrêter le train et sauver les cinq personnes, mais cela aura pour conséquence la mort de « fatman ». Il a été observé qu'en moyenne, dans le premier dilemme, la majorité des participants choisissent de détourner le trolley pour sauver le plus grand nombre de vies humaines. En revanche, dans le cas du footbridge dilemma, seule une minorité des participants choisissent de sacrifier « fatman ».
Afin d'identifier la raison pouvant expliquer une telle différence de comportement des individus entre ces deux dilemmes, Greene (Greene et al., 2001) a définie deux catégories de dilemmes moraux. Ainsi, le dilemme du trolley classique est considéré comme impersonnel car l'implication physique et émotionnelle de l'acteur est moindre. Tandis que le dilemme du footbridge est de type personnel, car l'acteur a une implication physique directe (lorsqu'il pousse « fatman »). Cette étude a démontré que face à des dilemmes de type personnel, les participants montraient une augmentation d'activité des zones cérébrales sous-tendant le traitement des émotions, alors que face à des dilemmes de type impersonnel, les participants montraient une augmentation d'activité des zones cérébrales associées à la mémoire et au calcul.

Cadres éthiques et agents

Dans cette étude nous souhaitons déterminer et mesurer la façon dont l'humain considère un agent artificiel. En effet, lorsqu'il s'agit de décision éthique et de robots, il semble que le jugement moral humain diffère (Malle et al., 2015) selon la nature de l'acteur. Dans le cadre de dilemmes éthiques, Malle a ainsi démontré que les individus ne jugeaient pas un choix de la même manière quand il a était fait par un humain versus un robot.

Nous distinguons ici deux cadres éthiques:

1. Conséquentialiste (Canto-Sperber, 1996) : Ce cadre est issu du téléologisme (du grecque telos « la fin » et logos « la raison »). Ce cadre éthique raisonne donc sur les conséquences des décisions plutôt que sur les moyens de les obtenir. Ainsi, un acteur privilégiera d'un point de vue conséquentialiste l'obtention des meilleures conséquences, sans se soucier des moyens mis en œuvre.
2. Déontologique (Canto-Sperber, 1996) : Du grecque deon « le devoir » et logos « la raison », ce cadre s'applique à suivre des principes, devoirs, qui souvent s'opposent au conséquentialisme en se focalisant sur l'action et la décision plutôt que sur les conséquences.

À titre d'illustration, dans le dilemme du footbridge, une personne ayant un profil plus conséquentialiste choisira de pousser « fatman » pour sauver le plus de vies. À l'inverse, une personne ayant un profil plus déontologique préférera ne rien faire car pousser « fatman » est un acte criminel inacceptable.

Aucune expérimentation ne s'est pour le moment intéressée à l'influence d'un agent artificiel sur la prise de décision de l'humain faisant face à un dilemme éthique. Cette étude est nécessaire car la façon dont l'humain considère la machine dans ce genre de situation est inconnue.
Nous savons que des informations fournies à un opérateur peuvent avoir un impact significatif sur sa prise de décision (Bommer et al., 1987). En nous basant sur les travaux de Greene, nous pouvons étudier les processus de prise de décision en mesurant les variations de l'activité cérébrale en fonction des conditions et en les liants aux choix qui ont été fait.
Dans cette expérimentation, nous allons confronter le participant à plusieurs dilemmes éthiques. Dans un premier temps, nous souhaitons classer les participants en deux groupes : un groupe ayant un profil conséquentialiste et un groupe ayant un profil plus déontologique. Pour cela, les participants devront prendre une décision pour différents dilemmes moraux personnels et impersonnels sans que l'agent ne soit présent (condition contrôle).
Dans un second temps, les mêmes dilemmes seront présentés au participant, mais cette fois-ci, avec l'agent artificiel. Ce dernier (de type conséquentialiste ou déontologique) aura pré-sélectionné une alternative et proposera un argument court (une à deux phrases) justifiant ce choix. Le participant devra valider ou non cette décision. La nature des dilemmes moraux sera manipulée. Ils seront de type personnel ou de type impersonnel (voir description ci-dessous) et l'agent de type conséquentialiste ou déontologique (i. e. qui choisira une décision conséquentialiste/déontologique soutenu par un argument de même type[1]). Nous allons étudier l'impact de l'agent artificiel sur les processus de prise de décision humains. Pour ce faire, nous utiliserons un appareil fNIRs dans le but d'identifier les structures cérébrales sous-tendant 1) la prise de décision morale et 2) l'influence de l'agent dans l'évaluation des alternatives d'un dilemme moral. Un questionnaire post-expérimental nous permettra de mesurer le degré d'appréhension de l'agent artificiel de chaque participant.

Contexte et intérêt scientifique

Cette étude vise à évaluer l'influence que peut avoir un agent artificiel sur la prise de décision de l'humain face à un dilemme éthique. On va pour ce faire identifier si un participant se rapproche plus du cadre éthique déontologique ou conséquentialiste grâce à la première étape de l'expérimentation, puis on pourra analyser

Objectifs

Nous cherchons à analyser l'influence de l'agent artificiel et des informations qu'il fournit sur la cognition et la prise de décision de l'humain, dans un contexte de dilemme éthique.

Objectif 1. *Déterminer les préférences éthiques des participants.* Lors de la première partie de l'expérimentation, il sera demandé aux participants de prendre une décision pour chaque dilemme sans influence de l'agent. Ces réponses permettront d'identifier si un participant se rapproche plus du cadre déontologique ou conséquentialiste. Grâce aux mesures neurophysiologiques, nous pourrons aussi confirmer les résultats neurologiques de Greene.

Objectif 2. *Identifier les participants qui sont sensibles aux arguments des agents et ceux qui ne le sont pas.* Une fois qu'un participant sera affilié au groupe conséquentialiste/déontologique, si il change de décision lorsqu'on lui propose un argument de type opposé, il sera considéré comme sensible.

Objectif 3. *Déterminer les corrélats neuronaux associés à la sensibilité et à la non-sensibilité aux arguments des agents et les comparer entre eux.* Lorsque le participant prendra une décision différente de sa décision initiale, on pourra observer les corrélats neuronaux associés et les comparer à ceux observés lorsqu'un participant est insensible.

Objectif 4. *Déterminer les causes de non-sensibilité aux arguments des agents.* Lorsque le participant sera catégorisé comme insensible, le questionnaire post-expérimental permettre de connaître la raison de cette insensibilité (aversion envers les machines, arguments non-convainquant, etc.).

[1] Les arguments présentés dans les deux cadres éthiques seront issus de travaux précédents (Bonnemains et al., 2016) sur la modélisation de l'éthique pour les systèmes autonomes.

Hypothèses

On suppose que la présence de l'agent, ainsi que des informations qu'il fournit, auront une influence sur l'activité cérébrale sous-tendant le traitement des émotions et le calcul mental.

On prédit que les participants auront un temps de réponse plus élevé lors de la deuxième partie de l'expérimentation car ils devront étudier l'argument de l'agent artificiel. De plus, on prédit qu'une partie des participants sera sensible aux arguments de l'agent et changera de décision entre la première et la deuxième partie de l'expérimentation pour un même dilemme.

On prédit que les participants catégorisés comme conséquentialistes montreront une activité des structures cérébrales associées à la mémoire et au calcul plus importante que celle des structures cérébrales associées au traitement des émotions et inversement pour les participants catégorisés comme déontologiques.

Si une personne est identifiée comme sensible aux arguments de l'agent, on prédit que le changement dans le choix de la décision sera lié à une variation dans l'activité des structures cérébrales associées à l'émotion et au calcul mental. Ainsi, un changement vers un choix conséquentialiste sera associé à une augmentation de l'activité des structures cérébrales sous-tendant le calcul mental et une diminution de l'activité des structures cérébrales sous-tendant le traitement de émotions et inversement pour un changement vers un choix déontologique.

Conflits d'intérêts

Aucun conflit d'intérêts à signaler.

2. MATERIEL ET METHODE

A. Participants

Nombre exact de participants ou « fourchette » approximative et critères utilisés pour fixer ce nombre :

Nous souhaitons tester 60 participants. Le groupe devra être constitué de personnes d'âge comparable avec un même nombre d'hommes et de femmes.

Neutralité des participants :

Nous recruterons les participants à l'ISAE-SUPAERO ou l'ONERA pour des raisons pratiques (proximité du lieu d'expérimentation, assurance, etc.). Les participants ne devront pas avoir réalisé d'expérimentation sur l'éthique afin d'éviter des biais d'apprentissage.

B. Recrutement

Mode de recrutement : Les participants, étudiants ou personnel de l'ISAE-SUPAÉRO et de l'ONERA sont recrutés par messagerie électronique et via annonces papiers dans les établissements cités précédemment.

Lieu de recrutement : ISAE-SUPAÉRO et ONERA

Critères de sélection : Les sujets devront être âgés de 18 à 35 ans, avoir une vision normale ou corrigée à la normale, être affilié à la sécurité sociale, avoir signé un consentement éclairé et ne jamais avoir participé à des expérimentations impliquant des considérations éthiques.

Critères de non inclusion : Personnes souffrant d'affections neuropsychologiques, trouble important de la vision, prise de médicaments ou substances psychotropes.

Indemnisation éventuelle des sujets : Les sujets ne seront pas indemnisés pour leur participation.

C. Matériel

Afin d'obtenir un matériel expérimental viable, nous effectuerons deux pré-tests permettant d'éprouver la validité de nos dilemmes et arguments. Ces pré-tests seront effectués en ligne afin de toucher le plus grand nombre de personnes possible.

Pré-test 1 : Homogénéisation de la pertinence des arguments

Problématique : Obtenir des arguments de pertinence comparable.
Pour un même dilemme, il est important que les arguments conséquentialiste et déonthologique soient aussi convaincant l'un que l'autre. C'est pourquoi cette pré-expérimentation demandera aux participants d'évaluer la pertinence pour chaque dilemme d'ensembles d'arguments conséquentialistes et déontologiques. Cette évaluation sera réalisée sur une échelle de Likert (1 à 7) :

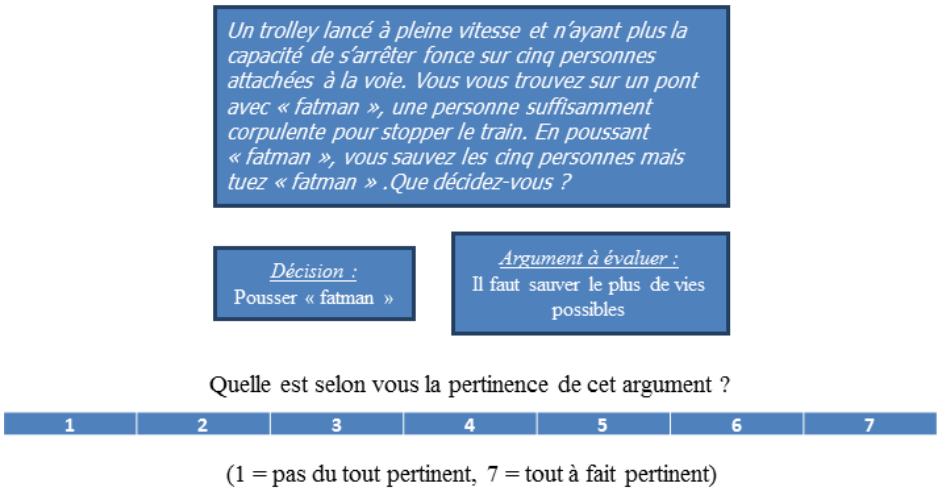


Figure 2. Illustration d'une demande de notation pour un argument conséquentialiste.

Cette évaluation sera réalisée en deux temps : évaluation des arguments d'une des deux catégories tirée au hasard, puis, après évaluation d'arguments d'autres dilemmes, évaluation des arguments de la seconde catégorie. Exemple :

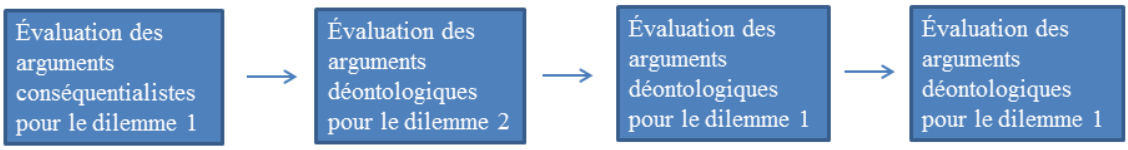


Figure 3. Illustration du déroulement de l'évaluation d'arguments pour deux dilemmes.

A la fin de l'expérimentation, il sera possible d'extraire une note de pertinence pour chaque argument. L'objectif est de sélectionner pour chaque dilemme un argument conséquentialiste et un argument déontologique ayant des notes de pertinence les plus proches possibles. Ce sont ces arguments qui seront retenus pour être présentés lors de l'expérimentation finale, afin que l'expérimentation ne soit pas biaisée par un argument plus convainquant que l'autre.

Pré-test 2 : Catégorisation personnel/impersonnel des dilemmes

Problématique : Comment sans mesure neurophysiologique déterminer si un dilemme est de type personnel ou impersonnel ?
Détail : Dans son article, Greene ne détaille pas réellement comment il catégorise ses dilemme comme personnel/impersonnel. Nous pensons que cette catégorisation est liée au degré d'implication ressenti par le participant. C'est pourquoi nous allons demander aux participants d'évaluer à quel point ils se sentent responsables (sur une échelle de Likert de 0 à 7) face à un dilemme.
Cette évaluation se fera pour chaque dilemme en quatre étapes :

- Avant toute prise de décision, le participant devra évaluer à quel point le dilemme lui semble intense.

- Puis, le participant devra choisir une des décisions du dilemme.
- Il devra ensuite évaluer quelle est sa responsabilité s'il prend cette décision.
- Enfin, il devra évaluer quelle est sa responsabilité s'il prend l'autre décision.

La première évaluation sera un critère déterminant. Si un dilemme est de forte intensité pour une majorité des participants, alors il sera désigné comme personnel. A l'opposé, si il est noté de faible intensité, il sera désigné impersonnel.

La seconde évaluation sera utilisée pour évaluer une hypothèse selon laquelle un dilemme est personnel s'il existe une différence importante de responsabilité entre les deux décisions d'un dilemme.

À la fin de ce pré-test, nous sélectionnerons comme personnels les cinq dilemmes ayant reçues les notes d'intensité les plus élevées, et comme impersonnels les cinq dilemmes ayant reçues les notes d'intensité les moins élevées.

Expérimentation

10 dilemmes éthiques, 5 dilemmes personnels et 5 dilemmes impersonnels. Pour chaque dilemme, deux décisions, une déontologique et une conséquentialiste. Chaque décision est rattachée à un argument de même type qu'elle. Exemple :

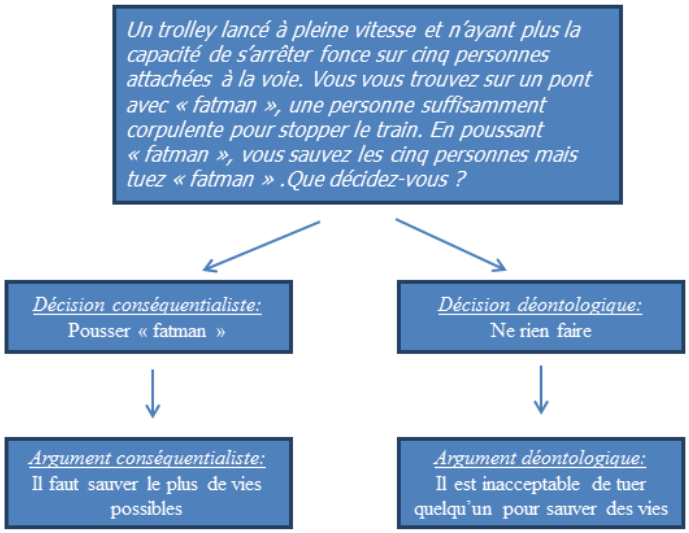


Figure 1. Illustration des composants d'un dilemme pour l'expérimentation.

