Ethics-based Cooperation in Multi-Agent Systems

Nicolas Cointe[1], Grégory Bonnet[2], and Olivier Boissier[3]

[1] Department of Engineering Systems and Services, Delft University of Technology
[2] Université de Lyon, MINES Saint-Etienne, CNRS, Laboratoire Hubert Curien, UMR 5516
[3] Équipe MAD – GREYC UMR CNRS 6072, Université de Caen Basse-Normandie

Abstract. In the recent literature in Artificial Intelligence, ethical issues are increasingly discussed. Many proposals of ethical agents are made. However, those approaches consider mainly an agent-centered perspective, letting aside the collective dimension of multi-agent systems. For instance, when considering cooperation among such agents, ethics could be a key issue to drive the interactions among the agents. This paper presents a model for ethics-based cooperation. Each agent uses an ethical judgment process to compute images of the other agents' ethical behavior. Based on a rationalist and explicit approach, the judgment process distinguishes a theory of good, namely how values and moral rules are defined, and a theory of right, namely how a behavior is judged with respect to ethical principles. From these images of the other agents' ethics, the judging agent computes trust used to cooperate with the judged agents. We illustrate these functionalities in an asset management scenario with a proof-of-concept implemented in the JaCaMo Multi-Agent Platform.

Keywords: Computational Ethics, Ethical Judgment, Agent Cooperation

1 Introduction

The increasing use of autonomous agents in various fields as health care, financial markets, transportation and so on, raises many issues. Besides achieving goals optimally, the ethical and moral dimensions of an agent's decisions should be considered in their reasoning. For instance, in a multi-agent based asset management, many models are available to evaluate the potential profit of an investment. Nevertheless it is still not possible for an agent to judge the morality and ethics of this investment, taking into account the moral values and ethical principles of the investor. Moreover, the heterogeneity of these elements raises many issues when agents need to collaborate with other agents while respecting their own ethics.

For instance, given its own ethics, how could an agent computes its ethical conformity with other agents from their observed behaviors? How could this agent decides to trust agents based on these images? The goal of this paper is to answer such questions. To this aim, a model is proposed to base cooperation among agents on trust built from the use of ethical judgments on actions of the others.

At first the ethical judgment process proposed in [14] is incorporated in a BDI architecture to enable agents to judge the other agents' behaviors. Using this process, a mechanisms is used to build an image of the other agents by computing and aggregating evaluations of these judgments. Then, a mechanism allows an agent to decide on trusting the other agents for delegation or cooperation. Finally, the components of this

proposal are instantiated to the asset management domain and demonstrate its use in an application implemented in the JaCaMo multi-agent oriented framework [7].

The paper is organized as follows. Sec. 2 introduces and describes the models of computational ethics and trust used in this article. Then, Sec. 3 shows how ethical judgment can be used to depict in terms of ethics the behavior of the others. Then, Sec. 4 presents the construction and use of trust. Finally the use of these contributions is illustrated in a proof of concept in Sec. 5 before concluding.

2 Foundations

We introduce in this section the necessary concepts to deal with ethics-based cooperation in multi-agent systems. Sec. 2.1 presents the concept of *trust* as a way to ground interaction and cooperation among agents. Sec. 2.2 introduces ethics and shows how it is related to trust. Finally, Sec. 2.3 synthesizes the requirements for defining an ethical-based cooperation in a multi-agent systems.

2.1 Trust in multi-agent systems

In decentralized and open systems, a way to deal with unreliable or unknown agents is to use trust [11,16,31]. Trust allows the agents to assess the interactions they observe or they make in order to decide if interacting with a given agent is a priori acceptable. This acceptance notion means that the investigated agent behaves well and is reliable according to the investigator criteria.

Many definitions of trust exist but, in accordance with [11], we consider *trust* as *a disposition to cooperate with a trustee*. Here, trust is an action that might be motivated by desires, depending on the context. It can be used as a condition to perform actions as delegating actions, sharing resources and information, or any kind of cooperation. To build trust, the agents first build an image of the investigated agents [16].

An *image* is *an evaluative belief that tells whether the target is good or bad with respect to a given behavior*. In the literature, images are aggregated from the experiences, i.e. the observed behavior of the target agent and its consequences. We can distinguish two kinds of approaches:

- statistical images [1,9,17,25,35] where the image is a quantitative aggregation of feedbacks about interactions. This aggregation estimates the trends of an agent to behave well from another agent's point-of-view. It can be represented by Bayesian networks, Beta density functions, fuzzy sets, Dempster-Shafer functions and other quantitative formalisms.
- logical images [10,11,27,34] where the image is a mental state rooted in every cooperation action that is produced by interactions. A persistent image allows to infer trust beliefs that can be used as preconditions to cooperate.

An agent can lack of observations and interactions in order to build a correct image of a target. A way to deal with this issue is to use reputation [23,30]. It consists in using third party agents' image of the target (that can depend on the initial agent's image has about the third parties) in order to assess a collective point-of-view about the target. Both images and reputations are used to lead to a trust action [31]. Most of the time, trust is dynamic and changes with respect to the evolution of images and reputations.

