Ethical Judgment of Agents' Behaviors in Multi-Agent Systems

Nicolas Cointe
Institut Henri Fayol, EMSE
LabHC, UMR CNRS 5516
F-42000, Saint-Etienne,
France
nicolas.cointe@emse.fr

Grégory Bonnet
Normandie University
GREYC, CNRS UMR 6072
F-14032 Caen, France
gregory.bonnet@unicaen.fr

Olivier Boissier
Institut Henri Fayol, EMSE
LabHC, UMR CNRS 5516
F-42000, Saint-Etienne,
France
olivier.boissier@emse.fr

ABSTRACT

The increasing use of multi-agent technologies in various areas raises the necessity of designing agents that judge ethical behaviors in context. This is why several works integrate ethical concepts in agents' decision making processes. However, those approaches consider mainly an agent-centered perspective, letting aside the fact that agents are in interaction with other artificial agents or human beings that can use other ethical concepts. In this article, we address the problem of producing ethical behaviors from a multi-agent perspective. To this end, we propose a model of ethical judgment an agent can use in order to judge the ethical dimension of both its own behavior and the other agents' behaviors. This model is based on a rationalist and explicit approach that distinguishes theory of good and theory of right. A proof-of-concept implemented in Answer Set Programming and based on a simple scenario is given to illustrate those functionalities.

General Terms

Theory

Keywords

Multi-Agent Systems, Ethical Judgment,
Computational Ethics

1. INTRODUCTION

The increasing presence of autonomous agents in various fields as health care, high-frequency trading, transportation and so on, may rise many issues if these agents are not able to consider and follow some rules, and adapt their behavior. These rules can be some simple constraints as a communication protocol or prohibition of certain behaviors, or some more complex ones as preferences of the user or the description of a code of deontology. For instance, the understanding of codes of conduct may ease the cooperation between a practitioner and a medical agent or a patient, considering some concepts as medical secrecy or dignity respect. Even if some works propose implementations of action restrictions [34], simple prohibitions or obligations [7], some codes of conducts use more complex notions such as moral values or ethical principles, and need further work. An explicit implementation of such concepts, like generosity or altruism, needs specific structures and processes in the architecture of the agent. Consequently, the interest in designing ethical autonomous agents has recently been raised in the Artificial Intelligence community [29] as highlighted by numerous articles [20, 23, 25, 26] and conferences[1]. However, all these works consider ethics in an individual – single-agent – point of view whereas numerous real-world applications such as transportation or high-frequency trading involve multiple agents, enforcing the need to consider the collective – multi-agent – point of view.

An individual point of view can be enough for an agent to act ethically within an agent organization. However, to evaluate the behavior of another agent (e.g. to collaborate with or to punish), agents need to be able to judge the ethics of the others. In this article, we are interested in the question of ethical judgment, i.e. the assessment of appropriateness or not of agents' behavior with respect to moral convictions and ethical principles. We propose a generic model for the judgment of behaviors that can be used by an agent both to decide on its behavior and to judge the behavior of others.

The remainder of the article is organized as follows. Section 2 introduces some key concepts of moral philosophy and a short state of the art on common approaches in computational ethics. We detail in Section 3 our ethical judgment model. Section 4 then illustrates the use of this model by an agent while interacting with other agents. Section 5 offers a proof of concept in ASP (Answer Set Programming). We compare our work to existing approaches in Section 6 and conclude in Section 7 by pointing out the importance of computational ethics for multi-agent systems, and by giving some perspectives about the next steps of our work.

2. ETHICS AND AUTONOMOUS AGENTS

We introduce first in Section 2.1 moral philosophy concepts on which we base our approach and review in section 2.2 existing autonomous agent architectures that propose ethical behaviors. Finally, Section 2.3 points out the principles of our approach.

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

[1] Symposium on Roboethics - www.roboethics.org,
International Conference on Computer Ethics and Philosophical Enquiry - philevents.org/event/show/15670,
Workshop on AI and Ethics, AAAI conference - www.cse.unsw.edu.au/~tw/aiethics,
International Conference on AI and Ethics - wordpress.csc.liv.ac.uk/va/2015/02/16/.

2.1 Moral philosophy concepts

From ancient philosophers to recent works in neurology [10] and cognitive sciences [16], many studies have been interested in the capability of human beings to define and distinguish between the fair, rightful and good options and the invidious, iniquitous and evil options. From the various discussions in moral philosophy on concepts like *morals*, *ethics*, *judgment* or *values*, we consider the following definitions:

Definition 1. Morals consists in a set of moral rules which describes the compliance of a given behavior with mores, values and usages of a group or a single person. These rules associate a good or bad value to some combinations of actions and contexts. They could be specific or universal, i.e. related or not to a period, a place, a folk, a community, etc.

Everyone knows many moral rules as "Lying is evil", "Being loyal is good" or "Cheating is bad". This kind of rules grounds our ability to distinguish between good and evil. Morals can be distinguished from law and legal systems in the sense that there is not explicit penalties, officials and written rules [15].

Moral rules are often supported and justified by some moral values (e.g. freedom, benevolence, wisdom, conformity). Psychologists, sociologists and anthropologists almost agree that moral values are central in the evaluation of actions, people and events [31].