Matériel de mesures neurophysiologiques

Un système de spectroscopie proche infrarouge fonctionnelle sera placé (fNIRS) sur le front du participant (voir Figure 2et Annexe 2). Si le front des participants est trop petit certaines électrodes seront sacrifiées.

Figure 5. Illustration de spectroscopie proche infrarouge fonctionnelle de Bionic ©.

Le participant sera installé dans une salle à faible luminosité, assis sur une chaise devant une table sur laquelle se trouveront deux écrans et un clavier. Sur l'écran de gauche sera affichée la fenêtre du logiciel détaillant le dilemme éthique et les décisions correspondantes. Sur l'écran de droite sera affiché l'argument soutenant la décision de l'agent artificiel. Le participant interagira avec le logiciel via le clavier.

D. Méthode

Déroulement de l'expérience pour un participant

Etape1 : Le participant est accueilli dans la salle d'expérimentation. Après avoir lu et signé la notice d'information et de consentement (Annexe 1) et le questionnaire d'inclusion (Annexe 2), il sera demandé au participant de lire les consignes de l'expérience. Le participant pourra poser toutes les questions nécessaires.

Etape 2 : Le participant s'installe ensuite confortablement face à l'ordinateur sur lequel il réalisera l'expérimentation. Enfin, l'expérimentateur procède à la mise en place de la fNIRs.

Etape 3 : Le participant réalise la tâche (décrite en suivant) pour laquelle il s'est porté volontaire.

Etape 4 : L'appareil fNIRs est retiré puis l'on demande au participant de remplir un questionnaire mesurant l'appréhension associée aux robots (Schaefer, 2013), pour savoir ce qu'il a pensé de l'agent artificiel lors de l'expérimentation. Ce questionnaire a pour but de déterminer si un participant n'ayant pas accepté les décisions de l'agent l'a fait par simple opposition face aux arguments de l'agent ou par opposition aux systèmes artificiels en règle générale.

Etape 5 : Un débriefing avec le participant est réalisé par l'expérimentateur afin de détecter tout inconfort.

Description de la tâche

Dans cette expérience, le participant se met à la place d'un opérateur assisté par un agent artificiel pour prendre une décision face à un dilemme éthique.
Dans une première partie, le participant se retrouve face à plusieurs dilemmes et il doit uniquement prendre une décision parmi celles proposées sans assistance de l'agent artificiel (condition de contrôle).
Dans une seconde partie, ces dilemmes lui sont présentés à nouveau (dans un ordre différent), mais cette fois l'agent artificiel lui présente une décision sélectionnée ainsi qu'un argument soutenant cette décision. Le participant doit

cette fois choisir de valider la sélection de l'agent artificiel ou de s'opposer à cette sélection (ce qui revient à prendre la décision opposée).

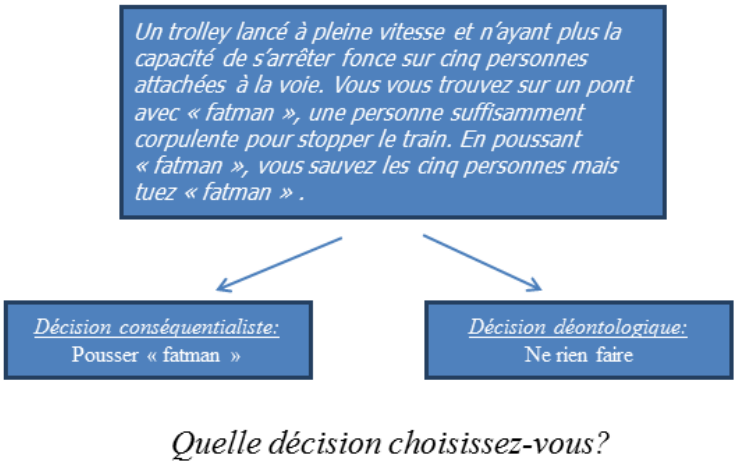


Figure 6. Illustration d'une demande de sélection d'une décision sans agent artificiel

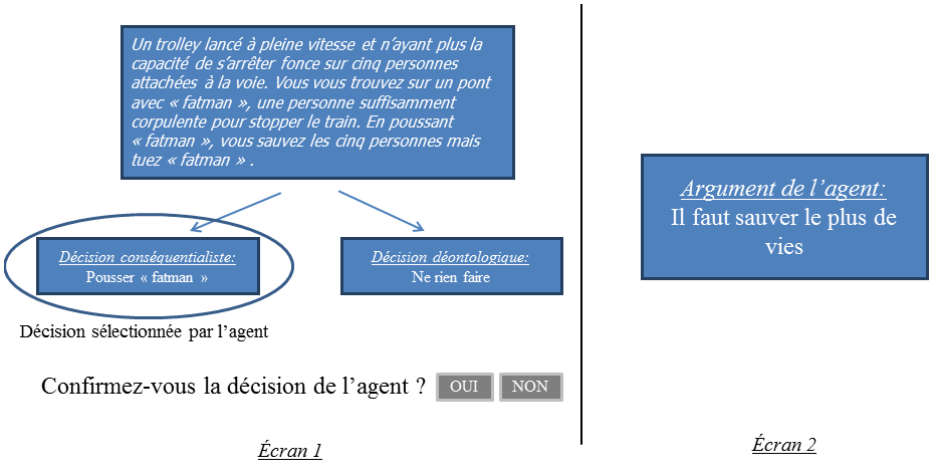


Figure 7. Illustration d'une demande de validation de décision prise par l'agent artificiel avec argument conséquentialiste.

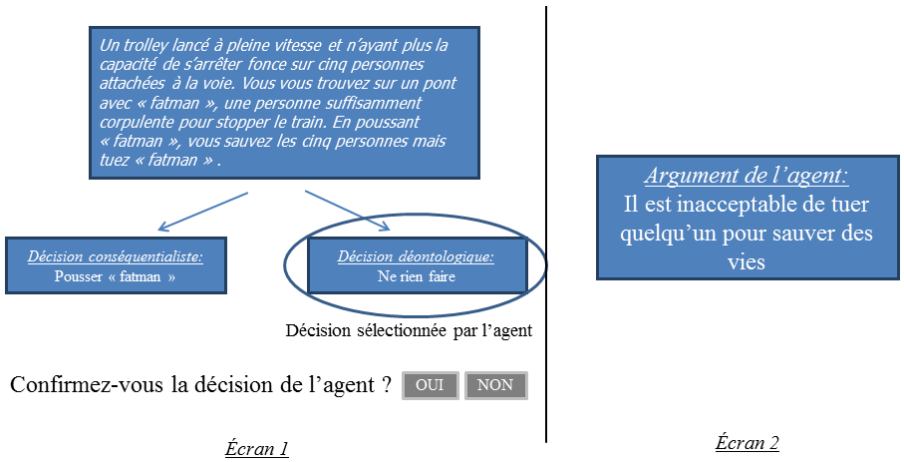


Figure 8. Illustration d'une demande de validation de décision prise par l'agent artificiel avec argument déontologique.

Le participant devra répondre à une série de 10 dilemmes, avec pour chaque dilemme la prise de décision sans agent, la validation ou non pour l'agent (i. e. argument) conséquentialiste et pour déontologique, soit un total de 30 essais. Après chaque réponse, un écran noir de 10 secondes sera affiché pour permettre au participant de revenir à son activité

neurophysiologique normale.

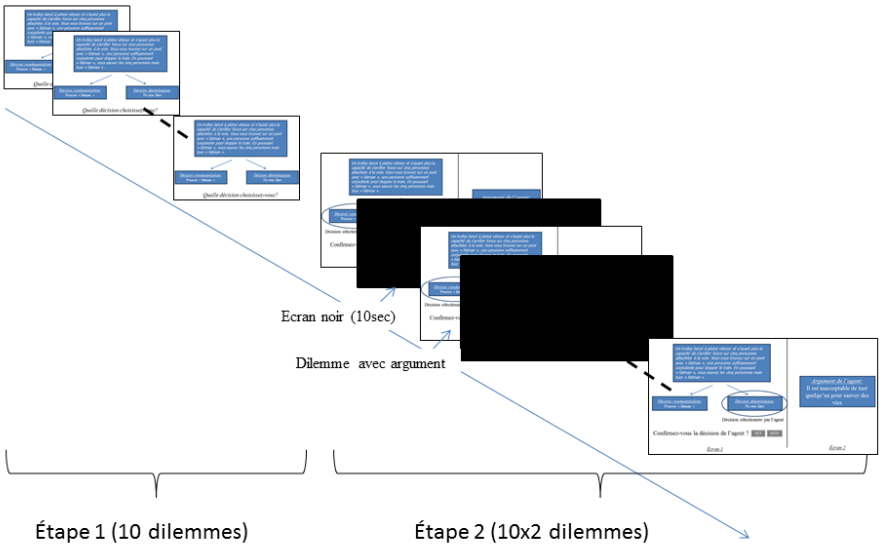


Figure 8. Illustration d'un essai

Analyse des données :

Résultats comportementaux

Les choix réalisés par le participant lors de la première partie vont permettre d'affilier à chaque participant une affinité conséquentialiste (rationnelle) ou déontologique (émotionnelle).

Lors de la deuxième partie, le taux de déviation par rapport au choix initiale de la condition contrôle (première partie) sera analysé afin d'identifier les participants sensibles aux arguments des agents.

Le temps de réponse traduira un processus décisionnel plus ou moins complexe. Plus le temps de réponse sera élevé, plus le processus de décision sera en théorie complexe.

Corrélats neuronaux

Les résultats obtenus à la première partie nous permettrons de confirmer les résultats de Greene.

Les mesures de la deuxième partie vont permettre de déterminer les corrélats neuronaux associés à la déviation de prise de décision (par rapport la condition contrôle). Ces résultats pourront être comparés aux corrélats neuronaux associés à la stabilité de la prise de décision (choix qui ne change pas par rapport à la condition contrôle).

Calendrier des évaluations ou observations :

Début des expérimentations : Novembre 2017
Fin des expérimentations : Novembre 2018
Durée prévue : 1 heure

E. Bénéfices et risques prévisibles et connus pour la santé physique et mentale (estime de soi, etc.) et la vie sociale (réputation)

Cette étude vise à caractériser l'influence d'un agent artificiel sur l'humain en situation de dilemme éthique. C'est

une première étape importante dans l'intégration de concepts éthique dans la conception des systèmes autonomes, c'est aussi une étude pertinente dans le cadre de l'amélioration des systèmes homme-robot. Le seul risque pour le participant est l'émergence d'émotions vives en rapport avec ce que sollicitent les dilemmes éthiques face au vécu personnel du participant. Cependant, la grande quantité d'études déjà réalisées avec des dilemmes éthiques démontrent que les participants n'ont aucune séquelle après ce genre d'expérimentation.

1- Présentez les bénéfices de votre étude. Il peut s'agir de bénéfices en termes d'avancées scientifiques, d'amélioration de la qualité de vie des participants, de leur estime de soi, etc.

OUI*	Votre protocole utilise-t-il une mise en scène expérimentale destinée à dissimuler une partie de l'objectif ou de la méthodologie aux sujets ou de faire croire à d'autres objectifs ou d'autres méthodologies**?** *Si oui, ce dossier doit présenter une description de la mise en scène utilisée et une explication de la façon de la dévoiler aux sujets à la fin de l'étude et de leur préciser les véritables objectifs de l'étude. En outre, on doit amener des arguments montrant que la dissimulation de certains aspects du protocole est indispensable au regard des objectifs et des enjeux, et qu'aucun des aspects dissimulés aux sujets n'est susceptible de menacer leur sécurité ou leur dignité.*
OUI	Questions ou situations pouvant mettre les participants mal à l'aise ?
NON	Matériaux considérés par le participant comme menaçants, choquants, répugnants ?
NON	Possibilité d'atteinte à la vie privée du participant, de sa famille, incluant l'utilisation d'informations personnelles ?
NON	Utilisation de stimuli physiques (auditifs, visuels, haptiques, etc.) autre que des stimuli associés à des activités normales ?
NON	Privation de besoins physiologiques (boire, manger, dormir, etc.)
NON	Manipulation de paramètres psychologiques ou sociaux comme la privation sensorielle, l'isolement social ou le stress psychologique ?
NON	Efforts physiques au-delà du niveau considéré comme modéré pour le participant ?
NON	Exposition à des drogues, produits chimiques ou agents potentiellement toxiques ?

*Il ne sera pas explicité que le participant sera face à des dilemmes éthiques. Cependant, il sera détaillée dans le formulaire de consentement éclairé que le participant sera confronté à des situations pouvant susciter une réaction émotionnelle vive selon son vécu.

F. Vigilance/ Arrêt prématuré de l'étude

La participation à cette étude se base sur le volontariat. Aussi, tout participant pourra quitter l'étude à n'importe quel moment sur simple demande sans qu'il lui en soit tenu rigueur.

3. TRAITEMENT DES DONNEES – RESPECT DE LA VIE PRIVEE DU PARTICIPANT

A. Confidentialité

Procédé d'anonymisation : La confidentialité sera garantie par le fait que chaque sujet sera désigné par un identifiant prenant la forme d'un numéro aléatoire dans les analyses et documents numériques ou papier. L'identité des personnes ne sera pas relevée et il n'y aura donc aucune correspondance possible avec des coordonnées réelles des volontaires.

Personnes ayant accès aux données :

- Frédéric Dehais, professeur à l'ISAE-SUPAERO
- Vincent Bonnemains, doctorant à l'ONERA
- Catherine Tessier, chercheur à l'ONERA
- Eve Fabre, post-doctorante à l'ISAE-SUPAÉRO
- Florian Ginisty, étudiant à l'ISAE-SUPAERO

B. Archivage

Type de données archivées (préciser si données identifiantes, directement ou par recoupement) : Durée de l'archivage :

Type de données archivées : Les données sont archivées selon le code d'anonymat et ne sont pas identifiables par recoupement, aucune table de correspondance n'étant créée.
Durée de l'archivage : Les données numériques et papier seront archivées pendant une durée maximale de 10 ans après la fin de la recherche.

Lieu de l'archivage :

Lieu de l'archivage : Les disques durs et données papier seront stockés dans une armoire sécurisée dans la salle d'expérimentation. Une copie numérique et cryptée sera stockée sur le serveur de l'ISAE-SUPAERO et le serveur de l'ONERA

Possibilité de destruction à la demande du participant *(voir cas de figure section 4)* :

Possibilité de destruction à la demande du participant : Sur simple demande du participant et sans justification, l'ensemble des données le concernant seront détruites si la demande est réalisée au cours de l'expérimentation. Ensuite, les données ne sont pas identifiables.