2.2 Ethical behaviors

This paper mainly focuses on building images of the ethical behavior of other agents. Moral theories are based on two components [33]: theory of the good (or morals) and theory of the right (or ethics). Due to the lack of formal definitions of the components of morals and ethics in the litterature, we admit the consensual definitions provided in this section. Even if still debatable due to the wide diversity of existing contradictory theories in philosophy and psychology, we consider that they offer a sound framework to define moral and ethics.

- A *theory of the good* is a set of moral rules and values which allow to assess the goodness or badness of an action itself. Moral rules give moral valuations to behaviors (e.g. "Lying is evil" or "Being honest is good"), and values give them more abstract qualification (e.g. "Telling what we believe is being honest").
- A *theory of the right* uses a set of ethical principles to recognize a fair or, at least, acceptable option in comparison with the other available actions in a given situation. Philosophers proposed various ethical principles, such as Kant's Categorical Imperative [22] or Thomas Aquinas' Doctrine of Double Effect [26]. For example even if stealing can be considered as immoral (regarding Divine Commands), some philosophers agree that it is acceptable for starving people to rob food (regarding Doctrine of Double Effect).

As a moral behavior is based on a theory of the good, an ethical behavior uses a theory of the right to conciliate morals, desires and capacities of the agent [28]. Interestingly, being moral or ethical is a behavior characterization such as being reliable is in trust systems. Consequently, it can be interesting to define a trust notion based on moral or ethical behaviors, which can enhance cooperation.

2.3 Requirement for ethical based cooperation

Works dealing with ethical behaviors in autonomous agents often focus on modelling moral reasoning [6,19,20,32] as a direct translation of some well-known moral theories, or on modelling moral agency in a general way [5,24]. However, those work do not clearly make the distinction between theory of the good and theory of the right. Some other works deal with ethical agent architecture. In the litterature, we find *implicit ethical architectures* [3,4] which design the agent's behavior either by implementing for each situation a way to avoid potential unethical behaviors, or by learning from human expertise. We also find *cognitive ethical architectures* [12,13,14,15] which 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. However, all those approaches do not take into account the collective dimension of agent systems, apart [29] which considers morals as part of agent societies.

More precisely, the architecture given in [14] makes a clear separation between theory of the good and theory of the right, and provides beliefs on various components of moral theories (moral rules, values or ethical principles for instance). Moreover, the

architecture given in [29] allows – but without operationalization – moral facts (judgments over other agents or blames for instance) to be viewed as beliefs that can be used in the agents' decisions.

In order to build ethics-based cooperation, we need an operational model of ethical judgment such as proposed in [14]. Inspired by [29], we reuse and extend this model with beliefs on moral and ethical images of other agents. We use those image beliefs to build trust beliefs to drive a cooperation based on morals or ethics.

3 Judgments building

This section explains how a BDI agent can use the judgment process introduced in [14] (see Sec. 3.1) and presents how an agent computes its own qualitative representation of the ethics (see Sec. 3.2) and morals (see Sec. 3.3) of the other agents, regarding the judging agent's *goodness knowledge* (i.e. knowledge on morals) and *rightness knowledge* (i.e. knowledge on ethics).

3.1 Judging other agents

Let us consider the judgment process introduced in [14]. Adapting it to our needs as expressed in Sec. 2, the generic reasoning done in the ethical judgment process generates the set of rightful actions for a given situation, regarding a set of knowledge.

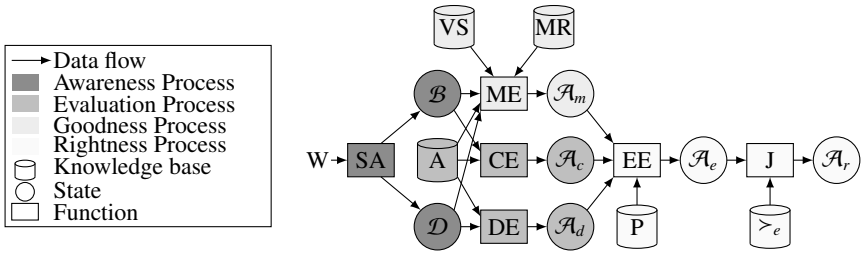

Fig. 1. Ethical judgment process as depicted in [14]

As depicted in Fig. 1, the judgment process is organized into three parts: (*i*) awareness and evaluation process, (*ii*) goodness process and (*iii*) rightness process. Since this judgment process may use sets of knowledge issued from another agent, we index all these sets with an agent id $a_i \in \mathbb{A}$ (e.g. $\mathcal{A}_{r_\mathsf{a}}$) with $\mathbb{A}$ the set of the agents. When a is the agent executing the process, this agent is using the process to decide about its own behavior, when a is different, the agent uses this process to judge the behavior of a.

Awareness and evaluation processes The *evaluation* process evaluates the set of actions $\mathcal{A}_\mathsf{a}$ (actions are pairs of conditions and consequences bearing on desires and beliefs) that it considered both desirable ($\mathcal{A}_{d_\mathsf{a}}$) and executable ($\mathcal{A}_{c_\mathsf{a}}$) from a's point-of-view, with respect to $\mathcal{D}_\mathsf{a}$ the set of desires and $\mathcal{B}_\mathsf{a}$ the set of beliefs of a. $\mathcal{B}_\mathsf{a}$ and $\mathcal{D}_\mathsf{a}$ are produced by the situation assessment SA of the current state. Here, DE and CE are respectively *desirability evaluation* and *capability evaluation* functions. In the sequel, we call *contextual knowledge* of a (CK_a), the union of $\mathcal{B}_\mathsf{a}$ and $\mathcal{D}_\mathsf{a}$.