A set of moral rules and moral values establishes a *theory of the good* which allows humans to assess the goodness or badness of a behavior and *theories of the right* which define some criteria to recognize a fair or, at least, acceptable option (also respectively named *theory of values* and *theories of right conduct* [32]). For example even if stealing can be considered as immoral (regarding a theory of the good), some philosophers agree that it is acceptable for a starving orphan to rob an apple in a supermarket (regarding a theory of the right). Humans commonly accept many situations where it is right and fair to satisfy needs or desires, even if it is not acceptable from a set of moral rules and values. The description of this conciliation is called *ethics* and, relying on some philosophers as Paul Ricoeur [28], we admit the following definition:

Definition 2. Ethics is a normative practical philosophical discipline of how humans should act and be toward the others. Ethics uses *ethical principles* to conciliate morals, desires and capacities of the agent.

Philosophers proposed various ethical principles, such as Kant's Categorical Imperative [18] or Thomas Aquinas' Doctrine of Double Effect [24], which are sets of rules that allow to distinguish an ethical option from a set of possible options. Traditionally, three major approaches are considered in the literature:

- **Virtue ethics**, where an agent is ethical if and only if he[2] acts and thinks according to some values as wisdom, bravery, justice, and so on [17].
- **Deontological ethics**, where an agent is ethical if and only if he respects obligations and permissions related to possible situations [2].
- **Consequentialist ethics**, where an agent is ethical if and only if he weighs the morality of the consequences of each choice and chooses the option which has the most moral consequences [33].

However, in some unusual situations, an ethical principle is unable to give a different valuation (a preference) between two options. Those situations, called *dilemmas*, are choices between two options, each supported by ethical reasons, given that the execution of both is not possible [22]. Each option will bring some regret. Many famous dilemmas, such as the trolley problem [12], are perceived as failures in morals or ethics or, at least, as an interesting question in the human ability to judge ethically and to provide a rational explanation of this judgment. In this article, we consider dilemma as a choice for which an ethical principle is not able to indicate the best option, regarding a given theory of good. When facing a dilemma, an agent can consider several principles in order to find a suitable solution. That is why an autonomous artificial agent must be able to understand a broad range of principles, and must be able to judge which principle leads to the most satisfying decision.

Indeed, the core of ethics is the judgment. It is the final step to make a decision and it evaluates each choices, with respect to the agent's desires, morals, abilities and ethical principles. Relying on some consensual references [1] and our previous definitions, we consider the following definition:

Definition 3. Judgment is the faculty of distinguishing the most satisfying option in a situation, regarding a set of ethical principles, for ourselves or someone else.

If an agent is facing two possible choices with both good and/or bad effect (e.g. kill or be killed), the ethical judgment allows him to make a decision in conformity with a set of ethical principles and preferences.

2.2 Ethics and autonomous agents

Considering all these notions, many frameworks have been developed in order to design autonomous agents embedded with an individual ethics. They are related to *ethics by design*, *ethics by casuistry*, *logic-based ethics* and *ethical cognitive architecture*.

Ethics by design consists in designing an ethical agent by an a priori analysis of every situation the agent may encounter and by implementing for each situation a way to avoid potential unethical behaviors. This approach can be a direct and safe implementation of rules (e.g. the military rules of engagement for an armed drone [4]). Its main drawback is the lack of explicit representation of any generic ethical concepts (as morals, ethics, etc). Moreover, it is not possible to measure a kind of similarity or distance between two ethics by design because they are not explicitly described. As a result, conceiving cooperative heterogeneous agents with different desires and principles, but without an explicit representation of ethics, is difficult and only permits the implementation of a strict deontological principle by a direct implementation of rules.

Casuistry aims first at inferring ethical rules from a large set of ethical judgment examples produced by some experts and second at using these rules to produce an ethical behavior [3]. Even if this approach offers a generic architecture for every application field, the human expertise is still necessary to describe many situations. Moreover, the agent's

[2] In this section, we consider agents in terms of philosophy, not only in terms of computer sciences.

ethical behavior is still not guaranteed (due to under- or over-learning). The agent's knowledge is still not explicitly described and ethical reasoning is made by proximity and not deduction. Consequently, cooperation between heterogeneous agents with different desires and principles is still difficult.

Logic-based ethics is the direct translation of some well-known and formally defined ethical principles (as Kant's Categorical Imperative or Thomas Aquinas' Doctrine of Double Effect) into logic programming [13, 14, 30]. The main advantage of this method is the formalization of theories of the right, even if theories of good are commonly simply considered as parameters. Consequently, it only permits to judge an option with respect to a single ethical principle.

Finally, *cognitive ethical architectures* consist in full explicit representations of each component of the agent, from the classical beliefs (information on the environment and other agents), desires (goals of the agent) and intentions (the chosen actions) to some concepts as heuristics or emotional machinery [5, 8, 9]. Even this kind of agent is able to use explicit norms and to justify its decisions, explicitly reasoning on other agents' ethics is not implemented.

2.3 Requirements for judgment in MAS

The approaches presented in the previous section propose interesting methods and models to design a single ethical autonomous agent. However in a multi-agent system, agents may need to interact and work together to share resources, exchange data or perform actions collectively. Previous approaches often consider other agents of the system as environmental elements whereas, in a collective perspective, agents need to represent, to judge and to take into account the other agents' ethics. We identify two major needs to design ethical agents in MAS:

- Agents need an explicit representation of ethics as suggested by the theory of mind. Indeed, the ethics of others can only be understood through an explicit representation of individual ethics [19]. In order to express an conciliate as many moral and ethical theories as possible, we propose both to split their representations in several parts and to use preferences on ethical principles. Thus, we propose to represent both theories of the good, split between moral values and moral rules, and theories of the right, split between ethical principles and the agents' ethical preferences. Such representations also ease the agents' configuration by non-specialists of artificial intelligence and ease the communication with other agents, including humans.