4. FORMULAIRE DE CONSENTEMENT ECLAIRE INCLUANT L'INFORMATION A DONNER AUX PARTICIPANTS

ANNEXE 1 – NOTICE D'INFORMATION ET CONSENTEMENT ECLAIRE

Titre du projet :

Étude de l'influence sur un opérateur humain des informations fournies par un agent artificiel lors d'une situation de dilemme éthique.

Chercheur titulaire responsable scientifique du projet :

Fréderic Dehais, professeur à l'ISAE-SUPAERO
Département de Conception et de Conduite de Véhicules Aéronautiques et Spatiaux (DCAS) Université de Toulouse.
E-mail : frederic.dehais@isae-supaero.fr
Tel. : 05 61 33 83 72

Lieu de recherche :

Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO)

But du projet de recherche :

Analyser l'influence cognitive et décisionnelle des informations fournies par un agent artificiel

Ce que l'on attend de vous (méthodologie)

Si vous acceptez de participer à cette étude, vous serez amené à étudier des situations et à choisir une décision possible face à cette situation. Votre activité cérébrale sera mesurée par des appareils fNIRS et vos réponses seront enregistrées. Ces enregistrements n'ont aucune valeur diagnostique, et nous ne pourrons vous délivrer aucune information de nature médicale. L'expérience se décompose en plusieurs phases vous permettant de prendre en main les consignes et la tâche à accomplir. Les données ne seront enregistrées que pendant la phase expérimentale suivant la phase d'entraînement. L'expérience durera en tout environ 1 heure.

Vos droits de vous retirer de la recherche en tout temps

En acceptant de participer à cette expérience, vous comprenez que :
1. Votre participation est tout à fait volontaire, et ne sera pas rétribuée
2. Vous pouvez vous retirer ou demander à cesser l'expérimentation, à tout moment, pour quelque motif que ce soit.
3. Un arrêt éventuel de votre participation n'aura aucun impact sur vos notes, votre statut ou vos relations futures avec l'équipe du Département de Conception et de Conduite de Véhicules
Aéronautiques et Spatiaux (DCAS) de l'ISAE-SUPAERO.

Vos droits à la confidentialité et au respect de la vie privée

En acceptant de participer à cette recherche, vous comprenez que :
1. Les données qui seront obtenues seront traitées avec la plus entière confidentialité.
2. Votre identité sera masquée, à l'aide d'un numéro aléatoire.
3. De ce fait, aucun lien ne pourra être établi entre votre identité (qui ne sera pas demandée) et les données recueillies.
4. Les données seront conservées dans un endroit sécurisé, et seul le responsable scientifique, ainsi que les chercheurs adjoints y auront accès.
5. Compte tenu de l'anonymisation totale de vos données, leur destruction ou rectification ne sera pas possible une fois la campagne d'expérimentation terminée.

Bénéfices

Sur le plan scientifique, cette étude a pour ambition d'analyser l'impact sur la cognition et la décision d'informations fournies par un agent artificiel. Cette expérimentation ne présente aucun danger connu.

Risques possibles

À notre connaissance, cette recherche n'implique aucun risque physique. Cependant, les situations décrites peuvent être source d'inconfort ou d'émotion vive. Vous êtes libres d'arrêter l'expérimentation à tout moment en cas d'inconfort. L'expérimentateur est à votre disposition si une gêne psychologique persiste. L'enregistrement des données cérébrales ne présente aucun danger, il s'agit d'une technique non-invasive, c'est-à-dire une technique qui n'a aucun impact sur le fonctionnement normal du corps et plus particulièrement ici du cerveau.

Diffusion

Cette recherche pourra faire l'objet de publications scientifiques, dans un colloque ou dans une revue scientifique spécialisée. Les données personnelles restent strictement confidentielles et ne peuvent être reliées à une identité.

Vos droits de poser des questions en tout temps

Vous pouvez poser des questions au sujet de la recherche en tout temps en communiquant avec le responsable scientifique du projet par courrier électronique à vincent.bonnemains@onera.fr .

Consentement à la participation

En signant le formulaire de consentement, vous certifiez que vous avez lu et compris les renseignements ci-dessus, qu'on a répondu à vos questions de façon satisfaisante et qu'on vous a avisé que vous étiez libre d'annuler votre consentement ou de vous retirer de cette recherche à tout moment, sans préjudice.

A remplir par le participant :

J'ai lu et compris les renseignements ci-dessus et j'accepte de plein gré de participer à cette recherche.
Nom, Prénom – Date – Signature

Un exemplaire de ce document vous est remis, un autre exemplaire est conservé par le responsable scientifique de la recherche.

ANNEXE 2 – QUESTIONNAIRE D'INCLUSION

Moi ……., certifie sur l'honneur que :

- Mon âge est compris entre 18 et 35 ans
- Ma vision a été évaluée par un spécialiste au cours des deux dernières années et a été considérée comme « normale »
- Je suis affilié à la sécurité sociale
- Je ne souffre pas d'affection neuropsychologique
- Je n'ai jamais participé à des expérimentations impliquant des considérations éthiques
- Je ne prends aucun médicament ni substance psychotrope.

A remplir par le participant :

J'ai lu et compris les renseignements ci-dessus et les certifie exactes.
Nom, Prénom – Date – Signature

Un exemplaire de ce document vous est remis, un autre exemplaire est conservé par le responsable scientifique de la recherche.

Références bibliographiques

MA Aroskar (1980) Anatomy of an ethical dilemma: The theory. AJN The American Journal of Nursing 80(4):658–660

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

A. Berten (1996), *Déontologisme*, p. 477-483, M. Canto-Sperber (dir.), Dictionnaire d'Éthique et de Philosophie Morale, Paris, PUF

M. Bommer, C. Gratto, J. Gravander & M. Tuttle. (1987). A behavioral model of ethical and unethical decision making. *Journal of business ethics*, *6*(4), 265-280.

V. Bonnemains, C. Saurel, and C. Tessier. Embedded ethics. Some technical and ethical challenges. Accepted to Journal of Ethics and Information Technology, Special Issue on "Ethics in Artificial Intelligence", 2017

D. M. Bartels et D. A. Pizarro, The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, vol. 121, no 1, p. 154-161, 2011

M. Canto-Sperber (1996). Dictionnaire d'éthique et de philosophie morale, Conséquentialisme, p. 313-320

M. Canto-Sperber (1996). Dictionnaire d'éthique et de philosophie morale, Déontologisme, p. 377-383

W. D. Casebeer, Moral cognition and its neural constituents. *Nature Reviews Neuroscience*, vol. 4, no 10, p. 840-847, 2003

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

P. Foot (1967) The Problem of Abortion and the Doctrine of Double Effect. Oxford Review 5:5–15

P. Foot (1983). Moral realism and moral dilemma. *The Journal of Philosophy*,*80*(7), 379-398.

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

B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis et C. Cusimano, 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, 2015

B. F. Malle et M. Scheutz, Inevitable Psychological Mechanisms Triggered by Robot Appearance: Morality Included?. In : *2016 AAAI Spring Symposium Series*. 2016

K. Schaefer (2013). The perception and measurement of human-robot trust.

J. Thomson (1985) The Trolley Problem. *Yale Law Journal*94:1395-1415.

A. Woodfield (1976) Teleology. Cambridge University Press

L. Yilmaz, A. Franco-Watkins, TS Kroecker, 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, DOI 10.1109/COGSIMA.2016.7497784, 2016

Appendix D

Moral dilemmas

D.1 Le tramway

Un tramway fonce sur cinq personnes piégées sur la voie et il est impossible de l'arrêter. Vous avez en face de vous un levier que vous pouvez actionner pour dévier le tramway sur une voie secondaire, où une seule personne est piégée. Sachant que tout individu se trouvant sur le passage du tramway mourra, que feriez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Actionner le levier"

D.2 La passerelle

Un tramway fonce vers cinq personnes piégées sur la voie. Tout individu se trouvant sur le passage du tramway mourra. Vous vous trouvez sur une passerelle située au-dessus de la voie entre le tramway et les cinq personnes en compagnie d'un individu suffisamment imposant pour arrêter la course du tramway. Sachant que votre sacrifice n'aurait aucun effet, la seule possibilité d'arrêter le tram serait de pousser l'individu corpulent sur la voie. Cela tuerait cet individu mais les 5 personnes sur la voie seraient sauvées. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Pousser l'individu"

D.3 La trappe

Un tramway fonce vers cinq personnes piégées sur la voie. Un individu suffisamment imposant pour arrêter la course du tramway se trouve sur une passerelle située au-dessus de la voie entre le tramway et les cinq personnes. Vous avez la possibilité d'actionner une trappe sur laquelle se trouve cet individu. Cet individu mourra alors percuté par le tram, mais les cinq personnes seront sauvées. Sachant que votre sacrifice n'aurait aucun effet (vous n'avez pas la corpulence suffisante pour arrêter le tramway). Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Actionner la trappe"

D.4 Transplantation d'organes

Vous êtes médecin et cinq de vos patients sont sur le point de mourir du fait de la défaillance d'un de leurs organes. La seule façon de les sauver serait de prélever les organes d'un individu sain et de les greffer à vos patients. Cela provoquerait la mort de l'individu sain. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Transplanter les organes"

D.5 Le choix de Sophie

En période de guerre, vous et vos deux enfants vivez en territoire occupé par l'ennemi. Un médecin pratique des expérimentations douloureuses sur l'humain menant inévitablement à la mort. Ce médecin souhaite utiliser l'un de vos enfants comme cobaye. Il vous laisse vingt-quatre heures pour choisir lequel de vos enfants lui livrer. Si vous ne le faites pas, vos deux enfants seront utilisés pour ces expérimentations. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Livrer un de vos enfants"

D.6 Le choix du maire

En période de guerre, vous êtes maire d'un petit village occupé par l'ennemi. Un médecin pratique des expérimentations douloureuses sur l'humain menant inévitablement à la mort. Ce médecin souhaite utiliser un des deux enfants du village comme cobaye. Il vous laisse vingt-quatre heures pour choisir lequel des deux enfants lui livrer. Si vous ne choisissez pas, les deux enfants seront utilisés pour ces expérimentations. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Choisir un des deux enfants"

D.7 Le choix de Corinne

Vous revenez de l'école maternelle avec votre fils quand vous êtes confrontés à deux terroristes armés. Ces derniers menacent de tirer sur votre fils, à moins que vous ne saisissiez le code de la porte d'entrée du bureau dans lequel vous travaillez. Si vous leur ouvrez, les terroristes rentreront dans le bâtiment et tueront vos collègues mais vous et votre fils serez épargnés. Si vous refusez, les terroristes tueront votre fils et prendront la fuite. Que décidez-vous ?

deontological choice: "Donner le code"
consequentialist choice: "Refuser de donner le mot de passe"

D.8 Canot de sauvetage

Vous êtes à bord d'un canot de sauvetage après l'évacuation d'un paquebot de croisière. Le canot de sauvetage contient trop de personnes et menace de couler. La mer est agitée et si rien n'est fait toutes les personnes à bord mourront. Un passager blessé a embarqué parmi vous et ne survivra dans aucun cas. Si vous poussez ce passager hors du canot, vous et les autres rescapés resterez à flot et survivrez. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Pousser l'homme blessé"

D.9 Escalade

Vous partez faire de l'escalade en très haute montagne avec douze amis et êtes divisés en deux cordées de six personnes. Un de vos amis ne faisant pas partie de votre cordée glisse et se casse le bras sur la paroi rocheuse. Il est conscient mais n'a plus la capacité de continuer à monter. Il est un danger pour les cinq autres amis de sa cordée qui risquent à tout moment de dévisser et de faire une chute mortelle. Vous êtes la seule personne assez proche pour pouvoir couper la corde rattachant votre ami blessé à la cordée. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Couper la corde"

D.10 Spéléologie

Vous faites partie d'un groupe d'archéologues explorant une grotte marine à marée basse. Le seul moyen de rentrer et de sortir de cette grotte est de passer par un passage étroit dans la roche. Sur le chemin du retour, alors que votre guide tente de s'extirper de la grotte, un éboulement se produit le prenant au piège. Le guide est toujours vivant mais le niveau de la mer monte très rapidement et vous n'aurez pas le temps de dégager les pierres qui bouchent la sortie et bloquent votre guide. Si rien n'est fait toutes les personnes présentes dans la grotte mourront noyées à l'exception du guide dont la tête est hors de la grotte. Le seul moyen de sortir de la grotte à temps est d'utiliser un bâton de dynamite pour libérer l'entrée de la grotte, ce qui tuerait votre guide. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Utiliser la dynamite"

D.11 Sauvetage pompier spéléologique

Un groupe d'archéologues explore une grotte marine à marée basse. Lorsque leur guide tente de s'extirper de la grotte, un éboulement se produit le prenant au piège. Vous êtes en charge des opérations de sauvetage. Le guide est toujours vivant mais le niveau de la mer monte très rapidement. Si l'équipe que vous dirigez ne fait rien, toutes les personnes présentes dans la grotte mourront à l'exception du guide dont la tête est

hors de la grotte. Le seul moyen de dégager l'entrée de la grotte à temps est d'utiliser un bâton de dynamite qui tuera le guide. Que décidez-vous ?

deontological choice :"Donner l'ordre de dégager le guide"
consequentialist choice: "Donner l'ordre d'utiliser la dynamite"

D.12 Conflit éthnique

Vous êtes juge d'une commune composée de deux ethnies: l'ethnie A et l'ethnie B. Vous n'appartenez à aucune des deux ethnies. Durant la nuit, une bagarre éclate et un membre de l'ethnie B tue un membre de l'ethnie A. Le lendemain, aucun membre de l'ethnie B n'accepte de dénoncer le coupable. L'ethnie A menace de les attaquer sauf si vous condamnez à mort le meurtrier. Le coupable étant inconnu, la seule possibilité est de désigner aléatoirement une personne et de la condamner à mort afin éviter un bain de sang. Que décidez-vous?

deontological choice: "Ne rien faire"
consequentialist choice: "Désigner une personne"

D.13 Torturer un terroriste

Vous avez capturé un terroriste qui a caché une bombe dont l'explosion tuerait des centaines de personnes. Il refuse de vous indiquer l'endroit où il l'a placée. Le seul moyen pour que vous arriviez à le faire parler avant que la bombe n'explose est de le torturer. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Torturer le terroriste"

D.14 Noyade

Vous êtes à la plage avec votre fille et votre nièce. Celles-ci se jettent à l'eau. Au bout de quelques minutes, vous entendez des cris et comprenez que les deux enfants accrochées à une bouée sont piégées par le courant. Vous êtes la seule personne sur la plage. Vous décidez d'aller les chercher. À votre arrivée, vous comprenez que vous ne pourrez porter qu'une des fillettes à la fois. Les deux enfants sont épuisées. Votre

nièce est à bout de force et ne tiendra plus longtemps. Votre fille semble avoir plus de ressources mais vous n'êtes pas sûr qu'elle ait assez de forces pour tenir jusqu'à votre retour. Que décidez-vous ?

deontological choice: "Prendre votre nièce"
consequentialist choice: "Prendre votre fille"

D.15 Silence au village

Des soldats ennemis ont envahi votre village. Ils ont pour ordre d'en tuer tous les habitants. Vous ainsi que plusieurs personnes du village êtes caché(e)s dans le cellier d'une grande maison. Vous entendez des soldats entrer dans la maison lorsque votre bébé commence à pleurer. Vous posez votre main sur sa bouche pour étouffer ses pleurs. Si vous enlevez votre main, les soldats seront alertés de votre présence et tueront tout le monde. Si vous n'enlevez pas votre main, votre bébé mourra étouffé. Que décidez-vous ?

