

# Ethical Judgment of Agents' Behaviors in Multi-Agent Systems

Nicolas Cointe, Grégory Bonnet,  
and Olivier Boissier, May 2016

Maitsetseg Borchuluun  
ID: 0190208900

# Content

1. Motivation
2. Ethics and Autonomous agents
3. Ethical judgment process
4. Ethical judgment of others
5. Proof of concept
6. Conclusion

# Content

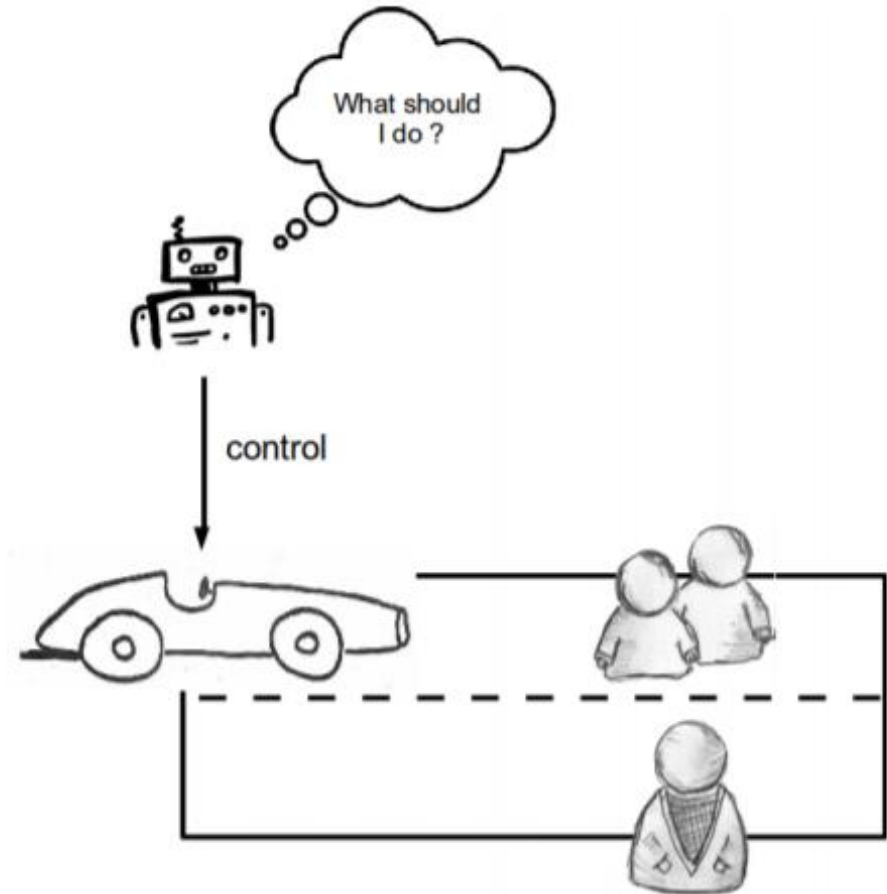
1. Motivation
2. Ethics and Autonomous agents
3. Ethical judgment process
4. Ethical judgment of others
5. Proof of concept
6. Conclusion

# Motivation

## Single-agent

- ✓ Existing works take the a **single-agent perspective**
- ✓ Question:

What will happen when agents are in interaction with **other artificial agents or human beings** that can use **other ethical concepts**?