- Agents need an explicit process of ethical judgment in order to allow them both individual and collective reasoning on various theories of good and right. According to previous definitions, we consider judgment as an evaluation of the conformity of a set of actions regarding given values, moral rules, ethical principles and preferences, and we propose different kinds of judgment based on the ability to substitute the moral or the ethics of an agent by another one. Thus, we propose that agents use judgment both as a decision making process as in social choice problems [21], and as the ability to judge other agents according to their behaviors.

In the sequel, we describe the generic model we propose to enable agents to judge the ethical dimension of behaviors being themselves or the others.

3. ETHICAL JUDGMENT PROCESS

In this section we introduce our generic judgment architecture. After a short global presentation, we detail each function and explain how they operate.

3.1 Global view

As explained in Section 2.1, ethics consists in conciliating desires, morals and abilities. To take these dimensions into account, the generic ethical judgment process (EJP) use *evaluation*, *moral* and *ethical* knowledge. It is structured along *Awareness*, *Evaluation*, *Goodness* and *Rightness* processes (see components in Fig. 1). In this article, we consider it in the context of a BDI model, using also mental states such as *beliefs* and *desires*. For simplicity reasons, we only consider ethical judgment reasoning on short-term view by considering behaviors as actions. This model is only based on mental states and is not dependent on a specific architecture.

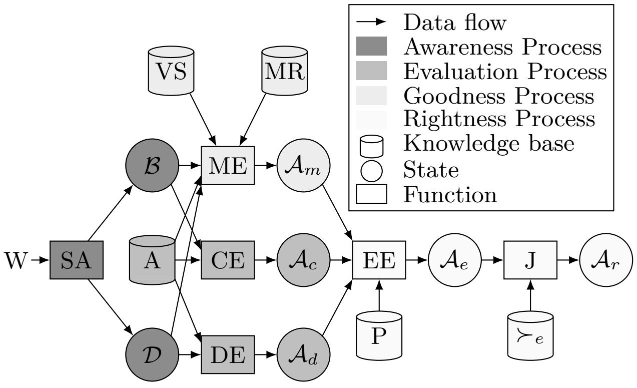

Figure 1: Ethical judgment process

Definition 4. An *ethical judgment process* EJP is defined as a composition of an Awareness Process (AP), an Evaluation Process (EP), a Goodness Process (GP), a Rightness Process (RP), an Ontology $\mathcal{O}$ ($\mathcal{O} = \mathcal{O}_v \cup \mathcal{O}_m$) of moral values ($\mathcal{O}_v$) and moral valuations ($\mathcal{O}_m$). It produces an assessment of actions from the current state of the world W with respect to moral and ethical considerations.

$$EJP = \langle AP, EP, GP, RP, \mathcal{O} \rangle$$

This model should be considered as a global scheme, composed of abstract functions, states and knowledge bases. These functions can be implemented in various ways. For instance, moral valuations from $\mathcal{O}$ may be discrete such as { good, evil } or continuous such as a degree of goodness.

3.2 Awareness and evaluation processes

In this process, agents must first assess the state of the world in terms of beliefs and desires through an awareness process.

Definition 5. The *awareness process* AP generates the set of beliefs that describes the current situation from the world

W, and the set of desires that describes the goals of the agent. It is defined as:

$$AP = \langle \mathcal{B}, \mathcal{D}, SA \rangle$$

where $\mathcal{B}$ is the set of beliefs that the agent has about W, $\mathcal{D}$ is the set of the agent's desires, and SA is a situation assessment function that updates $\mathcal{B}$ and $\mathcal{D}$ from W:

$$SA : W \rightarrow 2^{\mathcal{B} \cup \mathcal{D}}$$

From its beliefs $\mathcal{B}$ and desires $\mathcal{D}$ states, an agent executes the evaluation process EP to assess both desirable actions (i.e. actions that allow to satisfy a desire) and executable actions (i.e. actions that can be applied according to the current beliefs about the world).

Definition 6. The *evaluation process* EP produces desirable actions and executable actions from the set of beliefs and desires. It is defined as:

$$EP = \langle A, \mathcal{A}_d, \mathcal{A}_c, DE, CE \rangle$$

where A is the set of actions (each action is described as a pair of conditions and consequences bearing on beliefs and desires), $\mathcal{A}_d \subseteq A$ and $\mathcal{A}_c \subseteq A$ are respectively the sets of desirable and feasible actions, desirability evaluation DE and capability evaluation CE are functions such that:

$$DE : 2^{\mathcal{D}} \times 2^{A} \rightarrow 2^{\mathcal{A}_d}$$

$$CE : 2^{\mathcal{B}} \times 2^{A} \rightarrow 2^{\mathcal{A}_c}$$

The desirability evaluation is the ability to deduce the interesting actions to perform regarding the desires and knowledge on conditions and consequences of actions. Having defined the awareness and evaluation processes, we can turn now to the core of the judgment process that deals with the use of moral rules (resp. ethical principles) for defining the goodness process (resp. the rightness process).

3.3 Goodness Process

As seen in the state of the art, an ethical agent must assess the morality of actions given a situation assessment. To that purpose, we define the goodness process.

Definition 7. A *goodness process* GP identifies moral actions given the agent's beliefs and desires, the agent's actions and a representation of the agent's moral values and rules. It is defined as:

$$\mathrm{GP} = \langle VS, MR, \mathcal{A}_m, ME \rangle$$

where VS is the knowledge base of value supports, MR is the moral rules knowledge base, $\mathcal{A}_m \subseteq A$ is the set of moral actions[3]. The moral evaluation function ME is:

$$ME : 2^{\mathcal{D}} \times 2^{\mathcal{B}} \times 2^{A} \times 2^{VS} \times 2^{MR} \rightarrow 2^{\mathcal{A}_m}$$

In order to realize this goodness process, an agent must first be able to associate a finite set of moral values to combinations of actions and situations. The execution of the actions in these situations promotes the corresponding moral values. We consider several combinations for each moral value as, for instance, honesty could be both "avoiding telling something when it is incompatible with our own beliefs" (because it is lying) and "telling our own beliefs to someone when he believes something else" (to avoid lying by omission).