deontological choice: "Enlever sa main"
consequentialist choice: "Laisser sa main en place"

D.16 Silence à la garderie

Des soldats ennemis ont envahi le village. Ils ont pour ordre d'en tuer tous les habitants. Vous vous cachez dans la garderie dans laquelle vous travaillez avec un bébé ainsi que plusieurs de vos employés. Vous entendez des soldats entrer dans le bâtiment lorsque le bébé commence à pleurer. Un de vos employés pose sa main sur la bouche du bébé pour étouffer ses pleurs. Il vous regarde ensuite et vous demande s'il doit enlever sa main. S'il l'enlève, les soldats seront alertés de votre présence et tueront tout le monde. S'il laisse sa main, le bébé mourra étouffé. Que décidez-vous ?

deontological choice: "Donner l'ordre d'enlever sa main"
consequentialist choice: "Donner l'ordre de laisser sa main en place"

D.17 Fuite de gaz dans l'hopital

Vous êtes gardien de nuit dans un hôpital. Vous êtes à votre poste lorsqu'une fuite de gaz mortel a lieu. Le gaz se répand dans les conduits d'aération du service de soin

intensif. Les patients dans ce service ne peuvent pas être déplacés. Si rien n'est fait, le gaz se répandra dans plusieurs chambres et tuera cinq personnes. Cependant, vous avez la possibilité d'actionner une trappe qui redirigera le gaz vers une seule chambre et seule une personne mourra. Que décidez-vous ?

deontological choice: "Ne rien faire"
consequentialist choice: "Actionner la trappe"

D.18 Drone vs lance-missile

Vous êtes opérateur d'un drone armé évoluant en zone de conflit. Vous avez pour objectif de détruire un lanceur de missiles visant une usine stratégique inoccupée le jour de votre mission. Lors de votre mission, vous détectez un missile se dirigeant vers un village. En tant qu'opérateur, vous pouvez soit décider d'intercepter le missile ce qui évitera la mort de plusieurs dizaines de civils mais détruira votre drone et vous empêchera de réaliser la mission qui vous a été confiée, soit accomplir votre mission, protégeant ainsi l'usine stratégique alliée et l'intégrité de votre drone, mais pas le village. Que décidez-vous ?

deontological choice: "Intercepter le missile"
consequentialist choice: "Accomplir la mission"

D.19 Agent secret

Vous êtes un agent secret en mission à l'étranger. Vous découvrez qu'un groupe terroriste cherche à assassiner trois diplomates. Ces derniers se trouvent dans une voiture en direction de l'Ambassade. Ils sont suivis par un des terroristes qui profitera du premier feu rouge pour les tuer. Alors que vous les suivez, un camion-citerne contenant de l'essence s'approche. Si vous tirez sur le camion-citerne, il explosera et cela barrera la route au terroriste, ce qui vous laissera le temps de vous échapper. Le conducteur du camion-citerne sera tué dans l'explosion mais les trois diplomates seront sauvés.

deontological choice: "Ne rien faire"
consequentialist choice: "Tirer sur le camion"

D.20 La grue

Vous travaillez comme conducteur de grue sur un chantier de construction. Vous commencez à peine votre journée sur le chantier, quand vous vous rendez compte que le câble de votre grue est sur le point de rompre. La barre d'acier attachée au câble se trouve juste au-dessus d'une équipe de six ouvriers. Vous avez la possibilité de déplacer le bras de la grue de quelques mètres vers une autre parcelle du chantier où seul un ouvrier travaille. L'ouvrier serait écrasé par la barre en acier et mourra mais vous sauvez les six ouvriers.

deontological choice: "Ne rien faire"
consequentialist choice: "Déplacer le bras de la grue"

D.21 Infirmier

"Vous êtes un infirmier responsable d'une machine contrôlant le dosage d'un médicament dans le sang des patients. À cause d'un problème technique, la machine délivre des doses létales de médicament à quatre patients. Un patient se trouvant dans une autre chambre est connecté à la même machine mais n'a subi aucune variation de dosage. Ce patient ne peut survivre si le médicament ne lui est pas délivré en continu. Vous pouvez appuyer sur un bouton qui bloquera la distribution du médicament auprès de tous les patients. Le patient de l'autre chambre mourra, mais les quatre patients seront sauvés.

deontological choice: "Ne rien faire"
consequentialist choice: "Arrêter la machine"

D.22 Mission de reconnaissance

Vous faites partie de l'aviation militaire et vous êtes le commandant d'un groupe d'avions en mission de reconnaissance. Durant une de vos missions où vous réalisez le survol d'une zone habitée, vous vous rendez compte qu'un avion vient de lancer un missile par erreur qui en tombant détruira une maison où se trouve trois personnes. Vous avez la possibilité de dérouter le missile. Il tombera sur une route où se trouve une voiture avec une personne à l'intérieur. Il tuera la personne dans la voiture mais

les trois personnes seront sauvées.

deontological choice: “Ne rien faire”
consequentialist choice: “Dérouter le missile”

D.23 Savane

Vous êtes éthologue et étudiez le comportement des lions des savanes en Afrique Centrale. Depuis votre observatoire sur pilotis, vous apercevez quatre personnes prises en chasse par plusieurs lions affamés. Leur seule chance de s’en sortir est d’arriver à rejoindre l’échelle de votre observatoire pour se mettre en sécurité mais les lions sont sur le point de les dévorer. Une autre personne s’est réfugiée dans un arbre jouxtant votre observatoire. Vous avez la possibilité de faire tomber la personne devant les lions. Elle mourra dévorée mais cela laissera le temps aux quatre personnes cachées derrière le buisson de se réfugier dans votre observatoire.

deontological choice: “Ne rien faire”
consequentialist choice: “Pousser la personne sur les lions”

D.24 Fusillade

Vous êtes au restaurant avec quelques amis. Un criminel rentre dans le restaurant une mitraillette à la main. L’homme menace de tuer six personnes assises à une table si une somme d’argent exorbitante ne lui est pas versée immédiatement. De plus, toute personne qui bouge sera abattue Vous avez la possibilité de pousser le serveur à côté de vous afin de faire tomber l’assaillant qui pourra par la suite être neutralisé. Le serveur mourra sous les balles, mais les six personnes seront sauvées.

deontological choice: “Ne rien faire”
consequentialist choice: “Pousser le serveur sur le criminel”

Appendix E

Results of experiment 1

E.1 Participants' statistics

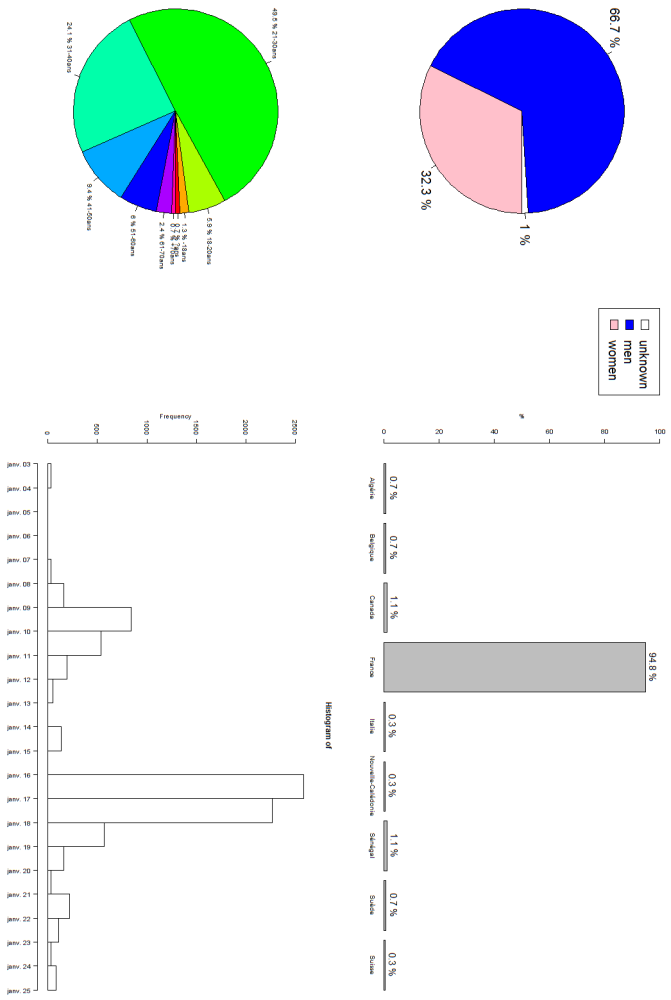


Fig. E.1 Age, gender and country of participants and dates of votes

E.2 Dilemmas' arguments

In this section, we present the available arguments for each dilemma and the relevancy score they have obtained. The argument selected for the main experiment is in bold font in the text, and is related to the white column in the graph.

E.2.1 Le tramway

Consequentialist arguments for decision "Actionner le levier":

- "Il vaut mieux sauver cinq personnes, quitte à en sacrifier une autre."
- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir cinq."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- **"Il est inacceptable de rediriger une menace vers une personne."**
- "Sacrifier une personne innocente est inacceptable."
- "Il vaut mieux ne pas agir que de faire du mal."

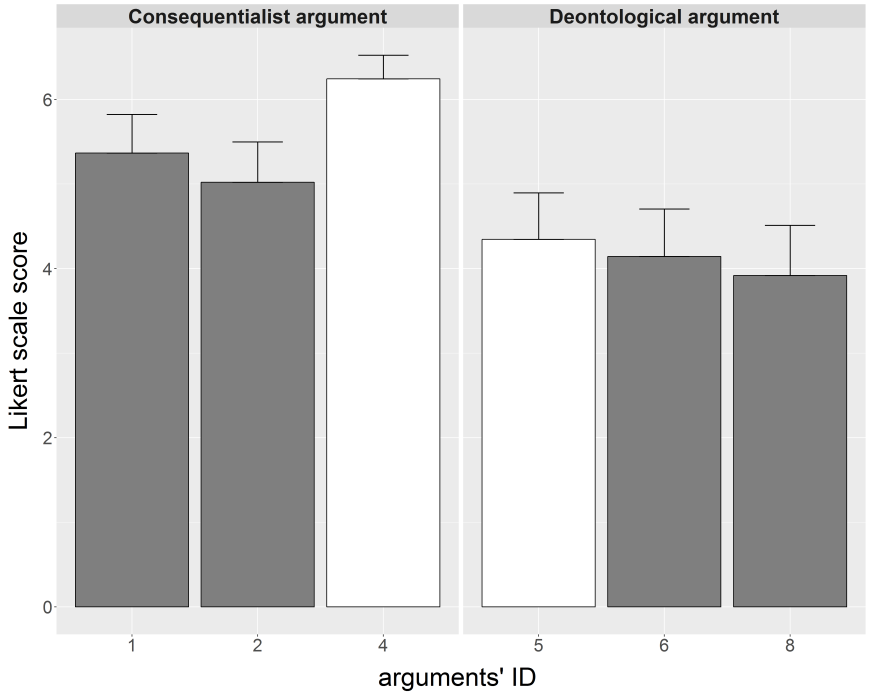


Fig. E.2 Relevance of arguments for dilemma "Le tramway"

E.2.2 La passerelle

Consequentialist arguments for decision "Pousser l'individu":

- "Il vaut mieux sauver cinq personnes, quitte à en sacrifier une autre."
- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir cinq."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- **"Il est inacceptable d'utiliser une personne comme un simple moyen."**
- "Il est inacceptable de sacrifier une personne."
- "Il vaut mieux laisser mourir que sacrifier."
- "Il est inacceptable de sacrifier quelqu'un pour sauver des vies."

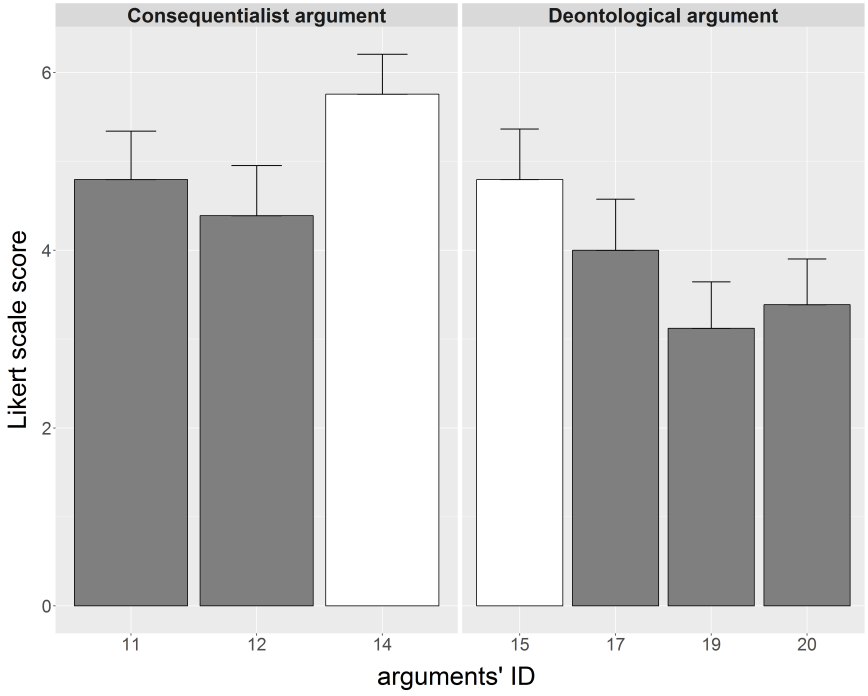


Fig. E.3 Relevance of arguments for dilemma "La passerelle"

E.2.3 La trappe

Consequentialist arguments for decision “Actionner la trappe”:

- “Il vaut mieux sauver cinq personnes, quitte à en sacrifier une autre.”
- “Il est préférable de sacrifier une personne plutôt que d’en laisser mourir cinq.”
- **“Il faut sauver le plus de personnes possible.”**

Deontologicial arguments for decision “Ne rien faire”:

- “Il est inacceptable d’utiliser une personne comme un simple moyen, il est donc préférable de ne pas agir.”
- **“Il est inacceptable de sacrifier une personne.”**
- “Il vaut mieux laisser mourir que sacrifier.”
- “Il est inacceptable de sacrifier quelqu’un pour sauver des vies.”