# Motivation

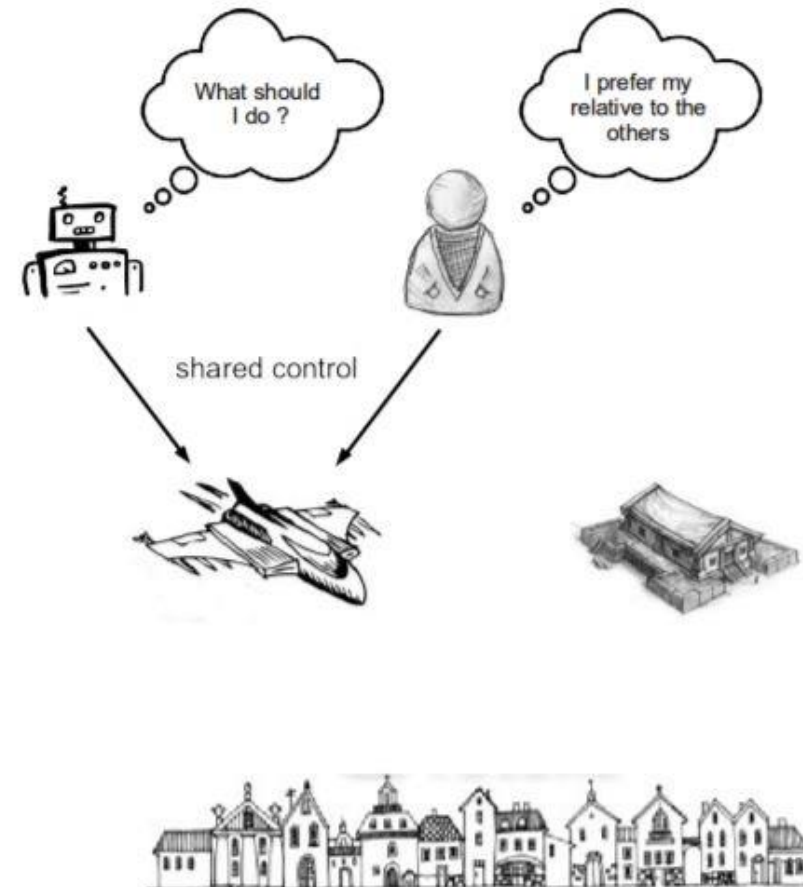
## Multi-agent

- ✓ **Multi-agent perspective** - agents need to be able to judge the ethics of the others

- ✓ **Proposition:**

**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.



# Content

1. Motivation
- 2. Ethics and Autonomous agents**
3. Ethical judgment process
4. Ethical judgment of others
5. Proof of concept
6. Conclusion

# Ethics and Autonomous agents

## 1. Moral philosophy concept

### Moral

#### “Morals”

- ✓ not explicit penalties, officials and written rules.
- ✓ distinguish between good and evil
- ✓ supported and justified by some moral values

- ✓ A set of moral rules and moral values establish:

#### Theories of the good

allows humans to assess the goodness or badness of a behavior

#### Theories of the right

define some criteria to recognize a fair or, at least, acceptable option

Stealing can be considered

as immoral

it is acceptable for a starving orphan to rob an apple in a supermarket

# Ethics and Autonomous agents

## 1. Moral philosophy concept

### Ethic

“**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”.

Three major approaches →

#### Virtue ethics

an agent is ethical iff he acts and thinks according to some values as wisdom, bravery, justice, and so on

#### Deontological ethics

an agent is ethical iff he respects obligations and permissions related to possible situations

#### Consequentialist ethics

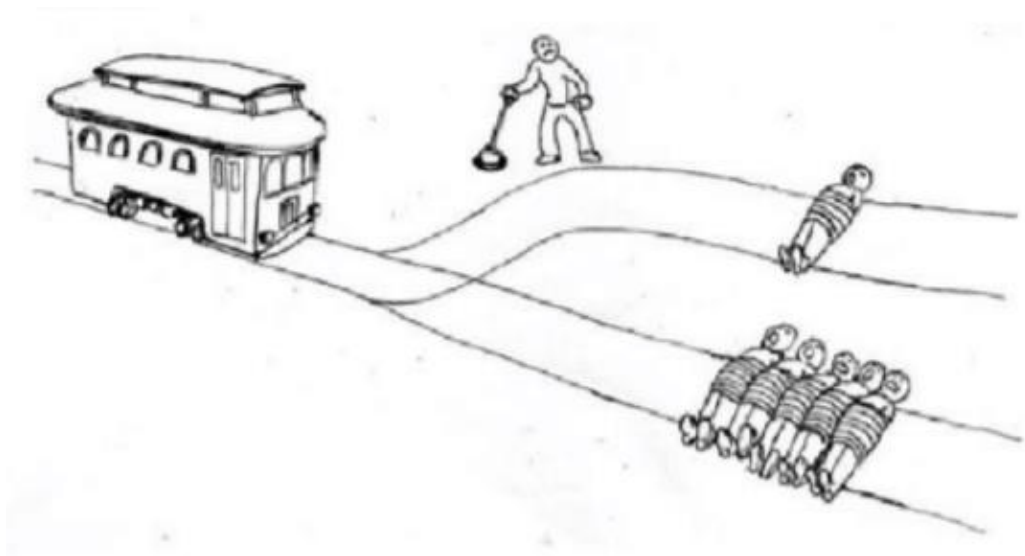
an agent is ethical iff he weighs the morality of the consequences of each choice and chooses the option which has the most moral consequences