Definition 8. A *value support* is a tuple $\langle s, v \rangle \in VS$ where $v \in \mathcal{O}_v$ is a moral value, and $s = \langle a, w \rangle$ is the support of this moral value where $a \subseteq A$, $w \subset \mathcal{B} \cup \mathcal{D}$.

The precise description of a moral value relies on the language used to represent beliefs, desires and actions. For instance, from this definition, generosity supported by "giving to any poor agent" and honesty supported by "avoiding telling something when it is incompatible with our own beliefs" may be represented by:

$$\langle \langle \mathit{give}(\alpha), \{\mathit{belief}(\mathit{poor}(\alpha))\} \rangle, \mathit{generosity} \rangle$$

$$\langle \langle \mathit{tell}(\alpha, \phi), \{\mathit{belief}(\phi)\} \rangle, \mathit{honesty} \rangle$$

where α represents any agent, $\mathit{poor}(\alpha)$ (resp. ϕ) is a belief representing the context for which executing the action $\mathit{give}(\alpha)$ (resp. $\mathit{tell}(\alpha, \phi)$) supports the value $\mathit{generosity}$ (resp. $\mathit{honesty}$).

In addition to moral values, an agent must be able to represent and to manage moral rules. A moral rule describes the association of a moral valuation (for instance in a set such as {moral, amoral, immoral}) to actions or moral values in a given situation.

Definition 9. A *moral rule* is a tuple $\langle w, o, m \rangle \in MR$ where w is a situation of the current world described by $w \subset \mathcal{B} \cup \mathcal{D}$ interpreted as a conjunction of beliefs and desires, $o = \langle a, v \rangle$ where $a \in A$ and $v \in V$, and $m \in \mathcal{O}_m$ is a moral valuation described in $\mathcal{O}_m$ that qualify o when w holds.

Some rules are very common such as "killing a human is immoral" or "being honest with a liar is quite good". For instance, those rules can be represented as follows:

$$\langle \{\mathit{human}(\alpha)\}, \langle \mathit{kill}(\alpha), \_ \rangle, \mathit{immoral} \rangle$$

$$\langle \{\mathit{liar}(\alpha)\}, \langle \_, \mathit{honesty} \rangle, \mathit{quite\ good} \rangle$$

A moral rule can be more or less specific depending on the situation w or on the object o. For instance "Justice is good" is more general (having less combinations in w or o, thus applying in a larger number of situations) than "To judge a murderer, considering religion, skin, ethnic origin or political opinion is bad". Classically, moral theories are classified in three approaches (refer to Section 2.1). Using both moral values and moral rules as defined above, we can represent such theories.

- A *virtuous* approach uses general rules based on moral values (e.g. "Being generous is good"),
- A *deontological* approach classically considers specific rules concerning actions in order to describe as precisely as possible the moral behavior (e.g. "Journalists should deny favored treatment to advertisers, donors or any other special interests and resist internal and external pressure to influence coverage"[4]),
- A *consequentialist* approach uses both general and specific rules concerning states and consequences (e.g. "Every physician must refrain, even outside the exercise of his profession, any act likely to discredit it"[5]).

[3] $\mathcal{A}_m \not\subseteq \mathcal{A}_d \cup \mathcal{A}_c$ because an action might be moral by itself even if it is not desired or feasible.

[4] Extract of [27], section "Act Independently".

[5] French code of medical ethics, article 31.

3.4 Rightness process

From the sets of possible, desirable and moral actions, we can introduce the rightness process aiming at assessing the rightful actions. As shown in Section 2, an ethical agent can use several *ethical principles* to conciliate these sets of actions.

Definition 10. A *rightness process RP* produces rightful actions given a representation of the agent's ethics. It is defined as:

$$RP = \langle P, \succ_e, \mathcal{A}_r, EE, J \rangle$$

where P is a knowledge base of ethical principles, $\succ_e \subseteq P \times P$ an ethical preference relationship, $\mathcal{A}_r \subseteq A$ the set of rightful actions and two functions EE (evaluation of ethics) and J (judgment) such that :

$$EE : 2^{\mathcal{A}_d} \times 2^{\mathcal{A}_p} \times 2^{\mathcal{A}_m} \times 2^P \to 2^{\mathcal{E}}$$

where $\mathcal{E} = A \times P \times \{\bot, \top\}$

$$J : 2^{\mathcal{E}} \times 2^{\succ_e} \to 2^{\mathcal{A}_r}$$

An ethical principle is a function which represents a philosophical theory and evaluates if it is right or wrong to execute a given action in a given situation regarding this theory.

Definition 11. An *ethical principle* $p \in P$ is a function that describes the rightness of an action evaluated in terms of capabilities, desires and morality in a given situation. It is defined as:

$$p : 2^A \times 2^{\mathcal{B}} \times 2^{\mathcal{D}} \times 2^{MR} \times 2^V \to \{\top, \bot\}$$

The ethics evaluation function EE returns the evaluation of all desirable ($\mathcal{A}_d$), feasible ($\mathcal{A}_p$) and moral ($\mathcal{A}_m$) actions given the set P of known ethical principles.