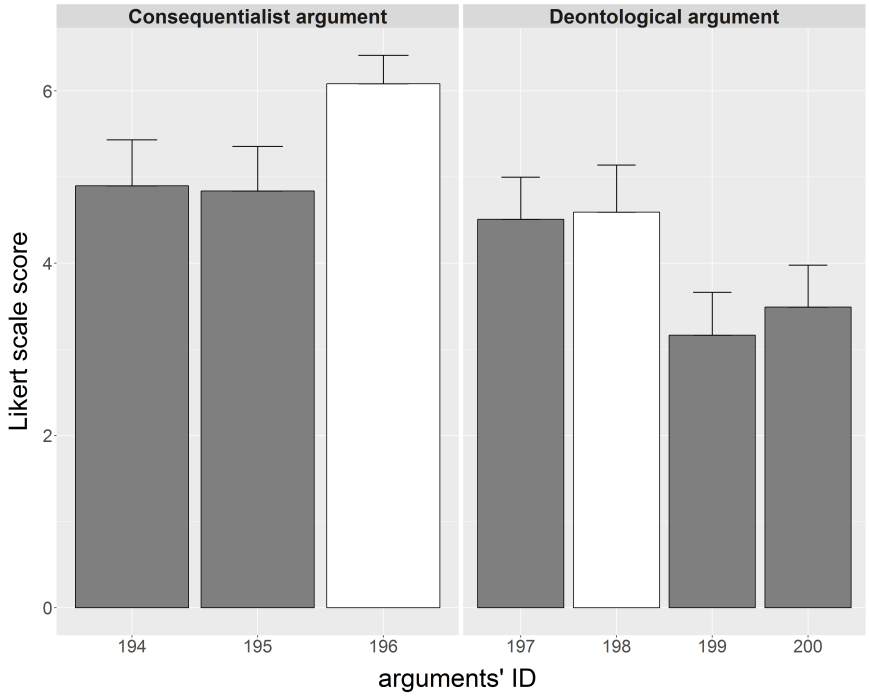


Fig. E.4 Relevance of arguments for dilemma “La trappe”

E.2.4 Transplantation d'organes

Consequentialist arguments for decision "Transplanter les organes":

- "Il est préférable de sauver cinq patients, quitte à sacrifier une personne."
- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir cinq."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- **"Il est inacceptable d'utiliser une personne comme un simple moyen."**
- "Il est inacceptable de sacrifier une personne."
- "Il vaut mieux laisser mourir que sacrifier."
- "Il est inacceptable de sacrifier quelqu'un pour sauver des vies."

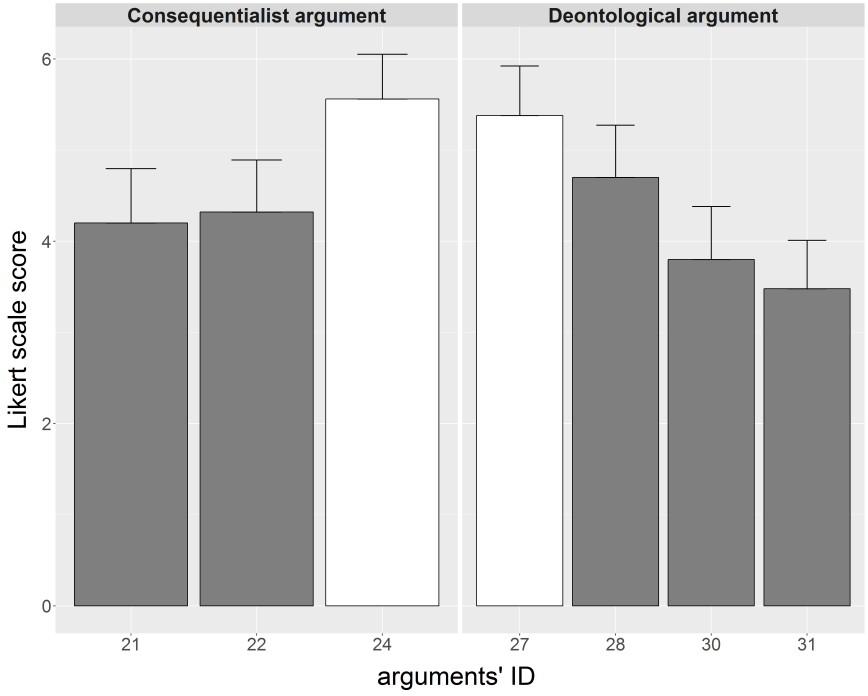


Fig. E.5 Relevance of arguments for dilemma "Transplantation d'organes"

E.2.5 Le choix de Sophie

Consequentialist arguments for decision “Livrer un de vos enfants”:

- “Il est préférable de sacrifier un de ses enfants plutôt que de n’en sauver aucun.”
- “Il est préférable de sauver un de ses enfants quitte à devoir sacrifier l’autre.”
- **“Il est inacceptable de laisser mourir un enfant qu’on pourrait sauver.”**

Deontologicial arguments for decision “Ne rien faire”:

- “Il est inacceptable d’utiliser un enfant comme un simple moyen.”
- **“Il est inacceptable de choisir entre ses deux enfants.”**
- “Il est inacceptable d’envoyer son enfant à la mort.”

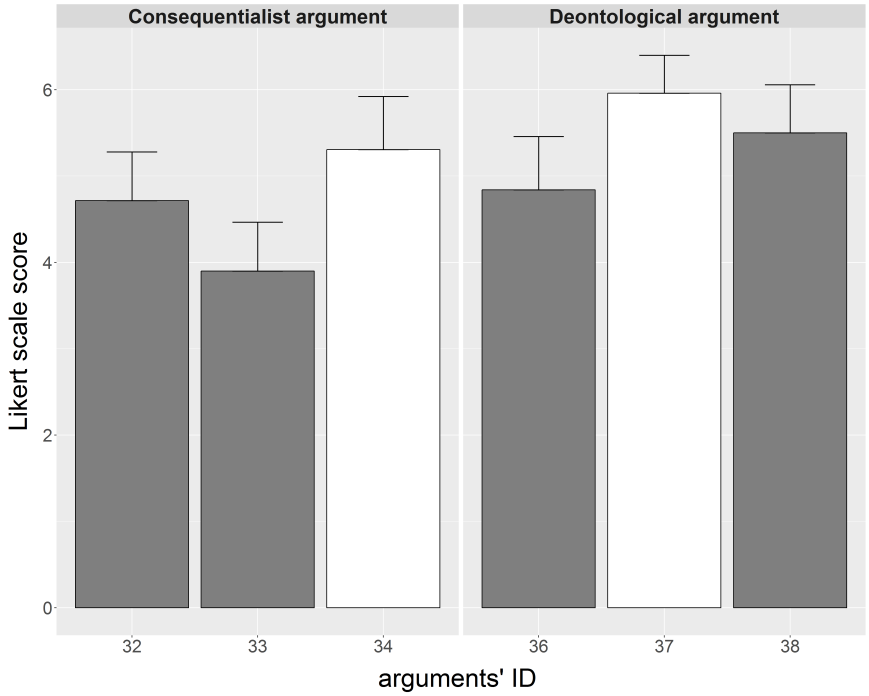


Fig. E.6 Relevance of arguments for dilemma “Le choix de Sophie”

E.2.6 Le choix du maire

Consequentialist arguments for decision “Choisir un des deux enfants”:

- “Il est préférable de sacrifier un enfant plutôt que de n’en sauver aucun.”

- "Il est préférable de sauver un enfant, quitte à devoir sacrifier l'autre."
- **"Il est inacceptable de laisser mourir un enfant qu'on pourrait sauver."**

Deontologicial arguments for decision "Ne rien faire":

- 'Il est inacceptable d'utiliser un enfant comme un simple moyen.'"
- "Il est inacceptable de choisir entre deux enfants."
- **"Il est inacceptable d'envoyer un enfant à la mort."**

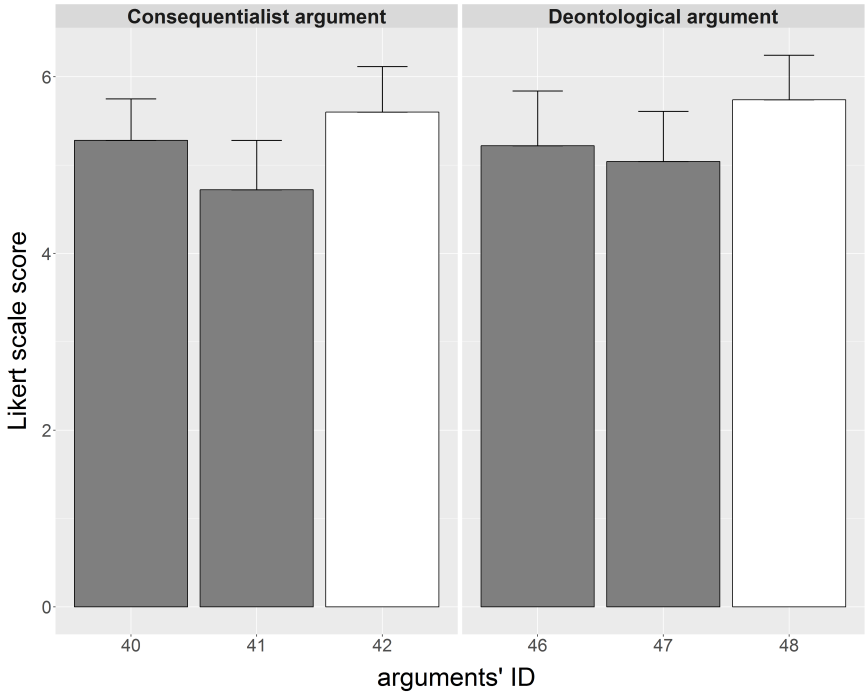


Fig. E.7 Relevance of arguments for dilemma "Le choix du maire"

E.2.7 Le choix de Corinne

Consequentialist arguments for decision "Refuser de donner le mot de passe":

- "Il est préférable de sauver le maximum de personnes, quitte à laisser mourir votre fils."
- **"Il faut sauver le plus de vies possible."**
- "Il est préférable de laisser mourir une personne et d'en sauver plusieurs."

- "Il est inacceptable de laisser mourir tous vos collègues."

Deontologicial arguments for decision "Donner le code":

- **"Il est inacceptable de laisser mourir son enfant."**
- "Il est inacceptable d'utiliser un enfant comme un moyen."
- "Un enfant doit toujours être protégé."

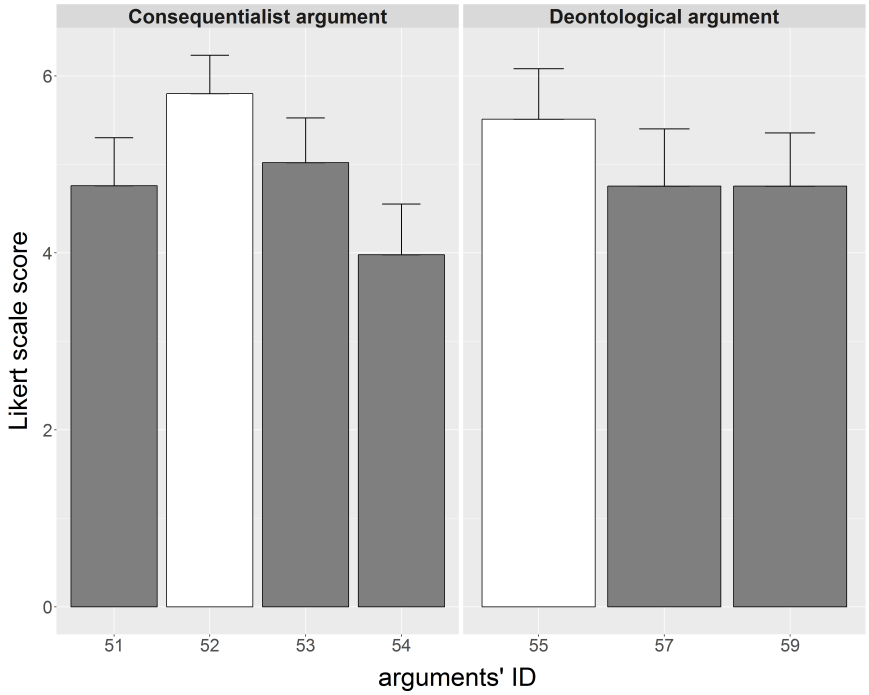


Fig. E.8 Relevance of arguments for dilemma "Le choix de Corinne"

E.2.8 Canot de sauvetage

Consequentialist arguments for decision "Pousser l'homme blessé":

- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir plusieurs."
- "Le passager est déjà condamné, il serait suicidaire de le garder à bord."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- **"Il est inacceptable de sacrifier une personne, quelle que soit sa condition."**

- "L'homme blessé a réussi à embarquer, on ne peut pas le jeter à eau."
- "Il est inacceptable de sacrifier quelqu'un pour sauver des vies."

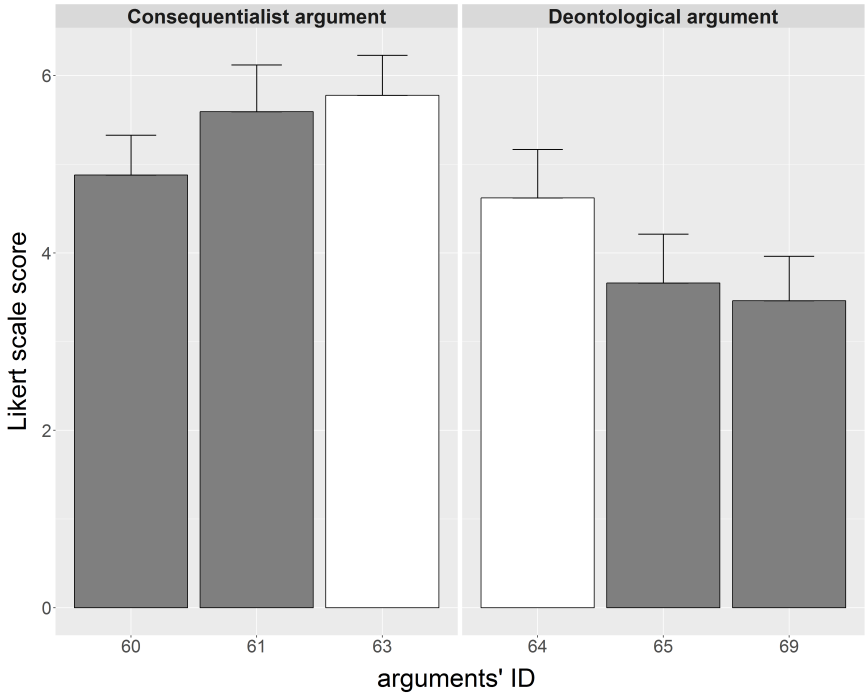


Fig. E.9 Relevance of arguments for dilemma "Canot de sauvetage"

E.2.9 Escalade

Consequentialist arguments for decision "Couper la corde":

- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir plusieurs."
- "Il serait suicidaire de garder une personne blessée dans la cordée."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- **"Il est inacceptable de sacrifier une personne."**
- "Votre ami fait partie du groupe, on ne peut pas le sacrifier."
- "Il est inacceptable de sacrifier quelqu'un qui a besoin d'aide."