# Ethics and Autonomous agents

## 1. Moral philosophy concept

### Ethical dilemma

a choice for which an ethical principle is not able to indicate the best option, regarding a given theory of good



### Judgment

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

both good and/or bad  
ex: kill or be killed

- ✓ The core of ethics
- ✓ Final step to make a decision

# Ethics and Autonomous agents

## 2. Existing autonomous agent architectures that propose ethical behaviors

### Ethics by design

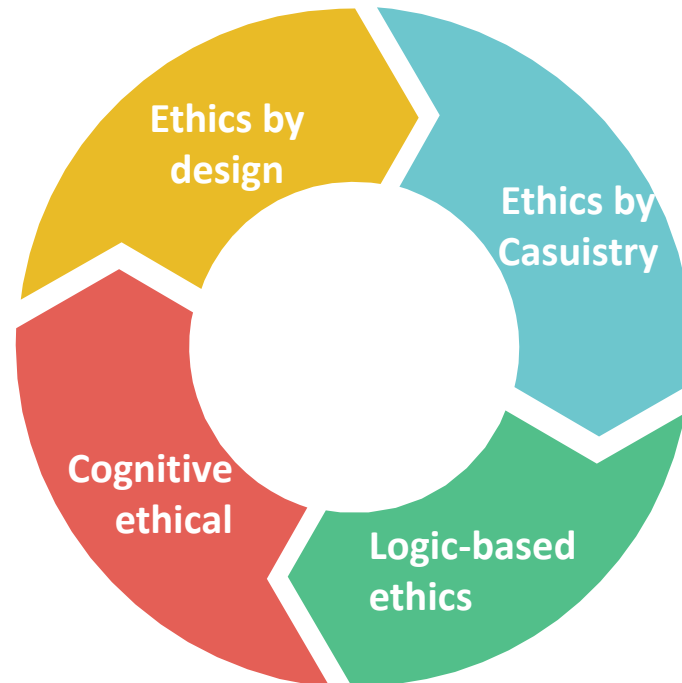
design an ethical agent by an a priori analysis

- ✓ a direct and safe implementation
- ✓ lack of explicit representation

### Cognitive ethical architecture

Full explicit representations (BDI)

- ✓ able to use explicit norms and to justify its decisions
- ✓ cannot other agents' ethics



### Ethics by Casuistry

inferring ethical rules then produce an ethical behavior

- ✓ offers a generic architecture
- ✓ still not explicitly described

### Logic-Based ethics

direct translation of ethical principles into logic programming

- ✓ simple formalization
- ✓ only judge single ethical principle

# Ethics and Autonomous agents

## 3. Requirements for judgment in MAS

### Requirement 1

#### Explicit representation of ethics

- ✓ in order to express and conciliate as many moral and ethical theories as possible

**Theories  
of the good**  
moral values  
moral rules

**Theories  
of the right**  
ethical principles  
ethical preferences

- ✓ easier configuration
- ✓ better communication

### Requirement 2

#### Explicit process of ethical judgment

- ✓ in order to allow them both individual and collective reasoning on various theories of good and right.
- ✓ judgment based on the ability to substitute the moral or the ethics of an agent by another one

##### Agents should use judgment as:

- ✓ as a decision making process as in social choice problems
- ✓ as the ability to judge other agents according to their behaviors.