Goodness Process The *goodness process* identifies moral actions $\mathcal{A}_{m\mathsf{a}}$ given a's contextual knowledge CK_{a}, actions A_{a}, value supports VS_{a} and moral rules MR_{a}. Moral actions are actions that, in the situations of CK_{a}, promote or demote the moral values of VS_{a}. A *value support* is a tuple $\langle s, v\rangle \in VS_{\mathsf{a}}$ where $v \in O_v$ is a moral value and $s = \langle \alpha, w\rangle$ is the support of this moral value where $\alpha \in A_{\mathsf{a}}$, $w \subset \mathcal{B}_{ai} \cup \mathcal{D}_{\mathsf{a}}$. O_v is the set of moral values used in the system[4]. A *moral rule* is a tuple $\langle w, o, m\rangle \in MR_{\mathsf{a}}$. The situation $w \in 2^{CK_{\mathsf{a}}}$ is a conjunction of beliefs and desires. The object $o = \langle \alpha, v\rangle$ where α is an action ($\alpha \in A_{\mathsf{a}}$) and v is a moral value ($v \in O_v$). Finally, m is the moral valuation ($m \in O_m$). For instance with $O_m = \{\texttt{moral}, \texttt{amoral}, \texttt{immoral}\}$ provides three moral valuation for o when w holds. It is important to notice that a *total order* is defined on O_m (e.g. `moral` is a higher moral valuation than `amoral`, which is higher than `immoral`). In the sequel, moral rules MR_{a}, value support VS_{a} and values O_v, knowledge used in the goodness process of the agent a are referred as the *goodness knowledge* (GK_{a}).

Rightness process Finally, the rightness process assess the rightful action $\mathcal{A}_{r\mathsf{a}}$ from the sets of possible $\mathcal{A}_{c\mathsf{a}}$, desirable $\mathcal{A}_{d\mathsf{a}}$ and moral $\mathcal{A}_{m\mathsf{a}}$ actions based on *ethical principles* P_{a} to conciliate these sets of actions according to ethical preference relationship $\succ_{e\mathsf{a}} \subseteq P_{\mathsf{a}} \times P_{\mathsf{a}}$. An *ethical principle* $p \in P_{\mathsf{a}}$ is a function which evaluates if it is *right* or *wrong* to execute a given action in a given situation regarding a philosophical theory. It describes the rightness of an action with respect to its belonging to $\mathcal{A}_{c\mathsf{a}}$, $\mathcal{A}_{d\mathsf{a}}$ and $\mathcal{A}_{m\mathsf{a}}$ in a given situation of CK_{a}. It is defined as $p : 2^{\mathcal{A}_{\mathsf{a}}} \times 2^{\mathcal{B}_{\mathsf{a}}} \times 2^{\mathcal{D}_{\mathsf{a}}} \times 2^{MR_{\mathsf{a}}} \times 2^{V_{\mathsf{a}}} \rightarrow \{\top, \bot\}$. Given a set of actions issued of the ethic evaluation function EE that applies the ethical principles, the judgment J is the last step which selects the set of rightful actions to perform, considering the set of ethical preferences $\succ_{e\mathsf{a}}$ defining a total order on the ethical principles. In this judgment process, the rightful actions are the ones that satisfy the most preferred principles in a lexicographic order. In the sequel, ethical principles P_{a} and preferences $\succ_{e\mathsf{a}}$ are referred as the rightness knowledge (RK_{a}).

3.2 Judging ethical conformity of behaviors

We extend now the previous judgment process to judge the ethics and morality of the behavior between t_0 and t of an agent a'. Inspired from [29] which considers beliefs on moral facts, the judgment process produces now beliefs (`ethical_conformity`, `moral_conformity`) stating the conformity to ethical principles or moral rules and values, that can be used in the agent's reasoning. Before defining these beliefs, let us define first an agent's behavior as follows:

Definition 1 (Behavior). *The* behavior $b_{\mathsf{a}',[t_0,t]}$ *of an agent* a' *on the time interval* $[t_0, t]$ *is the set of actions* α_k *that* a' *executed between* t_0 *and* t *as* $0 \leqslant t_0 \leqslant t$.

$$b_{\mathsf{a}',[t_0,t]} = \{\alpha_k \in A : \exists t' \in [t_0, t] \text{ s.t. } \mathsf{done}(\mathsf{a}', \alpha_k, t')\}$$

where $A = \bigcup_{\mathsf{a}_i=\mathsf{a}_1}^{\mathsf{a}_n} \mathcal{A}_{\mathsf{a}_i}$ *is the set of available actions in the multi-agent system composed of n agents, and* $\mathsf{done}(\mathsf{a}', \alpha_k, t')$ *means that* α_k *has been executed*[5] *by* a' *at time* t'.