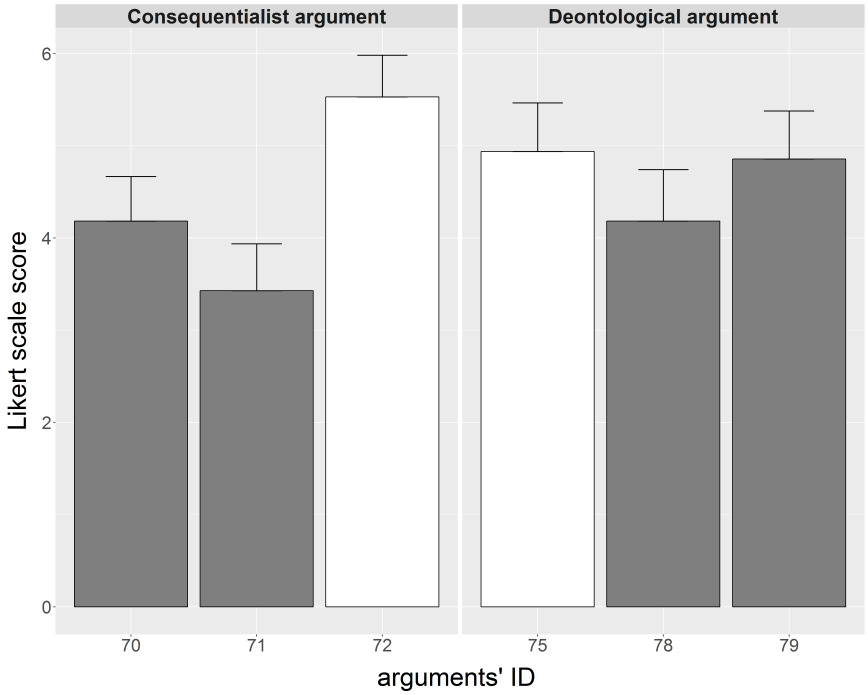


Fig. E.10 Relevance of arguments for dilemma “Escalade”

E.2.10 Spéléologie

Consequentialist arguments for decision “Utiliser la dynamite”:

- “Il est préférable de sauver le maximum de personnes, quitte à en sacrifier une autre.”
- “Il est préférable de sacrifier une personne plutôt que de laisser mourir tout le groupe.”
- “Il est inacceptable de laisser mourir le groupe, quel que soit le moyen utilisé.”
- **“Il faut sauver le plus de personnes possible.”**

Deontologicial arguments for decision “Ne rien faire”:

- **“Il est inacceptable de sacrifier une personne.”**
- “Il est inacceptable de sacrifier quelqu’un pour sauver d’autres personnes.”
- “Il est préférable de laisser mourir le groupe plutôt que de sacrifier une personne.”
- “Il est inacceptable de sacrifier quelqu’un pour sauver des vies.”

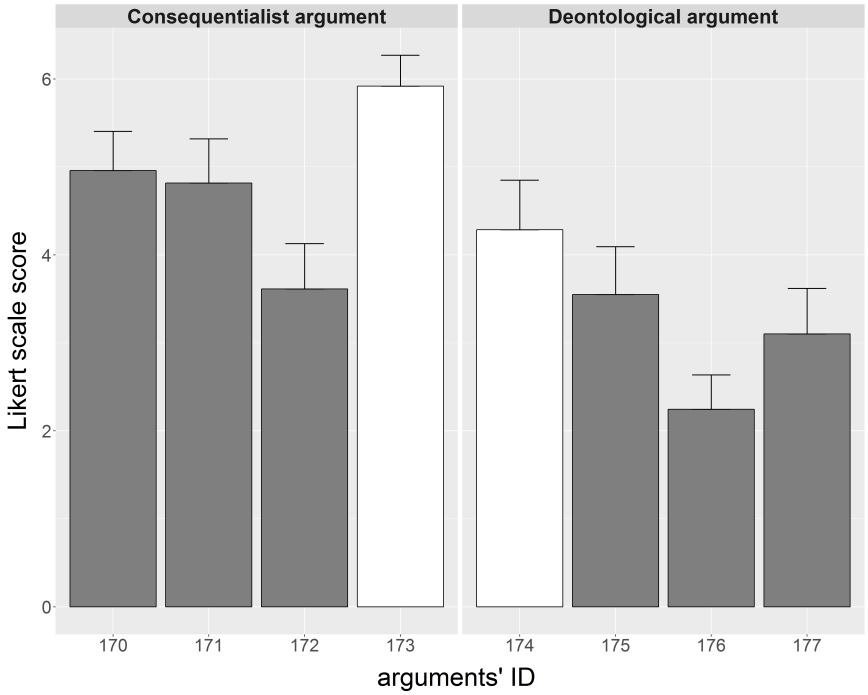


Fig. E.11 Relevance of arguments for dilemma "Spéléologie"

E.2.11 Sauvetage pompier spéléologique

Consequentialist arguments for decision "Donner l'ordre d'utiliser la dynamite":

- "Il est préférable de sauver le maximum de personnes, quitte à en sacrifier une autre."
- "Il est préférable de sacrifier une personne plutôt que de laisser mourir tout le groupe."
- "Il est inacceptable de laisser mourir le groupe, quel que soit le moyen utilisé."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Donner l'ordre de dégager le guide":

- **"Il est inacceptable de sacrifier une personne."**
- "Il est inacceptable de sacrifier quelqu'un pour sauver d'autres personnes."
- "Il est préférable de laisser mourir le groupe plutôt que de sacrifier une personne."
- "Il est inacceptable de sacrifier quelqu'un pour sauver des vies."

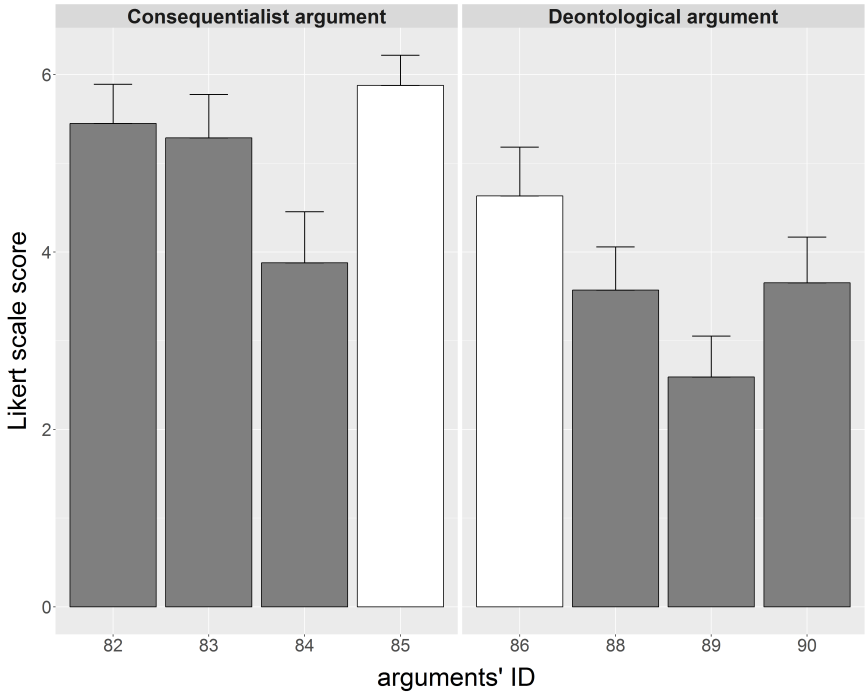


Fig. E.12 Relevance of arguments for dilemma “Sauvetage pompier spéléologique”

E.2.12 Conflit éthnique

Consequentialist arguments for decision “Désigner une personne”:

- “Il est préférable de sauver un maximum de personnes, quitte à sacrifier une personne.”
- “Il est préférable de sacrifier une personne plutôt que d’en laisser mourir un grand nombre.”
- “Il est inacceptable de laisser mourir des personnes alors qu’on pourrait les sauver.”
- **“Il faut sauver le plus de personnes possible.”**

Deontologicial arguments for decision “Ne rien faire”:

- “Il est inacceptable de sacrifier un innocent.”
- “Il est inacceptable de sacrifier quelqu’un pour sauver des vies.”
- **“Il est injuste de condamner à mort une personne sans preuves.”**

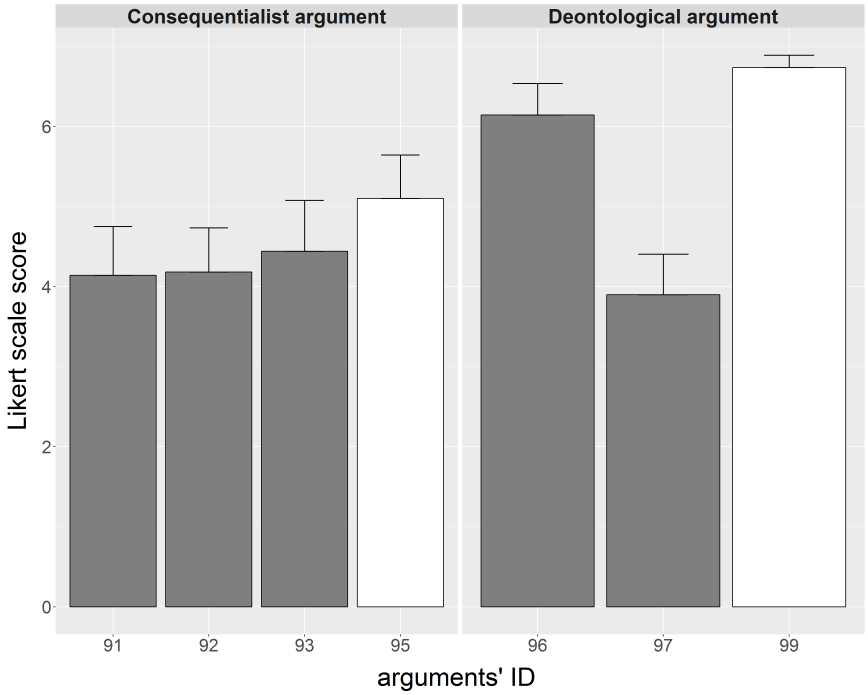


Fig. E.13 Relevance of arguments for dilemma "Conflit éthnique"

E.2.13 Torturer un terroriste

Consequentialist arguments for decision "Torturer le terroriste":

- "Il est préférable de porter atteinte à une personne plutôt que d'en laisser mourir un grand nombre."
- **"Il est inacceptable de laisser mourir des personnes alors qu'on pourrait les sauver."**
- "Il est préférable de sauver le plus de vies possible."

Deontologicial arguments for decision "Ne rien faire":

- "Il est inacceptable de porter atteinte à une personne, quel que soit le contexte."
- **"Il est inacceptable de torturer une personne, quel que soit le contexte"**
- "Il faut respecter l'intégrité des personnes inculpées."

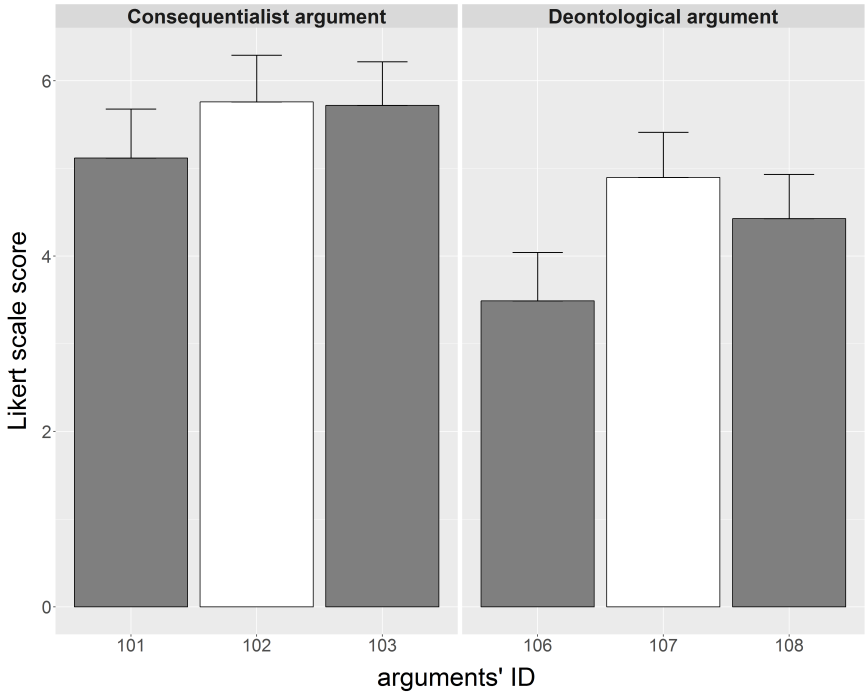


Fig. E.14 Relevance of arguments for dilemma “Torturer un terroriste”

E.2.14 Noyade

Consequentialist arguments for decision “Prendre votre fille”:

- “Il est préférable de sauver sa fille, quitte à laisser mourir sa nièce.”
- **“Il est plus important de sauver votre enfant.”**
- “Votre nièce a la plus faible chance de survie, il est préférable de sauver votre fille.”

Deontologicial arguments for decision “Prendre votre nièce”:

- “Il est inacceptable de ne pas sauver sa nièce.”
- “Il est inacceptable de choisir sa fille plutôt que sa nièce.”
- **“Il est inacceptable de ne pas aider un enfant sur le point mourir.”**

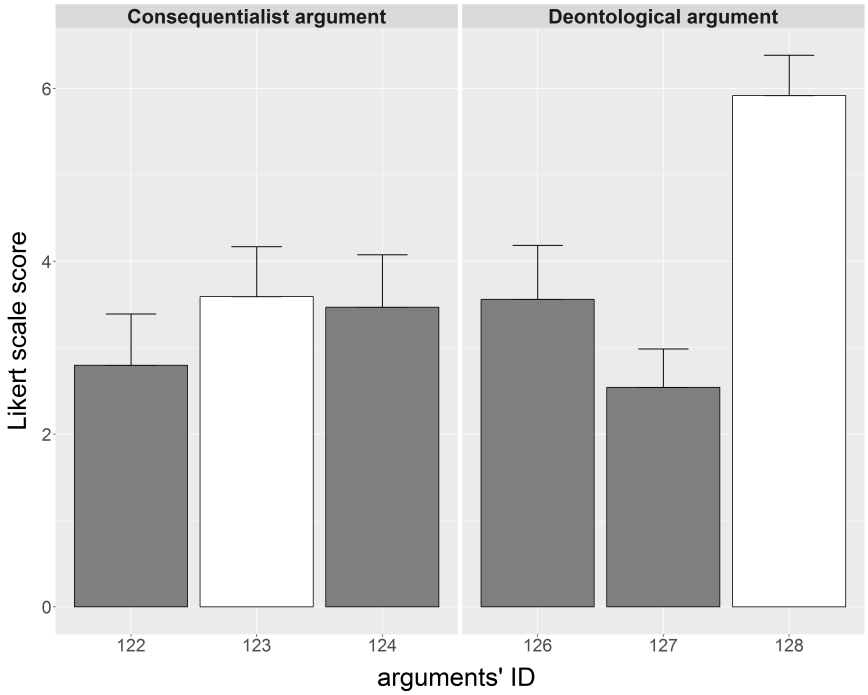


Fig. E.15 Relevance of arguments for dilemma "Noyade"

E.2.15 Silence au village

Consequentialist arguments for decision "Laisser sa main en place":

- "Il est préférable de sauver un maximum de personnes, quitte à sacrifier un bébé."
- "Il est préférable de sacrifier un bébé plutôt que de laisser mourir plusieurs personnes."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Enlever sa main":

- **"Il est inacceptable de sacrifier la vie d'un innocent."**
- "Il est inacceptable de sacrifier un bébé."
- "Il est préférable de mourir en ne sacrifiant pas son enfant."
- "Il est inacceptable de sacrifier une personne pour sauver des vies."