For instance, let us consider three agents in the following situation inspired by the one presented by Benjamin Constant to counter the Immanuel Kant's categorical imperative. An agent A hides in an agent B's house in order to escape an agent C, and C asks B where is A to kill C, threatening to kill B in case of non-cooperation. B's moral rules are "prevents murders" and "don't lie". B's desires are to avoid any troubles with C. B knows the truth and can consider one of the possible actions: tell C the truth (satisfying a moral rule and a desire), lie or refuse to answer (both satisfying a moral rule). B knows three ethical principles (which are abstracted in P by functions):

P1 If an action is possible, motivated by at least one moral rule or desire, do it,

P2 If an action is forbidden by at least one moral rule, avoid it,

P3 Satisfy the doctrine of double effect[6].

B's evaluation of ethics return the tuples given in Table 1 where each row represents an action and each column an ethical principle.

Principle Action	P1	P2	P3
tell the truth	$\top$	$\bot$	$\top$
lie	$\top$	$\bot$	$\bot$
refuse	$\top$	$\top$	$\top$

Table 1: Ethical evaluation of agent B's actions

Given a set of actions issued of the ethic evaluation function $\mathcal{E}$, the judgment J is the last step which selects the rightful action to perform, considering a set of ethical preferences (defining a partial or total order on the ethical principles).

To pursue the previous example, let us suppose that B's ethical preferences are P3 $\succ_e$ P2 $\succ_e$ P1 and J uses a tie-breaking rule based on a lexicographic order. Then "refusing to answer" is the rightful action because it satisfies P3 whereas "lying" doesn't. Even if "telling the truth" satisfies the most preferred principle, "refusing to answer" is righter because it satisfies also P2. Let us notice that judgment allows dilemma: without the tie-breaking rule both "telling the truth" and "refusing to answer" are the rightest actions.

4. ETHICAL JUDGMENT OF OTHERS

The judgment process described in the previous section is useful for an agent to judge it's own behavior, namely one action considering its own beliefs, desires and knowledge. However, it can also judge the behaviors of other agents in a more or less informed way by putting itself at their place, partially or not.

Given an EJP defined in the previous section, the states $\mathcal{B}$, $\mathcal{D}$, $\mathcal{A}_d$, $\mathcal{A}_p$, $\mathcal{E}$, $\mathcal{A}_m$ and knowledge of actions (A), goodness knowledge – *theory of good* – (MR, VS) and rightness knowledge – *theory of right* – (P, $\succ_e$) may be shared between the agents. The ontology $\mathcal{O}$ is assumed as common knowledge, even if we could consider in future works having several ontologies. The way they are shared can take many forms such as common knowledge, direct communications, inferences, and so on that are beyond the scope of this article. In any cases, we distinguish three categories of ethical judgments:

- *Blind ethical judgment* where the judgment of the judged agent is realized without any information about this agent, except a behavior,
- *Partially informed ethical judgment* where the judgment of the judged agent is realized with some information about this agent,
- *Fully informed ethical judgment* where the judgment of the judged agent is realized with a complete knowledge of the states and knowledge used within the judged agent's judgment process.

In all kinds of judgment, the judging agent reasons on its own beliefs or those of the judged one. This kind of judgment can be compared to the role of the human theory of mind [19] in the human judgment (the ability for a human to put himself in the place of another). Then, the judging agent uses its EJP and compares the resulting $\mathcal{A}_r$ and $\mathcal{A}_m$ to the behavior of the judged agent. If the action performed by the judged agent is in $\mathcal{A}_r$, it means that it is a rightful behavior, and if it is in $\mathcal{A}_m$, it means that is a moral behavior (being in both is stated as a rightful and moral behavior).

[6] Meaning doing an action only if the four following conditions are satisfied at the same time: the action in itself from its very object is good or at least indifferent; the good effect and not the evil effect are intended (and the good effect cannot be attained without the bad effect); the good effect is not produced by means of the evil effect; there is a proportionately grave reason for permitting the evil effect [24].

Both statements have to be considered with respect to the context of the situation, the theory of good and the theory of right that are used to judge. We consider that this ethical judgment is always relative to the states, knowledge bases and ontology used to execute the judgment process.

4.1 Blind ethical judgment

The first kind of judgment an agent can make is without any information about morals and ethics of the judged agent (for instance when agents are unable or do not want to communicate). Consequently, the judging agent a_j uses its own assessment of the situation ($\mathcal{B}_{a_j}$ and $\mathcal{D}_{a_j}$)[7], its own theory of good $\langle MR_{a_j}, VS_{a_j}\rangle$ and theory of right $\langle P_{a_j}, \succ_{e,a_j}\rangle$ to evaluate the behavior of the judged agent a_t. This is an *a priori* judgment and a_t is judged as not considering rightful actions, or moral actions if the action $\alpha_{a_t} \notin \mathcal{A}_{r,a_j}$ or $\alpha_{a_t} \notin \mathcal{A}_{m,a_j}$.

4.2 Partially informed ethical judgment

The second kind of judgment that an agent can do is grounded on partial information about the judged agent in case the judging agent is able to acquire parts of the knowledge of the judged agent (e.g. by perception or communication). Three partial ethical judgments can be considered knowing either (*i*) the situation (i.e. $\mathcal{B}_{a_t}$, $\mathcal{D}_{a_t}$, A_{a_t}) either (*ii*) the theory of good (i.e. $\langle VS_{a_t}, MR_{a_t}\rangle$) and A_{a_t}[8] or (*iii*) the theory of right (i.e.$\langle P_{a_t}, \succ_e, a_t\rangle$) of the judged agent.

Situation-aware ethical judgment.