[4] Let us notice that in [14] moral values and moral valuation are shared in the system. Agents distinguish themselves by moral rules and rightness processes.

[5] A behavior can deal with concurrency: several actions can have been done at the same time.

An agent a can judge the conformity of an action α_k executed by another agent a' with respect to its own goodness and rightness knowledge.

Definition 2 (Ethical conformity). *An action α_k is said to be* ethically conform *with respect to the judging agent a's contextual knowledge (CK_a), goodness knowledge GK_a and rightness knowledge RK_a at time t', noted:*

$$\texttt{ethical\_conformity}(\alpha_k, t')$$

iff α_k is in the set of rightful actions $\alpha_k \in \mathcal{A}_{r_\mathsf{a}}$ computed by the ethical judgment J_a of the judging agent a, based on $[CK_\mathsf{a}, GK_\mathsf{a}, RK_\mathsf{a}]$ at time t'.

Let us notice that the ethical conformity of an action can be applied to actions of the judging agent or to actions executed by another agent and observed by the judging agent. This ethical conformity can be judged with respect to the judging agent's contextual, goodness and rightness knowledge. It can be judged also with respect to the rightness or goodness knowledge from another agent as long as the judging agent has a representation of these knowledge. Finally, the ethical conformity is used to compute the set EC^+ of ethically conform (resp. the set EC^- of non ethically conform) actions of the observed behavior $b_{\mathsf{a}',[t_0,t]}$ of the judged agent a' between t_0 and t:

$EC^+_{b_{\mathsf{a}',[t_0,t]}}$=$\{\alpha_k \in b_{\mathsf{a}',[t_0,t]} \wedge t' \in [t_0, t]$ s.t. $\texttt{done}(\mathsf{a}', \alpha_k, t') \wedge \texttt{ethical\_conformity}(\alpha_k, t')\}$
$EC^-_{b_{\mathsf{a}',[t_0,t]}}$=$\{\alpha_k \in b_{\mathsf{a}',[t_0,t]} \wedge t' \in [t_0, t]$ s.t. $\texttt{done}(\mathsf{a}', \alpha_k, t') \wedge \neg\texttt{ethical\_conformity}(\alpha_k, t')\}$

These two sets provide information on the behavior of the judged agent and its compliance with the ethics of the judging agent. Nevertheless, it cannot assess why an observed behavior is judged as unethical. Indeed, the reason can be a difference between the judging and the judged agents' theory of the right, theory of the good or the assessment of the situation. In the sequel, we will denote: $EC_{b_{\mathsf{a}',[t_0,t]}} = EC^+_{b_{\mathsf{a}',[t_0,t]}} \cup EC^-_{b_{\mathsf{a}',[t_0,t]}}$

3.3 Judging moral conformity of behaviors

The moral conformity of an action with respect to a given moral rule is realized regarding a moral threshold $mt \in MV$ and a situation assessment

Definition 3 (Moral conformity). *An action α_k is said to be* morally conform *at time t' with respect to the judging agent a's contextual knowledge CK_a and goodness knowledge GK_a, considering the moral rule $mr \in MR_\mathsf{a}$, moral threshold $mt \in MV_\mathsf{a}$, noted:*

$$\texttt{moral\_conformity}(\alpha_k, mr, mt, t')$$

iff α_k belongs to $\mathcal{A}_{m_\mathsf{a}}$ with a moral valuation greater or equal to mt, given the considered moral rule mr, CK_a and GK_a at time t'.

Similarly to the ethical conformity, we use the moral conformity of an action to compute the set MC^+ (resp. MC^-) of morally conform (resp. non morally conform) actions of the observed behavior $b_{\mathsf{a}',[t_0,t]}$ of a' during $[t_0, t]$ with respect to mr and mt :

$$MC^{+}_{b_{\mathsf{a'},[t_0,t]},mr,mt}=\{a_k \in b_{\mathsf{a'},[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathsf{a'},a_k,t') \wedge moral\_conformity(a_k,mr,mt,t')\}$$

$$MC^{-}_{b_{\mathsf{a'},[t_0,t]},mr,mt}=\{a_k \in b_{\mathsf{a'},[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathsf{a'},a_k,t') \wedge \neg moral\_conformity(a_k,mr,mt,t')\}$$

We can generalize the above evaluation of the moral conformity with respect to a moral rule to a set of moral rules, considering the possibility to define a subset ms of moral rules $ms \subseteq MR_{\mathsf{a}}$. Such a set ms represents a cluster of rules such as rules based on some moral values, rules concerned by particular situations, and so on. In the sequel, we denote: $MC_{b_{\mathsf{a'}},ms,mt,[t_0,t]} = MC^{+}_{b_{\mathsf{a'}},ms,mt,[t_0,t]} \cup MC^{-}_{b_{\mathsf{a'}},ms,mt,[t_0,t]}$

4 Trust within ethical behavior

In this section, the conformity beliefs defined in the previous sections is used to compute the images of other agents (see Sec. 4.1). We then introduce how we use these images to build trust (cf. Sec. 4.2). Sec. 4.3 provides hints about how to use it.

4.1 Ethical and Moral images of an agent

Following Sec. 2.1, the *ethical* and *moral* images of an agent are evaluative beliefs that tell whether another agent has a conform behavior or not with respect to a given rightness (RK) and goodness (GK) knowledge.