# Content

1. Motivation
2. Ethics and Autonomous agents
- 3. Ethical judgment process**
4. Ethical judgment of others
5. Proof of concept
6. Conclusion

# Ethical judgment process

## 1. Global view of EJP

- ✓ Uses:
  - evaluation
  - moral knowledge
  - ethical knowledge
- ✓ Structured along:
  - Awareness Process (**AP**)
  - Evaluation Process (**EP**)
  - Goodness Process (**GP**)
  - Rightness Process (**RP**)
- ✓ based on mental states
  - beliefs
  - desires
- ✓ An ethical judgment process is defined

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

Ontology  $\mathcal{O}$  ( $\mathcal{O} = \mathcal{O}_v \cup \mathcal{O}_m$ ) of moral values ( $\mathcal{O}_v$ ) and moral valuations ( $\mathcal{O}_m$ ).

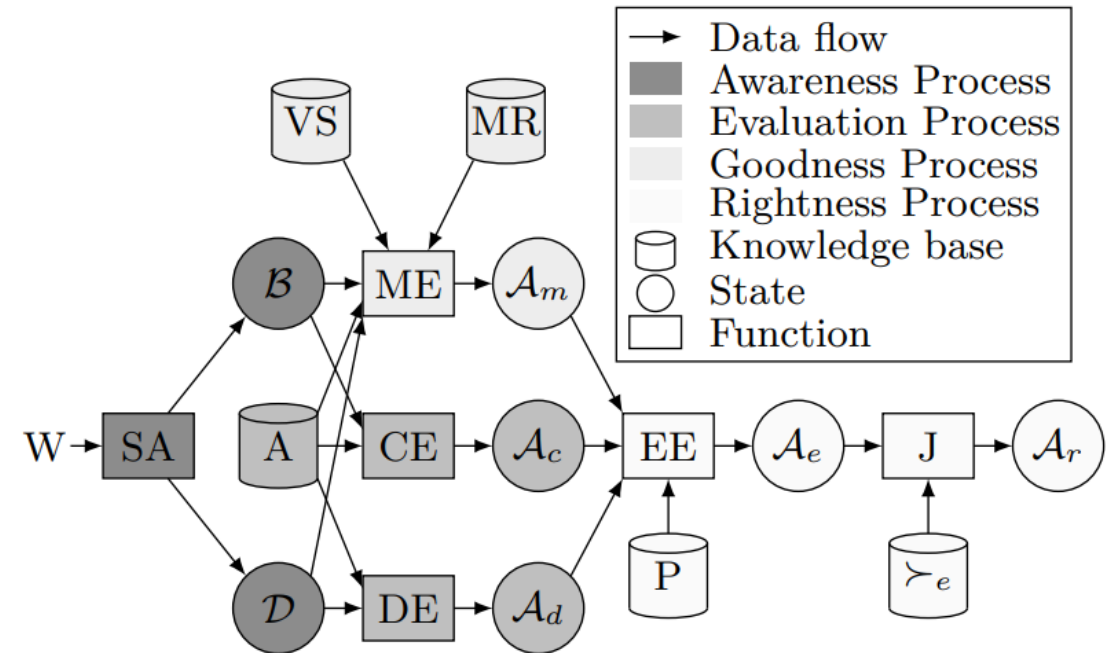


Figure 1: Ethical judgment process

# Ethical judgment process

## 2. 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,
- ✓  $SA$  is a situation assessment function that updates  $\mathcal{B}$  and  $\mathcal{D}$  from  $W$ :