Firstly, if the judging agent a_j knows the beliefs $\mathcal{B}_{a_t}$ and desires $\mathcal{D}_{a_t}$ of the judged agent a_t, a_j can put itself in the position of a_t and can judge if the action α executed by a_t belongs to $\mathcal{A}_{r,a_j}$, considering its own theories. Firstly, a_j is able to evaluate the morality of α by generating $\mathcal{A}_{m,a_t}$ from A_{a_t} and qualify the morality of a_t's behavior (i.e. if α is or not in $\mathcal{A}_{m,a_t}$). The agent a_j can go a step further by generating $\mathcal{A}_{r,a_t}$ from the generated $\mathcal{A}_{m,a_t}$ to check if α is conform to the rightness process, i.e. belongs to $\mathcal{A}_{r,a_t}$.

Theory-of-good-aware ethical judgment.

Secondly, if the judging agent is able to obtain the moral rules and values of the judged one, it is possible to evaluate the actions in a situation (shared or not), regarding these rules. From a simple moral evaluation perspective, the judging agent can compare the theories of the good by checking if moral values MV_{a_t} or moral rules MR_{a_t} are consistent with its own theory of good (i.e. the same definition as a_j's one or at least no contradiction). For a moral judgment perspective, the judging agent can evaluate the morality of a given action from the point of view of the judged one. Interestingly, this judgment allows to judge an agent that has different duties (due to a role or some special responsibilities for instance) as human being can judge a physician on the conformity between its behavior an a medical code of deontology.

Theory-of-right-aware ethical judgment.

Thirdly, let us now consider the case of a judging agent able to reason on ethical principles and preferences of other agents, considering a situation (shared or not) and a theory of good (shared or not)[9]. It allows to evaluate how the judged agent a_t conciliates its desires, moral rules and values in a situation by comparing the sets of rightful actions $\mathcal{A}_r, a_j$ and $\mathcal{A}_r, a_t$ respectively generated by the use of P_{a_j}, $\succ_{e,a_j}$ and P_{a_t}, $\succ_{e,a_t}$. For instance, if $\mathcal{A}_r, a_j = \mathcal{A}_r, a_t$ with an unshared theory of good, it shows that their theories of right produce the same conclusions in this context. This judgment can be useful for an agent to estimate how another can judge it with a given goodness process.

4.3 Fully informed judgment

Finally, a judging agent can consider both goodness and rightness process to judge another agent. This kind of judgment needs information about all the internal states and knowledge bases of the judged agent. This kind of judgment is useful to check the conformity of the behavior of another agent with the judge's information about its theories of good and right.

5. PROOF OF CONCEPT

In this section we illustrate how each part of the model presented in the previous sections works through a multi-agent system implemented in Answer Set Programming (ASP). The complete source code is downloadable on a cloud service[10]. This agent illustrates an example of ethical agent in a multi-agent system where agents have beliefs (about richness, gender, marital status and nobility), desires, and their own judgment process. They are able to give, court, tax and steal from others or simply wait. We mainly focus on an agent named `robin_hood`.

5.1 Awareness Process

In this example, the situation awareness function SA is not implemented and the beliefs are directly given in the program. The following code represents a subset of the beliefs of `robin_hood`:

```
agent(paul).                 -man(marian).
agent(prince_john)           rich(prince_john).
agent(marian).               man(prince_john).
-poor(robin_hood).           noble(prince_john).
-married(robin_hood).        poor(paul).
```

The set of desires $\mathcal{D}$ are `robin_hood`'s desires. In our implementation we consider two kinds of desires: desires to accomplish an action (`desirableAction`) and desires to produce a state (`desirableState`).

```
desirableAction(robin_hood,robin_hood,court,marian).
desirableAction(robin_hood,robin_hood,steal,A):-
  agent(A), rich(A).
desireState(prince_john,rich,prince_john).
-desireState(friar_tuck,rich,friar_tuck).
```

The two first desires concern actions: `robin_hood` desires to court `marian` and to steal from any rich agent. The next two desires concern states: `prince_john` desires to be rich, and `friar_tuck` desires to stay in poverty, regardless the action to perform.

[7] We use the subscript notation to denote the agent handling the represented set of information.

[8] In this case, A_{a_t} is necessary as, contrary to ethical principles, the moral rules can explicitely refer to specific actions.

[9] If both the situation and the theory of good are shared, it is a fully informed judgment (see 4.3).

[10] http://depositfiles.com/files/g2msnfncq

5.2 Evaluation Process

The agents' knowledge about actions A is described as labels associated to sets (possibly empty) of conditions and consequences. For instance, the action `give` is described as:

```
action(give).
condition(give,A,B):-
  agent(B), agent(A), A!=B, not poor(A).
consequence(give,A,B,rich,B):- agent(A), agent(B).
consequence(give,A,B,poor,A):- agent(A), agent(B).
```

A condition is a conjunction of beliefs (here the fact that `A` is not poor). The consequence of an action is a clause composed of the new belief generated by the action and the agent concerned by this consequence. The desirability evaluation DE (see Definition 6) deduces the set of actions $\mathcal{A}_d$. An action is in A_d if it was directly desired (in $\mathcal{D}$) or if its consequences are a desired state:

```
desirableAction(A, B, X, C):-
  desireState(A,S,D), consequence(X,B,C,S,D).
```

The capability evaluation CE (see Definition 6) evaluates from beliefs and conditions the set of actions $\mathcal{A}_c$. An action is possible if its conditions are satisfied.

```
possibleAction(A,X,B):- condition(X,A,B).
```