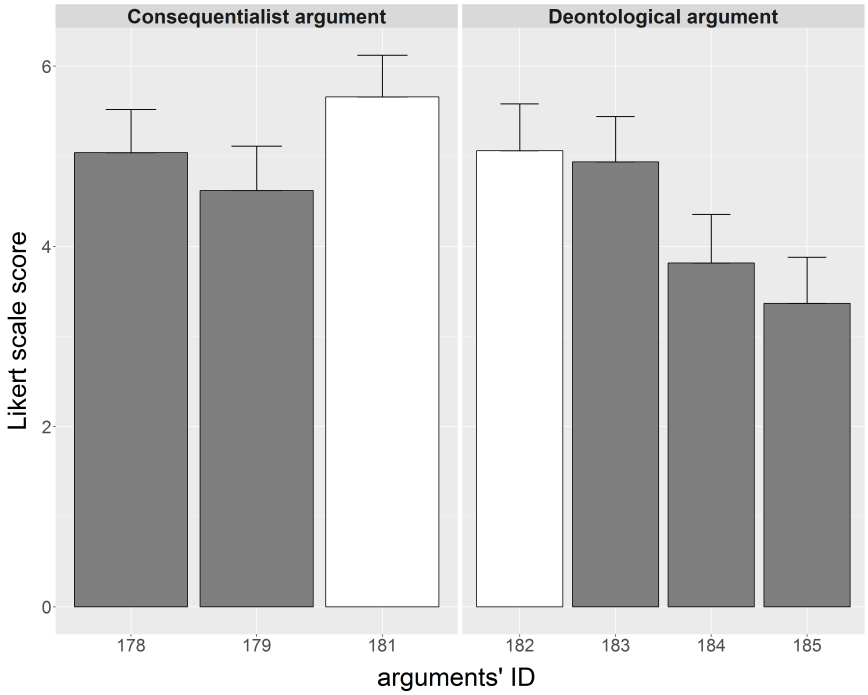


Fig. E.16 Relevance of arguments for dilemma “Silence au village”

E.2.16 Silence à la garderie

Consequentialist arguments for decision “Donner l’ordre de laisser sa main en place”:

- “Il est préférable de sauver un maximum de personnes, quitte à sacrifier un bébé.”
- “Il est préférable de sacrifier un bébé plutôt que de laisser mourir plusieurs personnes.”
- “Il est inacceptable de laisser mourir des personnes alors qu’on pourrait les sauver.”
- **“Il faut sauver le plus de personnes possible.”**

Deontologicial arguments for decision “Donner l’ordre d’enlever sa main”:

- **“Il est inacceptable de sacrifier un innocent.”**
- “Il est inacceptable de sacrifier un bébé pour sauver d’autres personnes.”
- “Il est préférable de mourir plutôt que d’étouffer un bébé.”
- “Il est inacceptable d’autoriser quelqu’un à sacrifier un bébé.”

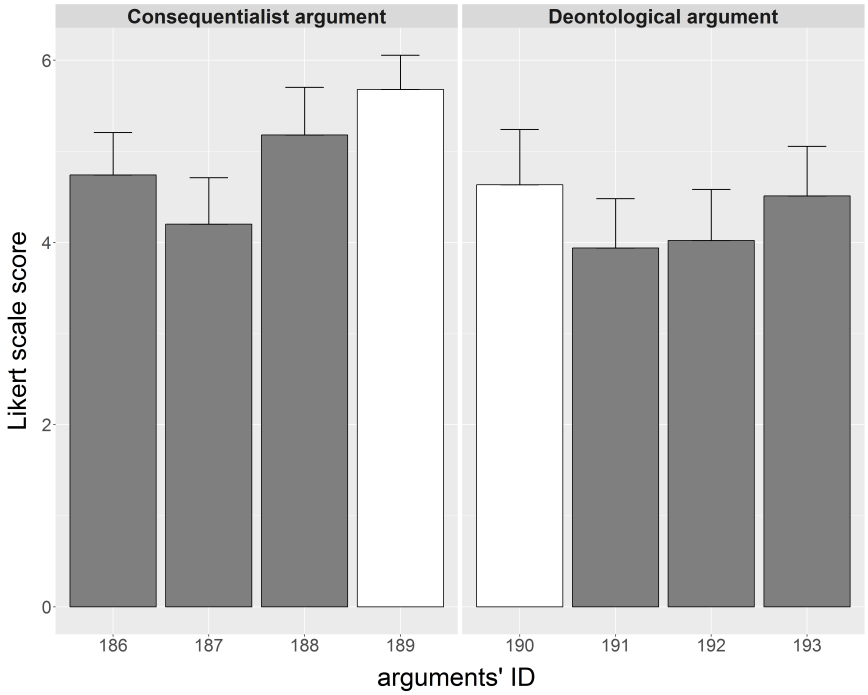


Fig. E.17 Relevance of arguments for dilemma "Silence à la garderie"

E.2.17 Fuite de gaz dans l'hopital

Consequentialist arguments for decision "Actionner la trappe":

- "Il vaut mieux sauver cinq personnes, quitte à en sacrifier une autre."
- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir cinq."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- "Il est inacceptable de rediriger une menace vers une personne."
- **"Sacrifier une personne innocente est inacceptable."**
- "Il vaut mieux ne pas agir que de faire du mal."

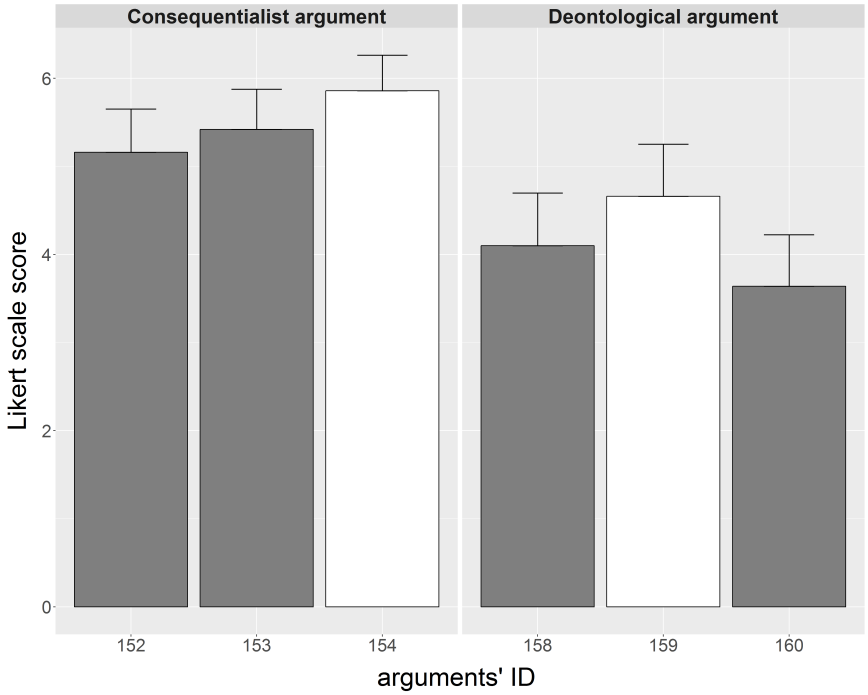


Fig. E.18 Relevance of arguments for dilemma “Fuite de gaz dans l’hopital”

E.2.18 Drone vs lance-missile

Consequentialist arguments for decision “Accomplir la mission”:

- “Les bénéfices de la sauvegarde de l’usine sont plus importants. Il est préférable de garder des points stratégiques dans le conflit plutôt que de protéger des civils.”
- **“Il est préférable de ne pas faillir à une mission stratégique de haute importance.”**
- “Il faut privilégier la protection de la nation.”

Deontologicial arguments for decision “Intercepter le missile”:

- “Il est inacceptable de laisser des civils être blessés.”
- “Protéger des civils est un concept fondamental en cas de conflit.”
- **“Il est inacceptable de ne pas protéger des êtres humains quand on le peut.”**

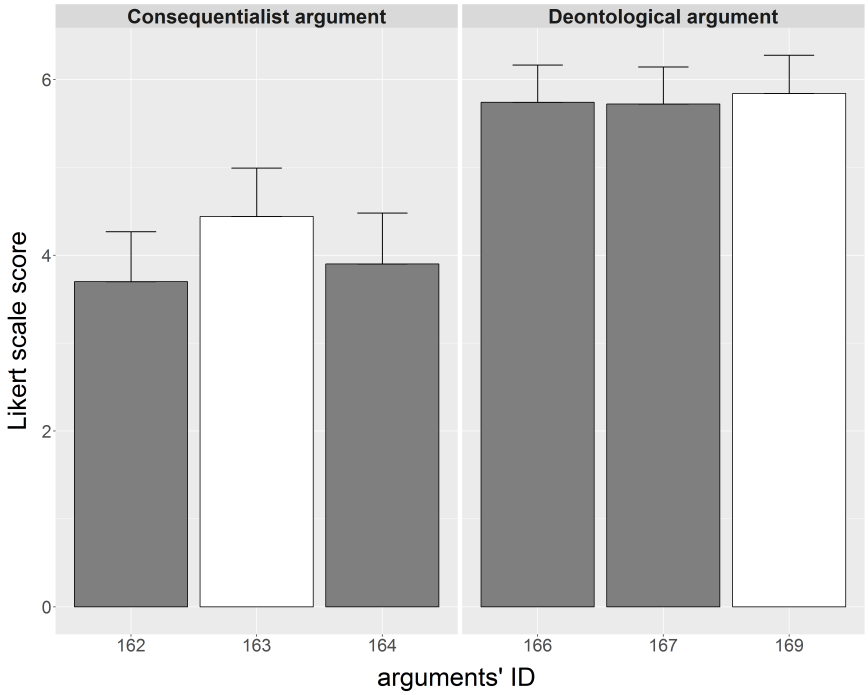


Fig. E.19 Relevance of arguments for dilemma "Drone vs lance-missile"

E.2.19 Agent secret

Consequentialist arguments for decision "Tirer sur le camion":

- "Il vaut mieux sauver trois personnes, quitte à en sacrifier une autre."
- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir trois."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- **"Il est inacceptable de sacrifier une personne, même pour sauver des personnes de haut rang."**
- "Sacrifier une personne innocente pour sauver des diplomates est inacceptable."
- "Il vaut mieux ne pas agir que de faire du mal."

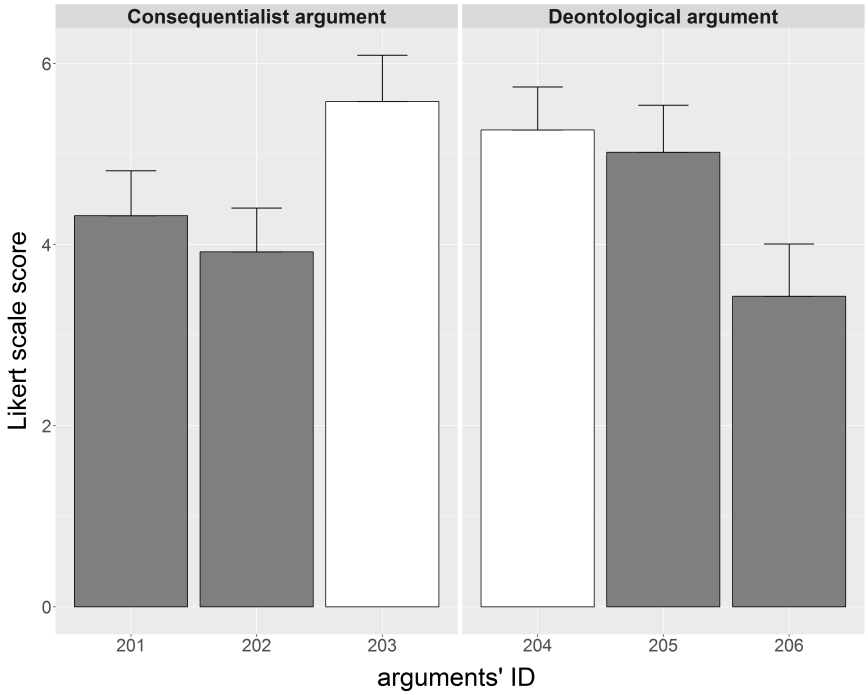


Fig. E.20 Relevance of arguments for dilemma “Agent secret”

E.2.20 La grue

Consequentialist arguments for decision “Déplacer le bras de la grue”:

- “Il vaut mieux sauver six personnes, quitte à en sacrifier une autre.”
- “Il est préférable de sacrifier une personne plutôt que d’en laisser mourir six.”
- **“Il faut sauver le plus de personnes possible.”**

Deontologicial arguments for decision “Ne rien faire”:

- “Il est inacceptable de rediriger une menace vers une personne.”
- **“Sacrifier une personne innocente est inacceptable.”**
- “Il vaut mieux ne pas agir que de faire du mal.”

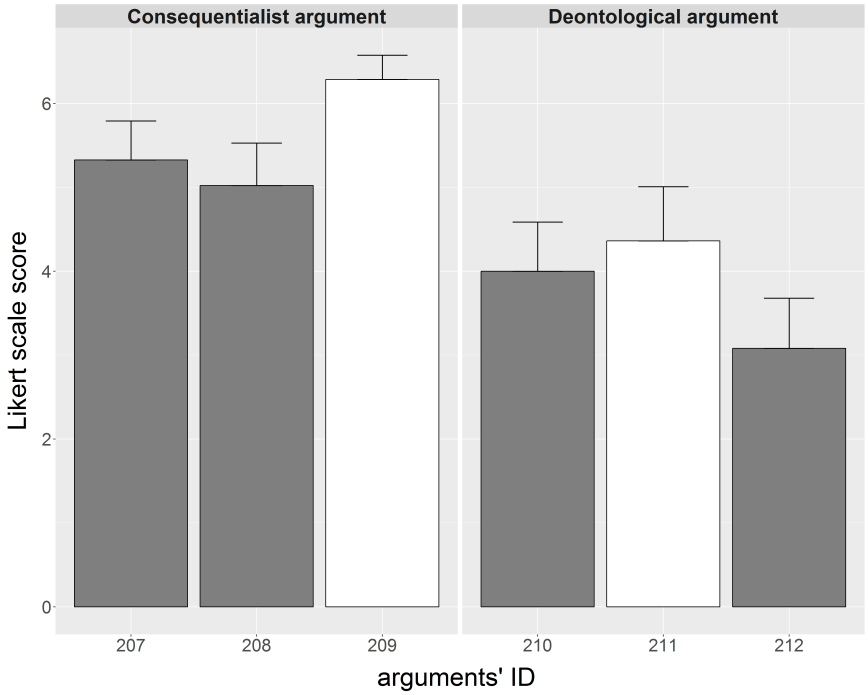


Fig. E.21 Relevance of arguments for dilemma "La grue"

E.2.21 Infirmier

Consequentialist arguments for decision "Arrêter la machine":

- "Il vaut mieux sauver quatre personnes, quitte à en sacrifier une autre."
- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir quatre."
- "Le patient est dépendant de la machine, il vaut mieux sauver les autres patients."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- "Il est inacceptable de rediriger une menace vers une personne."
- **"Sacrifier une personne innocente est inacceptable."**
- "Il vaut mieux ne pas agir que de faire du mal."
- "Il faut protéger les personnes les plus faibles."