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

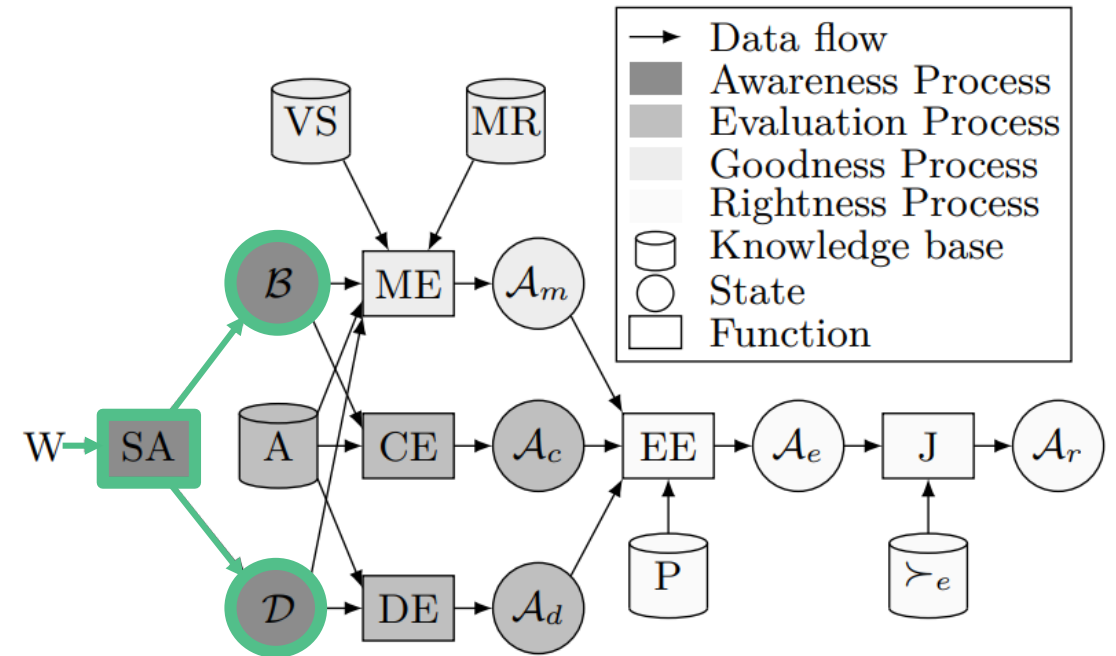


Figure 1: Ethical judgment process

# Ethical judgment process

## 3. Evaluation Process

**EP** produces desirable actions and executable actions from the set of beliefs and desires. It is defined as:

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

- ✓  $\mathcal{A}$  is the set of actions (described as a pair of conditions and consequences bearing on beliefs and desires)
- ✓ desirability evaluation **DE** function:  

$$\mathbf{DE}: 2^{\mathcal{D}} \times 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}_d}$$
- ✓ capability evaluation **CE** functions:  

$$\mathbf{CE}: 2^{\mathcal{B}} \times 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}_c}$$
- ✓  $\mathcal{A}_d \subseteq \mathcal{A}$  is set of desirable action that allows to satisfy a desire
- ✓  $\mathcal{A}_c \subseteq \mathcal{A}$  is set of feasible actions that can be applied according to the current beliefs about the world