5.3 Goodness Process

In the goodness process, value supports VS are implemented as (for instance):

```
 generous(A,give,B) :- A != B, agent(A), agent(B).
-generous(A,steal,B):- A != B, agent(A), agent(B).
-generous(A,tax,B)  :- A != B, agent(A), agent(B).
```

Then, we can express the agents' moral rules for each ethical approaches (see Sections 2.1 and 3.3). An example of moral rule in a virtuous approach is:

```
moral(robin_hood,A,X,B):-
  generous(A,X,B), poor(B), action(X).
```

The morality evaluation ME gives the set of moral actions $\mathcal{A}_m$ (see Section 3.3):

```
moralAction(A,X,B):- moral(A,A,X,B).
-moralAction(A,X,B):- -moral(A,A,X,B).
```

and produces as results:

```
moralAction(robin_hood,give,paul)
-moralAction(robin_hood,tax,paul)
```

In this example, we only present a virtuous approach. However, examples of deontological and consequentialist approaches are also given in our downloadable code.

5.4 Rightness Process

In order to evaluate each action, we define several naive ethical principles that illustrate priorities between moral and desirable actions. For instance, here is the `perfAct` (for perfect, i.e. a moral, desirable and possible action) principle:

```
ethPrinciple(perfAct,A,X,B):-
  possibleAction(A,X,B),
  desirableAction(A,A,X,B),
  not -desirableAction(A,A,X,B),
  moralAction(A,X,B),
  not -moralAction(A,X,B).
```

Principle Intention	perfAct	dutNR	desNR	dutFst	nR	desFst
give,paul	$\bot$	$\top$	$\bot$	$\top$	$\top$	$\bot$
give,little_john	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
give,marian	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
give,prince_john	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
give,peter	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
steal,little_john	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
steal,marian	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
steal,prince_john	$\bot$	$\bot$	$\top$	$\bot$	$\top$	$\top$
steal,peter	$\bot$	$\bot$	$\top$	$\bot$	$\top$	$\top$
court,marian	$\bot$	$\bot$	$\top$	$\bot$	$\top$	$\top$
wait,robin_hood	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$

Figure 2: Ethical evaluation $\mathcal{E}$ of the actions

If `paul` is the only poor agent, `marian` is not married and `robin_hood` is not poor, `robin_hood` obtains the evaluation given in Figure 2.

All principles are ordered with respect to `robin_hood`'s preferences:

```
prefEthics(robin_hood,perfAct,dutNR).
prefEthics(robin_hood,dutNR,desNR).
prefEthics(robin_hood,desNR,dutFst).
prefEthics(robin_hood,dutFst,nR).
prefEthics(robin_hood,nR,desFst).

prefEthics(A,X,Z):-
  prefEthics(A,X,Y), prefEthics(A,Y,Z).
```

The five first lines describe the order on the ethical principles. The last lines define transitivity for the preference relationship (here `perfAct` $\succ_e$ `dutNR` $\succ_e$ `desNR` $\succ_e$ `dutFst` $\succ_e$ `nR` $\succ_e$ `desFst`).

Finally, the judgment J is implemented as:

```
existBetter(PE1,A,X,B):-
  ethPrinciple(PE1,A,X,B),
  prefEthics(A,PE2,PE1),
  ethPrinciple(PE2,A,Y,C).

ethicalJudgment(PE1,A,X,B):-
  ethPrinciple(PE1,A,X,B),
  not existBetter(PE1,A,X,B).
```

Consequently, the rightful action a_r for `robin_hood` is `give,paul` which complies with `dutNR`.

5.5 Multi-agent ethical judgment

In order to allow a blind judgment, we introduce a new belief about the behavior of another agent:

```
done(little_john,give,peter).
```

Then `robin_hood` compares its own rightful action and this belief to judge `little_john` with:

```
blindJudgment(A,ethical,B):-
  ethicalJudgment(_,A,X,C), done(B,X,C), A!=B.

blindJudgment(A,unethical,B):-
```

```
  not blindJudgment(A,ethical,B),
  agent(A), agent(B),
  done(B,_,_), A!=B.
```

In this example, the action give to peter was not in $\mathcal{A}_r$ for robin_hood. Then little_john is judged unethical by robin_hood.

For a partial-knowledge judgment, we replace a part of robin_hood's knowledges ans states by those of little_john. With the beliefs of little_john (which believes that peter is a poor agent and paul is a rich one), robin_hood judged him ethical.

Finally, for a full-knowledge judgment, we replace all the beliefs, desires and knowledge bases of the agent robin_hood by little_john's one. Then, robin_hood is able to reproduce the whole Ethical Judgment Process of little_john and compare both judgments of a same action.

6. RELATED WORKS

We adopt in this article a full rationalist approach (based on reasoning, not emotions) but some other works propose other close approaches [34, 6]. The main specificity of our work is the avoidance of any representation of emotions to be able to justify the behavior of an agent in terms of moral values, moral rules and ethical principles to ease the evaluation of its conformity with a code of deontology or any given ethics.

On the first hand, [6] is an full intuitionistic approach which evaluates plans from emotional appraisal. The values are only source of emotions, and influence the construction of plans by an anticipatory emotional appraisal. In our point of view, values and goals (desires) must be separated because agents must be able to distinguish desires from moral motivations and may explain how to conciliate them.

On the other hand, [34] is a logic-based approach of modeling moral reasoning with deontic constraints. This model is a way to implement a theory of good and is used to implement model checking of moral behavior [11]. However, ethical reasoning is only considered in [34] as meta-level reasoning and is only suggested as the adoption of a less restrictive model of behavior. In this perspective, our work precisely focuses on the need of representing theory of the right as a set of principles to address the issue of moral dilemmas.