Definition 4 (Ethical Image (resp. Moral Image)). *An* ethical image *(resp.* moral image*) of an agent* $\mathsf{a'}$[6] *is the judgment of the behavior* $b_{\mathsf{a'},[t_0,t]}$ *of that agent in a situation with respect to an ethics (resp. to set of moral rules ms and a moral threshold mt), regarding the contextual CK, goodness GK and rightness RK knowledge of another agent* a*. This image states a conformity valuation* $cv \in CV$*, where* CV *is an ordered set of* conformity valuation[7]*. They are noted as* `ethical_image`$(\mathsf{a'},\mathsf{a},cv,t_0,t)$ *and* `morality_image`$(\mathsf{a'},\mathsf{a},cv,ms,mt,t_0,t)$

Indeed, while an agent can only have a single ethical image of other agents, it can have several moral images of the same agents depending on the chosen ms and mt.

To build these images, an agent a uses two aggregation functions *ethicAggregation* and *moralAggregation* applied respectively on evaluated actions regarding ethics $EC_{b_{\mathsf{a'},[t_0,t]}}$ and regarding moral $MC_{b_{\mathsf{a'},[t_0,t]}}$. Both aggregation functions compute the ratio of the weighted sum of positive evaluations with respect to ethics and with respect to morals. The weight of each action corresponds to a criterion (e.g. the time past from the date of the evaluation, the consequences of the action and so on).

[6] Let's notice that in the definition of these images, the second parameter refers to an agent. It means that the image is built with respect to the knowledge of this agent. The first parameter refers to the considered agent's behavior.

[7] As for morals, conformity valuations are for instance { `improper`, `neutral`, `congruent` }.

Definition 5 (Ethical aggregation function). *ethicAggregation* : $2^{\mathcal{A}} \rightarrow [0,1]$ *such that* $ethicAggregation(EC_{b_{\mathsf{a}'},[t_0,t]}) = \sum_{\alpha_k \in EC^{+}_{b_{\mathsf{a}'},[t_0,t]}} weight(\alpha_k) / \sum_{\alpha_k \in EC_{b_{\mathsf{a}'},[t_0,t]}} weight(\alpha_k)$

Definition 6 (Moral aggregation function). *moralAggregation* : $2^{\mathcal{A}} \rightarrow [0,1]$ *such that* $moralAggregation(MC_{b_{\mathsf{a}'},[t_0,t]}) = \sum_{\alpha_k \in MC^{+}_{b_{\mathsf{a}'},[t_0,t]}} weight(\alpha_k) / \sum_{\alpha_k \in MC_{b_{\mathsf{a}'},[t_0,t]}} weight(\alpha_k)$

In order to transform the quantitative evaluation into a qualitative one, every conformity valuation is associated to an interval in the range of the ethical and moral aggregation functions. Once the conformity valuation computed, the associated beliefs $\texttt{moral\_image}(\mathsf{a}', \mathsf{a}, ms, mt, cv, t_0, t)$ or $\texttt{ethical\_image}(\mathsf{a}', \mathsf{a}, cv, t_0, t)$ are produced. For instance, if `congruent` conformity evaluation is defined in $[0.75, 1]$, the behavior of an agent is considered as ethical if $ethicAggregation \geq 0.75$. Finally, those images can be used to influence interactions by building trust relationships, or to describe the morality of interactions, depending on the behavior of the others.

4.2 Building trust beliefs

According to the information on the moral and ethical images, an agent can decide to trust others or not. Trust can be absolute (trust in the rightness of the others' behavior) or relative to a set of moral rules (trust in their responsibility, carefulness, obedience to some sets of rules, and so on). We define two internal epistemic actions, with respect to ethical and moral images respectively, that build beliefs on trust.

Definition 7 (Trust function). *The* ethical trust function TB^{e}_{a} *(resp.* moral trust function TB^{m}_{a}*) is defined as:* $TB^{e}_{\mathsf{a}} : \mathbb{A} \rightarrow \{\top, \bot\}$ *(resp.* $TB^{m}_{\mathsf{a}} : \mathbb{A} \times 2^{\mathcal{MR}_{\mathsf{a}}} \times MV_{\mathsf{a}} \rightarrow \{\top, \bot\}$*)*

Here, those trust functions are abstract and must be instantiated. In example, when an agent a computes that the behavior of another agent a' is conform with $CK_{\mathsf{a}}, GK_{\mathsf{a}}$ and RK_{a} (i.e. the ethical image), the ethical trust function produces a belief $\texttt{ethical\_trust}(\mathsf{a}', \mathsf{a})$. Similarly, when the agent a computes that a''s behavior is conform with ms (i.e. the moral image of its behavior regarding ms is at least mt), the moral trust function produces a belief $\texttt{moral\_trust}(\mathsf{a}', \mathsf{a}, ms, mt)$.

4.3 Ethical trusting

Beliefs on images and trust can be be used as a part of the context to evaluate the morality and ethics of an action. To this end, we can express that the morality of an action that affect other agents depends on their image.