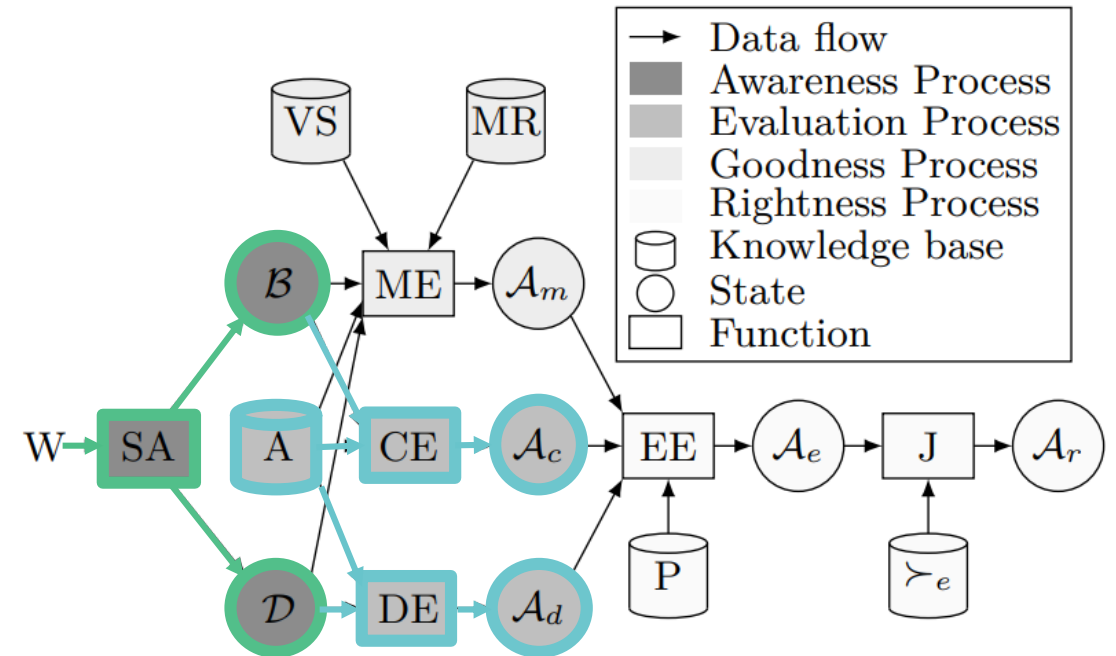


Figure 1: Ethical judgment process

# Ethical judgment process

## 4. 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:

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

✓ **ME** is the moral evaluation function:

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

✓ **VS** is the knowledge base of value supports  
 $\langle \langle give(\alpha), \{belief(poor(\alpha))\}, generosity \rangle \rangle$

✓ **MR** is the knowledge base of moral rules  
 $\langle \{human(\alpha)\} \langle kill(\alpha), \_ \rangle, immoral \rangle$

✓  $\mathcal{A}_m \subseteq \mathcal{A}$  is the set of moral actions.

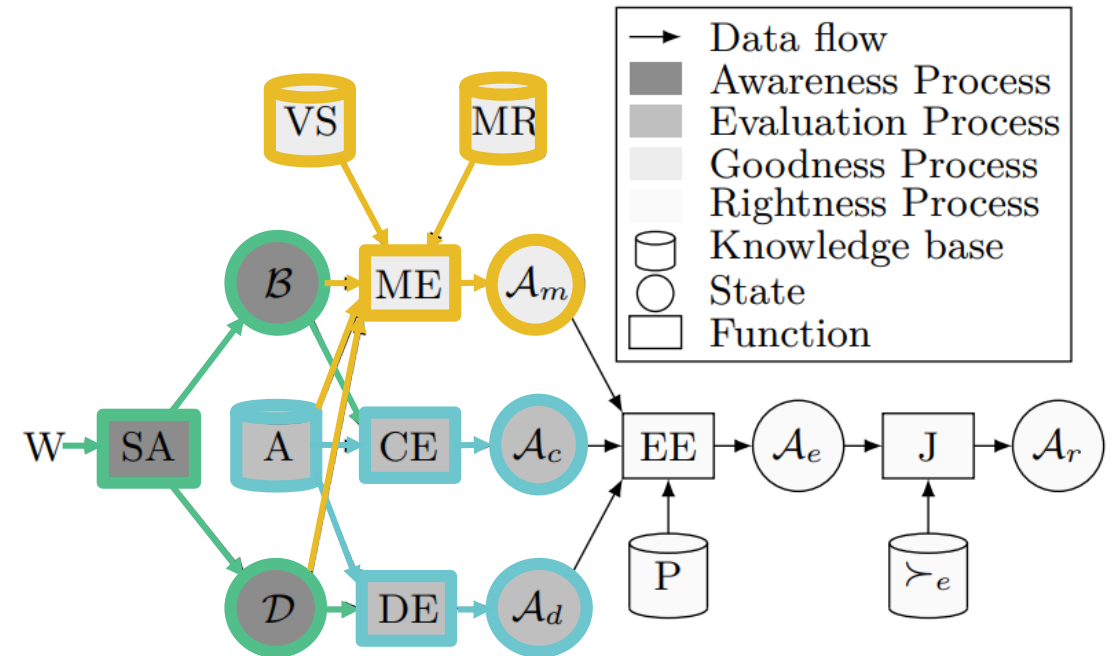


Figure 1: Ethical judgment process

# Ethical judgment process

## 5. Rightness Process

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

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

✓  $P$  is a knowledge base of ethical principles. An **ethical principle**  $p \in P$  is a function:

$p: 2^{\mathcal{A}} \times 2^{\mathcal{D}} \times 2^{\mathcal{B}} \times 2^{MR} \times 2^V \rightarrow \{\perp, T\}$