7. CONCLUSION

In order to act collectively in conformity with a given ethics an moral, an autonomous agent needs to be able to evaluate the rightness/goodness of both its own behavior and those of the others. Based on concepts in moral philosophy, we proposed in this article a generic judgment ability for autonomous agents. This process uses explicit representations of elements such as moral values, moral rules and ethical principles. We illustrated how this model allows to compare ethics of different agents. Moreover, this ethical judgment model has been designed as a module to be plugged on existing architectures to provide an ethical layer in an existing decision process. As this judgment process can be used with information on moral values, moral rules, ethical principles and preferences which are shared by a collective of agents, our approach defines a guideline for a forthcoming definition of collective ethics.

Even if this article presents a framework to implement a given ethics and to use it to provide judgment, the model is still based on a qualitive approach. Whereas we can define several moral valuations, there is neither degrees of desires, neither degrees of capability, nor degrees of rightfulness. Moreover, ethical principles need to be more precisely defined in order to capture the various set of theories suggested by philosophers.

Thus, our future works will be first directed towards exploring various uses of this ethical judgment through the implementation of existing codes of conduct (e.g. medical and financial deontologies) in order to assess the genericity of our approach. Secondly, we intend to extend our model to quantitative evaluations in order to assess how far from rightfulness or goodness a behavior is. Indeed, such extension would be useful to define a degree of similarity between two morals or two ethics to facilitate the distinction between different ethics from an agent perspective.

Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-13-CORD-0006.

REFERENCES

[1] Ethical judgment. Free Online Psychology Dictionary, August 2015.

[2] L. Alexander and M. Moore. Deontological ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring edition, 2015.

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

[4] R. Arkin. *Governing lethal behavior in autonomous robots*. CRC Press, 2009.

[5] K. Arkoudas, S. Bringsjord, and P. Bello. Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, pages 17–23, 2005.

[6] C. Battaglino, R. Damiano, and L. Lesmo. Emotional range in value-sensitive deliberation. In *12th International Conference on Autonomous agents and multi-agent systems*, pages 769–776, 2013.

[7] G. Boella, G. Pigozzi, and L. van der Torre. Normative systems in computer science - Ten guidelines for normative multiagent systems. In *Normative Multi-Agent Systems*, Dagstuhl Seminar Proceedings, 2009.

[8] H. Coelho and A.C. da Rocha Costa. On the intelligence of moral agency. *Encontro Português de Inteligência Artificial*, pages 12–15, October 2009.

[9] H. Coelho, P. Trigo, and A.C. da Rocha Costa. On the operationality of moral-sense decision making. In *2nd Brazilian Workshop on Social Simulation*, pages 15–20, 2010.

[10] A. Damasio. *Descartes' error: Emotion, reason and the human brain*. Random House, 2008.

[11] L.A. Dennis, M. Fisher, and A.F.T. Winfield. Towards verifiably ethical robot behaviour. In *1st International Workshop on AI and Ethics*, 2015.

[12] P. Foot. The problem of abortion and the doctrine of the double effect. *Oxford Review*, pages 5–15, 1967.

[13] J.-G. Ganascia. Ethical system formalization using

non-monotonic logics. In *29th Annual Conference of the Cognitive Science Society*, pages 1013–1018, 2007.

[14] J.-G. Ganascia. Modelling ethical rules of lying with Answer Set Programming. *Ethics and information technology*, 9(1):39–47, 2007.

[15] B. Gert. The definition of morality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall edition, 2015.

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

[17] R. Hursthouse. Virtue ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall edition, 2013.

[18] R. Johnson. Kant's moral philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer edition, 2014.

[19] K.-J. Kim and H. Lipson. Towards a theory of mind in simulated robots. In *11th Annual Conference Companion on Genetic and Evolutionary Computation Conference*, pages 2071–2076, 2009.

[20] P. Lin, K. Abney, and G.A. Bekey. *Robot ethics: the ethical and social implications of robotics*. MIT press, 2011.

[21] W. Mao and J. Gratch. Modeling social causality and responsibility judgment in multi-agent interactions. In *23rd International Joint Conference on Artificial Intelligence*, pages 3166–3170, 2013.

[22] T. McConnell. Moral dilemmas. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall edition, 2014.

[23] D. McDermott. Why ethics is a high hurdle for AI. In *North American Conference on Computing and Philosophy*, 2008.

[24] A. McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter edition, 2014.

[25] B.M. McLaren. Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems*, 21(4):29–37, 2006.

[26] J.M. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

[27] Society of Professional Journalists. Code of ethics, September 2014.

[28] P. Ricoeur. *Oneself as another*. University of Chicago Press, 1995.

[29] S. Russell, D. Dewey, M. Tegmar, A. Aguirre, E. Brynjolfsson, R. Calo, T. Dietterich, D. George, B. Hibbard, D. Hassabis, et al. Research priorities for robust and beneficial artificial intelligence. 2015. avaible on futureoflife.org/data/documents/.

[30] A. Saptawijaya and L. Moniz L. Moniz Pereira. Towards modeling morality computationally with logic programming. In *Practical Aspects of Declarative Languages*, pages 104–119. 2014.

[31] S.H. Schwartz. Basic human values: Theory, measurement, and applications. *Revue française de sociologie*, 47(4):249–288, 2006.

[32] M. Timmons. *Moral theory: an introduction*. Rowman & Littlefield Publishers, 2012.

[33] S.A. Walter. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter edition, 2015.

[34] V. Wiegel and J. van den Berg. Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, 1(3):233–242, 2009.