Firstly, ethical and moral trust can enrich the description of the moral rules or values. It is useful to represent that the others' behavior can have an impact on how a context is qualified. For instance, the *responsibility* value may be supported by delegating actions to ethically trusted agents only. Here, responsibility is defined as the capability to act safely with the appropriate agents. We can also explicitly express it is not responsible to delegate something to an agent known for its unethical behavior.

Secondly, specific moral trust beliefs can be used as elements of moral rules. For instance, assuming a *honesty* moral value and its value supports, an agent can express the

moral rule "It is immoral to not behave honestly towards an agent who is trusted as being honest". Here, "who is trusted as being honest" can be modeled by a `moral_trust` belief where the associated moral rules *ms* are all rules that refer to honesty.

Finally, as evaluating and judging others are actions, it is also possible to evaluate their morality or ethics. For instance, *tolerance* as a moral value might be supported by building an image on the others with a low moral threshold until the sets $EC_{\mathsf{a}',[t_0,t]}$ or $MC_{\mathsf{a}',[t_0,t]}$ are significant enough. The choice of the thresholds, the weights and the conversion of the aggregation into a conformity valuation can also be a way to represent various types of trust. As another example, *forgiveness* can a value supporting high weights on the most recent observations. It can allow then to specificy an ethics of trust as "It is immoral to build trust without tolerance and forgiveness" [21].

5 Proof of concept

This section illustrates how the elements presented in the previous sections have been implemented in a multi-agent system. We use the JaCaMo platform [7] where the agents are programmed in BDI architecture using the Jason Language and the shared environment is programmed with workspaces and artifacts from the Cartago Platform. The complete source code is available on our website[8]. The environment is a simulated asset market where assets are quoted, bought and sold by autonomous agents. Section 5.1 introduces ethical asset management and the features of our application. Morals and ethics are defined in Sec. 5.2. Images and trust building are shown in Sec. 5.3.

5.1 Asset market modeling

Trading assets leads to several practical and ethical issues[9]. This is all the more important in automated trading as decisions, made by autonomous agents to whom human users delegate the power to sell and buy assets, have consequences in real life [18]. As shown by [8], some investment funds are interested to make socially responsible and ethical trading, and they are growing and taking a significant position on the market. However, whereas the performance of such funds can be measured objectively, their ethical quality is more difficult to assess as it depends on the values of the observer.

In this proof-of-concept, we consider a market where autonomous trading agents can manage portfolios in order to sell or buy assets. Assets types are currencies – i.e. money – and equity securities – i.e. part of a company's capital stock. A market is represented as a tuple ⟨ `name`, `id`, `type`, `matching` ⟩ with the name of the market `name`, a unique identifier `id`, the type of exchanged assets `type` and the algorithm used to store and execute orders `matching`. On the market, each agent can execute `buy`, `sell` or `cancel` orders. They respectively correspond in exchanging an equity for a currency, exchanging a currency for an equity, and canceling an exchange order that has not been executed yet. Each equity is quoted in a state-of-the-art Central Limit Order Book (CLOB) [2] algorithm.

[8] `https://cointe.users.greyc.fr/projects/ethical_market_simulator`
[9] `http://sevenpillarsinstitute.org/`

By observing the market, the agents get beliefs on the market. Agents perceive each minute the volume (the quantity of exchanged assets), two moving means, representing the average price on the last twenty minutes and on the last forty minutes, the standard deviations of prices on the last twenty minutes, the closing prices on this period, and the up and down Bollinger bands (the average prices ± twice the standard deviations). Agents have also beliefs on the orders added and stored in the CLOB and their execution. The general form of all those beliefs is respectively:

```
indicators(Date,Mktplace,Asset,Close,Volume,Intensity,Mm,Dblmm,BUp,BDown)
onMarket(Date,Agent,Portfolio,Marketplace,Side,Asset,Volume,Price)
executed(Date,Agent,Portfolio,Marketplace,Side,Asset,Volume,Price)
```

A set of beliefs `own(PortfolioName,Broker,Asset,Quantity)` updated in real time represent the agents' portfolio. By reasoning on those beliefs as a contextual knowledge CK, an agent is able to infer the feasibility of passing a buy or sell order (simply by verifying if its own portfolio contains the assets to exchange) to produce $\mathcal{A}_p$. He can also reason on the desirability of these actions to produce $\mathcal{A}_d$. To this end, we implemented a simple but classical method of trading decision-making based on comparisons between the Bollinger bands and the moving means.

Are introduced in our experiment two types of agents:

- *Zero-intelligence agents* make random orders (in terms of price and volume) on the market to generate activity and simulate the "noise" of real markets. Each of them is assigned to one or every assets.
- *Ethical agents* implements the ethical judgment on their own actions as a decision process to make their decisions. they have a simple desirability evaluation function to speculate: if the price of the market is going up (the shortest moving mean is over the other one), they buy the asset, otherwise, they sell it. If the price goes out of the Bollinger bands, these rules are inverted.