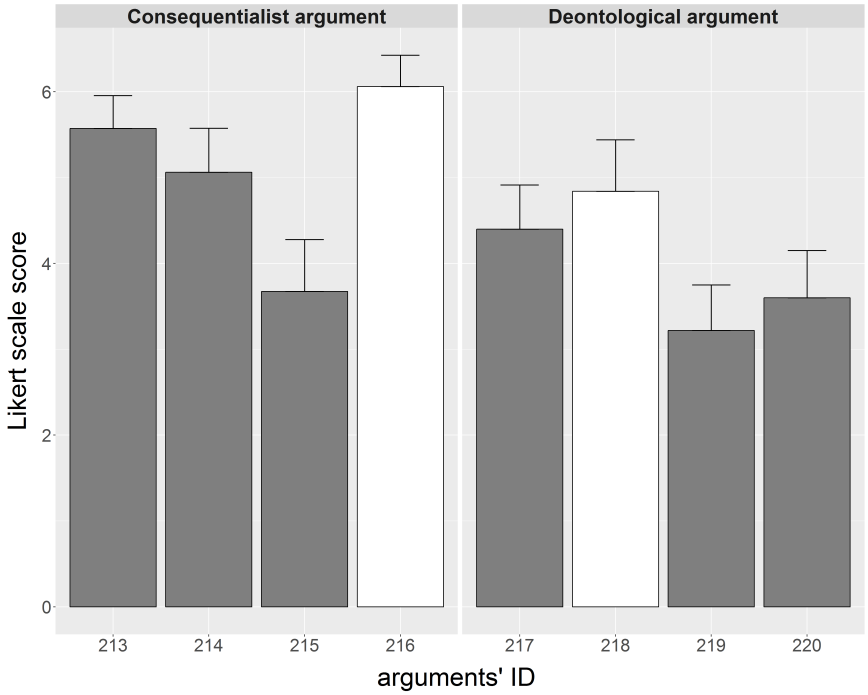


Fig. E.22 Relevance of arguments for dilemma “Infirmier”

E.2.22 Mission de reconnaissance

Consequentialist arguments for decision “Dérouter le missile”:

- “Il vaut mieux sauver trois personnes, quitte à en sacrifier une autre.”
- “Il est préférable de sacrifier une personne plutôt que d’en laisser mourir trois.”
- **“Il faut sauver le plus de personnes possible.”**

Deontologicial arguments for decision “Ne rien faire”:

- “Il est inacceptable de rediriger une menace vers une personne.”
- **“Sacrifier une personne innocente est inacceptable.”**
- “Il vaut mieux ne pas agir que de faire du mal.”

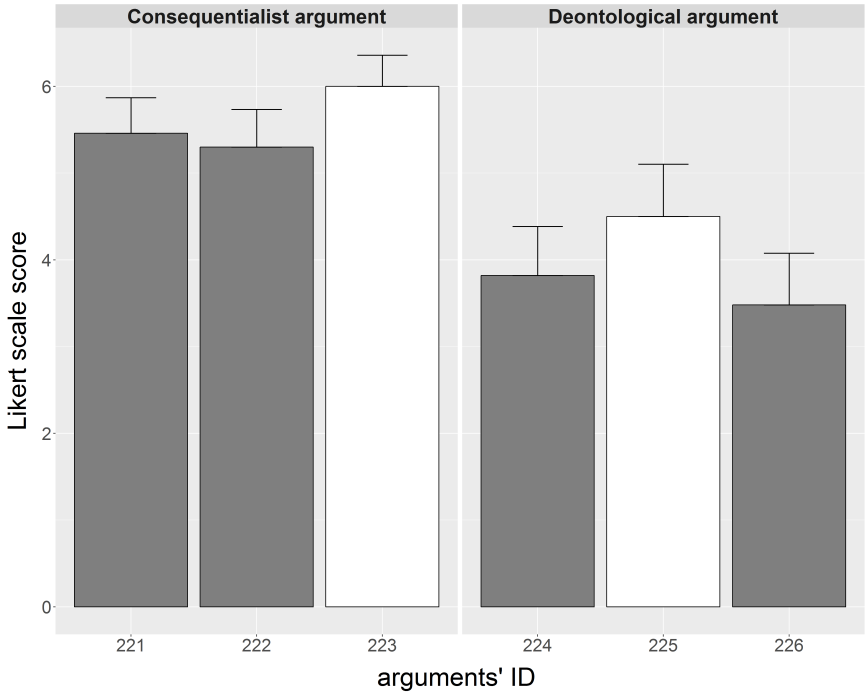


Fig. E.23 Relevance of arguments for dilemma "Mission de reconnaissance"

E.2.23 Savane

Consequentialist arguments for decision "Pousser la personne sur les lions":

- "Il vaut mieux sauver quatre personnes, quitte à en sacrifier une autre."
- "Il est préférable de sacrifier une personne plutôt que d'en laisser mourir quatre."
- "Sauver les quatres personnes des lions est la priorité."
- **"Il faut sauver le plus de personnes possible."**

Deontologicial arguments for decision "Ne rien faire":

- "Il est inacceptable de rediriger une menace vers une personne."
- "Sacrifier une personne innocente est inacceptable."
- **"Jeter une personne sur des lions est inacceptable."**
- "Il vaut mieux ne pas agir que de faire du mal."

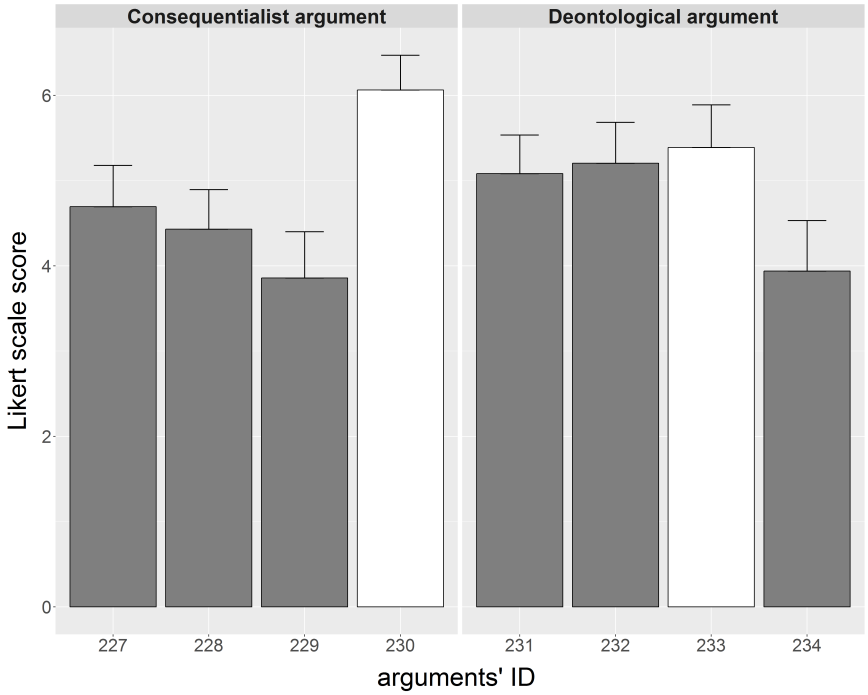


Fig. E.24 Relevance of arguments for dilemma “Savane”

E.2.24 Fusillade

Consequentialist arguments for decision “Pousser le serveur sur le criminel”:

- “Il vaut mieux sauver six personnes, quitte à en sacrifier une autre.”
- “Il est préférable de sacrifier une personne plutôt que d’en laisser mourir six.”
- **“Protéger le maximum de personnes est la priorité.”**
- “Il faut sauver le plus de personnes possible.”

Deontologicial arguments for decision “Ne rien faire”:

- “Il est inacceptable de rediriger une menace vers une personne.”
- “Sacrifier une personne innocente est inacceptable.”
- **“Il est inacceptable de provoquer la mort d’une personne.”**
- “Il vaut mieux ne pas agir que de faire du mal.”

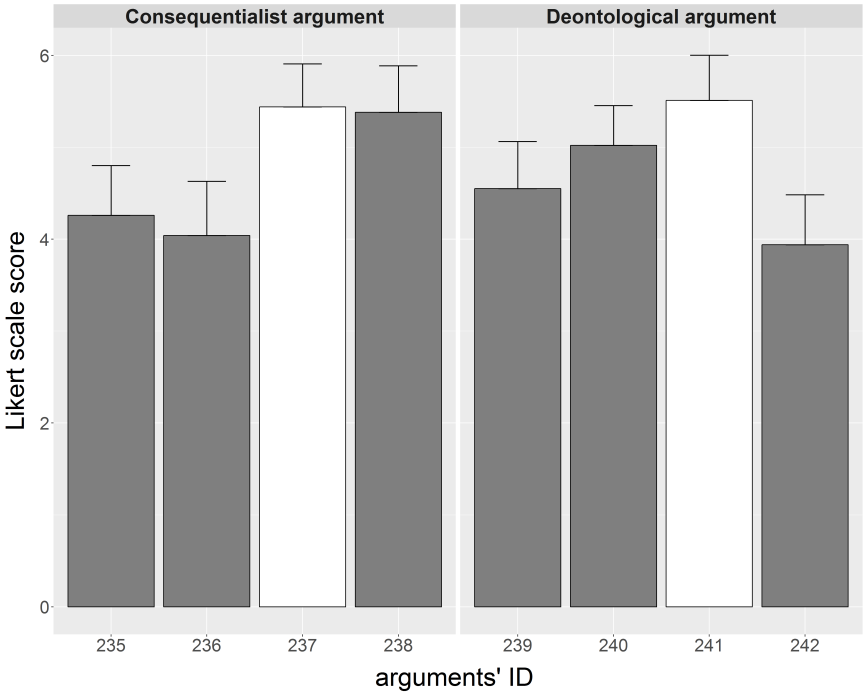


Fig. E.25 Relevance of arguments for dilemma "Fusillade"

Appendix F

Results of experiment 2

F.1 Participants' statistics

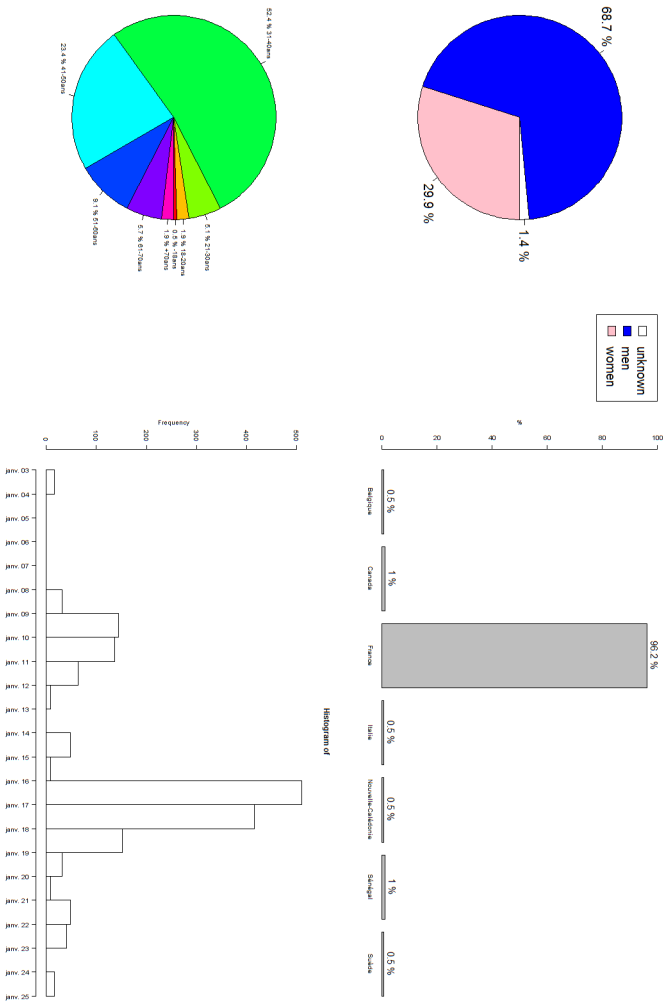


Fig. F.1 Age, gender and country of participants and dates of votes

F.2 Responsibility and intensity statistics

Fig. F.2 Assessment of responsibility and intensity for each dilemma through a Likert scale

F.3 Dilemma categorization statistics

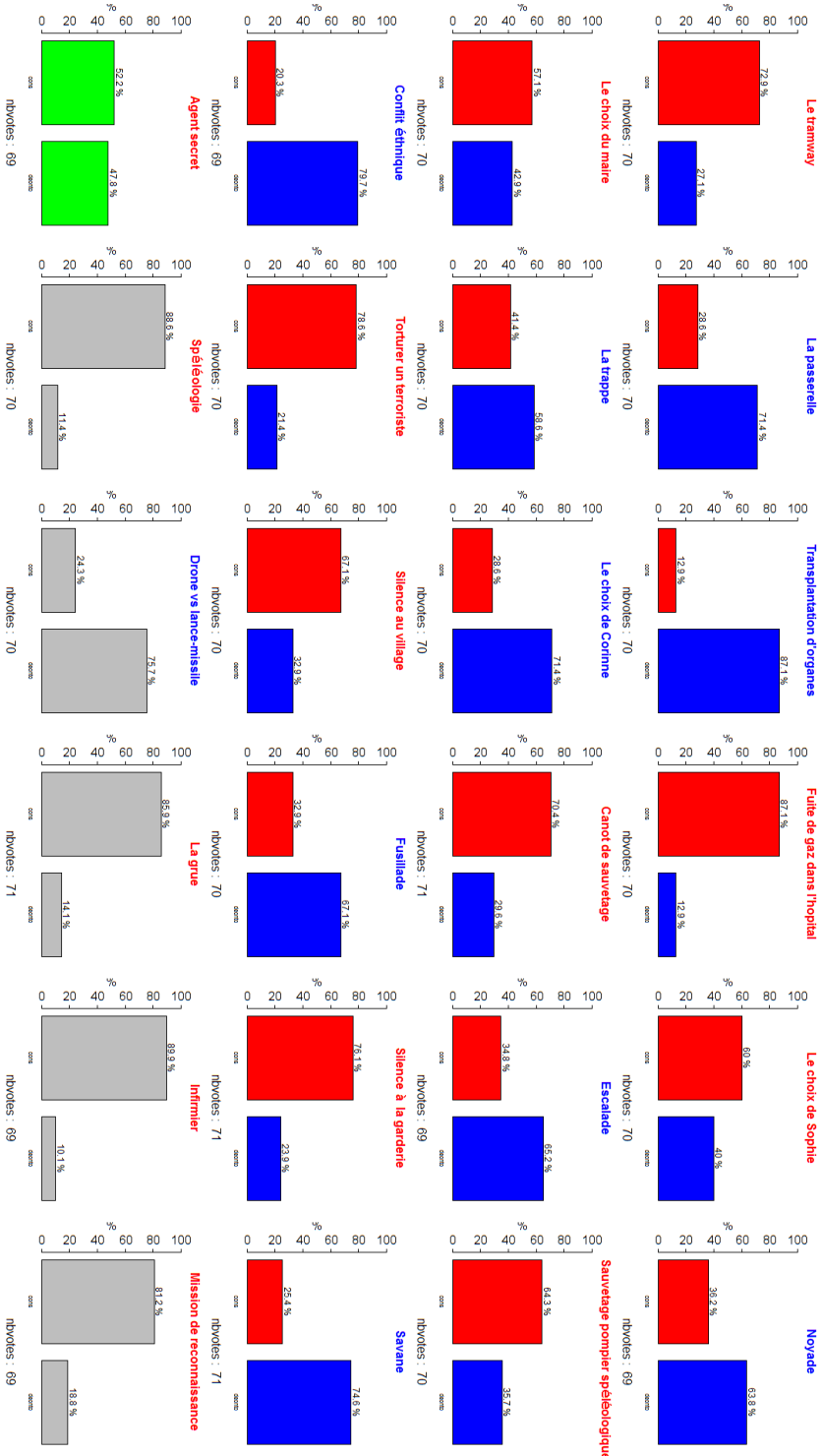


Fig. F.3 Categorization of dilemmas