5.2 Ethical settings

We consider that the ethical agents are initialized with a particular set of beliefs about activities of the companies (e.g. an energy producer using nuclear power plants) and some labels about their conformity with international standards (e.g. an electric infrastructure producer labeled FSC). Those beliefs are important to assess how it is moral to trade a given asset based on the company's activities. Indeed, to provide information on the morality of acting on a financial market, we implemented moral values and moral rules directly inspired from the literature available online[10]. The ethical agents know a set of organized values: for instance "environmental reporting" is considered as a subvalue of "environment". Values are represented as:

```
value("environment").
subvalue("promote_renewable_energy","environment").
subvalue("envirnmt_reporting","environment").
```

Agents have a set of value supports as "trading assets of nuclear energy producer is not conform with the subvalue *promotion of renewable energy*", represented as:

[10] `http://www.ethicalconsumer.org/`

```
valueSupport(buy(Asset,_,_,_),"envirnmt_reporting"):-label(Asset,"FSC").
```

Agents are also equipped with moral rules stating the morality of environmental considerations. For instance, "It is moral to act in conformity with the value *environment*" is simply represented as:

```
moral_eval(X,V1,moral):- valueSupport(X,V1) & subvalue(V1,"environment").
moral_eval(X,"environment",moral):- valueSupport(X,"environment").
moralSet("environment","value_environment").
```

We declare in the last line this moral rule as an element of a set of moral rules related to environmental values (in order to build images). In this example, an ethical agent is able to infer for instance that, regarding its beliefs and this goodness knowledge, trading the asset of the FSC labeled company is moral while trading the asset of the nuclear energy producer is both moral and immoral. Thus, the agent needs a rightness knowledge to discriminate if it is right or wrong to trade the second assets. Finally, ethical agents are equipped with ethical principles, such as the Aristotelian ethics (inspired from [20]) and more simple principles such as considering `perfectAct` "It is rightful to do a possible, moral and desirable action", the non shaming desire `desireNR` "It is rightful to do a possible, not immoral and desirable action" and the moral duty `dutyNR` "It is rightful to do a possible, moral and not undesirable action". Please see directly the file `rightness_process.asl` for more details. Each agent can have several ethical principles, and the rightful actions to execute are the ones that satisfy the preference over the principles according to a lexicographic order.

5.3 Image and trust building

Each time an action is executed on the market (i.e. a buy order matches with a sell order) the agents receive a message and evaluate their image of the agents implied in the transaction. As said in the previous section, evaluating the conformity of behaviors, building the image and the trust beliefs are actions. Thus, they are implemented as Jason plans. In the sequel, we will detail moral trust building. Ethical trust building is based on the same ideas. The following plan evaluates the conformity of the action with each moral rule of the set `MSet` and increments the value `X` stored in the belief `moralAggr(Agent,MSet,X)`.

In this implementation, we use a linear aggregation, (i.e. it associates the same weight with each action). Then, a conformity valuation is computed regarding the proportion of conform actions in order to build the image. We use here three conformity valuation (arbitrary `neutral` for an aggregated ratio in $[0.4, 0.6[$, `improper` if lower and `congruent` if higher). Finally, when the conformity valuation crosses a trust threshold, a plan updates the trust belief in the judged agent regarding the set of moral rules.

```
+!trust : moralImageOf(Agent,MoralSet,ConformityValuation)
  & trustThreshold(Threshold) & not trust(Agent,MoralSet)
  & not tOrderOnConformityValuation(Threshold,ConformityValuation)
  <- +trust(Agent,MoralSet); !trust.
```

Similarly, we have implemented a plan for ethical conformity which stores the number of conform and non conform actions regarding the rightness knowledge, a plan for ethical image building and a plan for ethical trust building.

5.4 Results

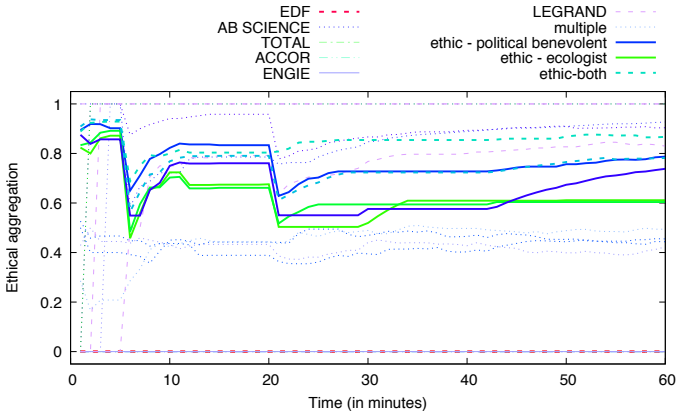

Fig. 2. Evolution of the output of an ethical aggregations functions

Fig. 2 shows the evolution of the ethical aggregations computed by an ethical agent on the others' behaviors. In this simuation, three groups of ethical agents are created with three different theories of good. Let us notice that the judge agent evaluates the behvior of both ethical an zero˙intelligence agents. Our model only evaluates the conformity of an observed behavior with an ethics, without trying to understand or reason on the intentions of the other agents. As expected, the ethical agents obeying the same ethics stay at a similar value (thick lines). The agents obliged to generate activity on such assets stay at 0.0 or 1.0 because they respectively can't do moral or evil actions regarding the judge's point of view. All the other agents slowly converge towards a value depending on their behavior. By the use of the "mind observer" provided by JaCaMo, the reader can observe the beliefs of the agents during the experiments.

6 Conclusion

In this paper, we mapped a model of ethical judgment process into a BDI agent model and defined mechanisms to build images depicting the conformity of a behavior with respect to an ethics or morals. We demonstrated how agents can use these images to decide about trusting other agents in order to cooperate and delegate actions. A proof of concept shows how this model has been implemented in a BDI framework to be used in an asset-management application. From a modeling point-of-view, this proposal addresses the problem of measuring how far from an ethics or a moral theory a behavior is, especially when ethics and morals lie in the hidden personal motivations and rules of a set of heterogeneous agents. With this model, agents may know which moral rules or values as well as ethics are concerned by this proximity. Thanks to the expressiveness of our model for ethics and morals, we envision as a future work to represent a large set of moral values that take the images of the others into account in their descriptions, such as forgiveness or intransigence.

Acknowledgment

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 724431).

References

1. A. Abdul-Rahman and S. Hailes. Supporting trust in virtual communities. In *33th IEEE International Conference on Systems Sciences*, pages 1–9, 2000.
2. I. Aldridge. *High-frequency trading: a practical guide to algorithmic strategies and trading systems*, volume 459. John Wiley and Sons, 2009.
3. M. Anderson and S.L. Anderson. Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot*, 42(4):324–331, 2014.
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. F. Berreby, G. Bourgne, and J.-G. Ganascia. Modelling moral reasoning and ethical responsibility with logic programming. In *LPAR proceedings*, pages 532–548, 2015.
7. Olivier Boissier, Rafael H Bordini, Jomi F Hübner, Alessandro Ricci, and Andrea Santi. Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6):747–761, 2013.
8. S. Bono, G. Bresin, F. Pezzolato, S. Ramelli, and F. Benseddik. Green, social and ethical funds in europe. Technical report, Vigeo, 2013.
9. J. Carbo, J. Molina, and J. Davila. Comparing predictions of SPORAS vs. a fuzzy reputation agent system. In *Conference on Fuzzy Sets and Fuzzy Systems*, pages 147–153, 2002.
10. J. Carter, E. Bitting, and A. Ghorbani. Reputation formalization for an information-sharing multi-agent system. *Computational Intelligence*, 18(2):515–534, 2002.
11. C. Castelfranchi and R. Falcone. *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons, 2010.
12. H. Coelho and A.C. da Rocha Costa. On the intelligence of moral agency. *Encontro Português de Inteligência Artificial*, pages 12–15, October 2009.
13. 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.
14. N. Cointe, G. Bonnet, and O. Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *AAMAS proceedings*, pages 1106–1114, 2016.
15. N. Cointe, G. Bonnet, and O. Boissier. Multi-agent based ethical asset management. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 52–57, 2016.
16. R. Conte and M. Paolucci. *Reputation in artificial societies: Social beliefs for social order*, volume 6. Springer Science & Business Media, 2002.
17. B. Esfandiari and S. Chandrasekharan. On how agents make friends: Mechanisms for trust acquisition. In *Deception, Fraud, and Trust in Agent Societies Workshop*, pages 27–34, 2001.
18. Directorate-General for Economic and Financial Affairs. Impact of the current economic and financial crisis on potential output. Occasional Papers 49, European Commission, June 2009.
19. J.-G. Ganascia. Ethical system formalization using non-monotonic logics. In *29th Annual Conference of the Cognitive Science Society*, pages 1013–1018, 2007.
20. J.-G. Ganascia. Modelling ethical rules of lying with Answer Set Programming. *Ethics and Information Technology*, 9(1):39–47, 2007.

21. H.J.N Horsburgh. The ethics of trust. *The Philosophical Quartely*, 10(41):343–354, 1960.
22. R. Johnson. Kant's moral philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer edition, 2014.
23. A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service proposition. *Decision Support Systems*, 43(2):618–644, 2007.
24. E. Lorini. On the logical foundations of moral agency. In *11th International Conference on Deontic Logic in Computer Science*, pages 108–122, 2012.
25. S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, Stirling, 1994.
26. A. McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter edition, 2014.
27. G. Muller, L. Vercouter, and O. Boissier. Towards a general definition of trust and its application to openness in MAS. In *6th Workshop on Deception, Fraud and Trust in Agent Societies*, pages 49–56, 2003.
28. P. Ricoeur. *Oneself as another*. University of Chicago Press, 1995.
29. A. Rocha-Costa. Moral systems of agent societies: Some elements for their analysis and design. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 32–37, 2016.
30. J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence*, 24(1):33–60, 2005.
31. Jordi Sabater-Mir and Laurent Vercouter. Trust and reputation in multiagent systems. *Multiagent Systems*, page 381, 2013.
32. A. Saptawijaya and L. Moniz Pereira. Towards modeling morality computationally with logic programming. In *Practical Aspects of Declarative Languages*, pages 104–119. 2014.
33. M. Timmons. *Moral theory: an introduction*. Rowman & Littlefield, 2012.
34. L. Vercouter and G. Muller. L.I.A.R.: Achieving social control in open and decentralized multiagent systems. *Applied Artificial Intelligence*, 24(8):723–768, 2010.
35. B. Yu and M.P. Singh. Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4):535–549, 2002.