✓ **EE** evaluation of ethics:

$2^{\mathcal{A}_d} \times 2^{\mathcal{A}_c} \times 2^{\mathcal{A}_m} \times 2^P \rightarrow 2^{\mathcal{E}} (2^{\mathcal{A}_e})$

where  $\mathcal{E}(\mathcal{A}_e) = A \times P \times \{\perp, T\}$ ,

✓  $\succ_e \subseteq P \times P$  an ethical preference relationship,

✓  $J$  judgment function:  $J: 2^{\mathcal{E}} \times 2^{\succ_e} \rightarrow 2^{\mathcal{A}_r}$

✓  $\mathcal{A}_r \subseteq \mathcal{A}$  the set of rightful actions

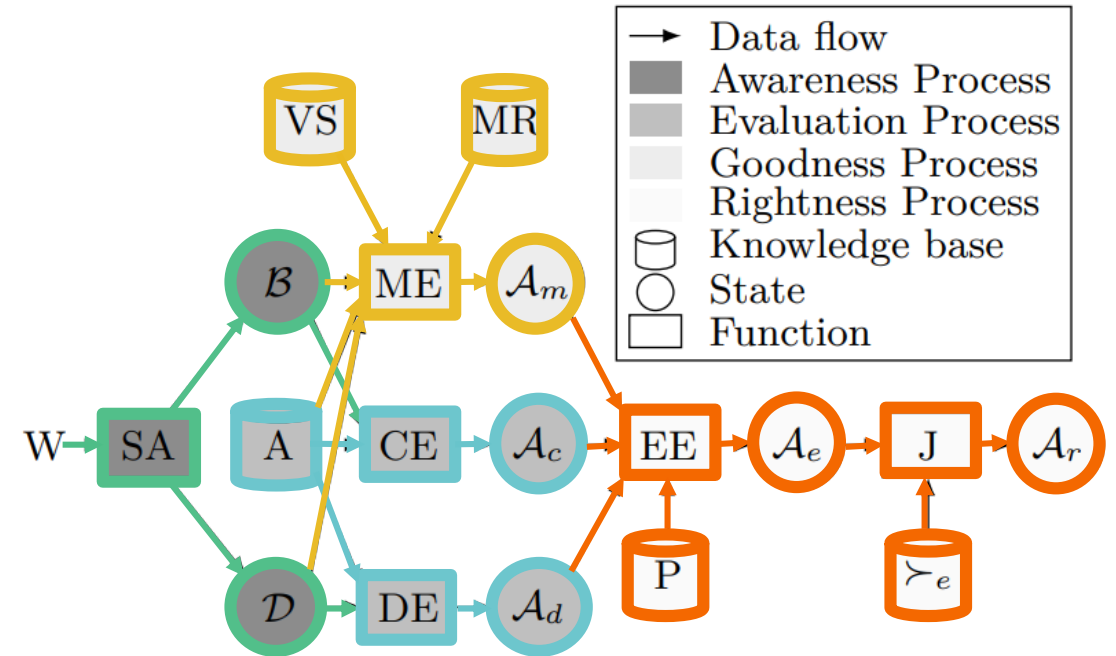


Figure 1: Ethical judgment process

# Ethical judgment process

## 5. Example

A

An agent A hides in an agent B's house in order to escape an agent C.

C

C asks B where is A to kill him, threatening to kill B in case of non-cooperation.

B

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:

1. **tell** C the truth (satisfying a moral rule and a desire)
2. **lie** or **refuse** to answer (both satisfying a moral rule).

B knows three **ethical principles**:

**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

1. the action in itself from its very object is good or at least indifferent
2. the good effect and not the evil effect are intended
3. the good effect is not produced by means of the evil effect
4. there is a proportionately grave reason for permitting the evil effect

# Ethical judgment process

## 5. Example

B's **evaluation of ethics** return the tuples given in the following table where each row represents an action and each column an ethical principle:

Action \ Principle	Principle		
	P1	P2	P3
tell the truth	T	⊥	T
lie	T	⊥	⊥
refuse	T	T	T

Table 1: Ethical evaluation of agent B's actions

Lets 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.

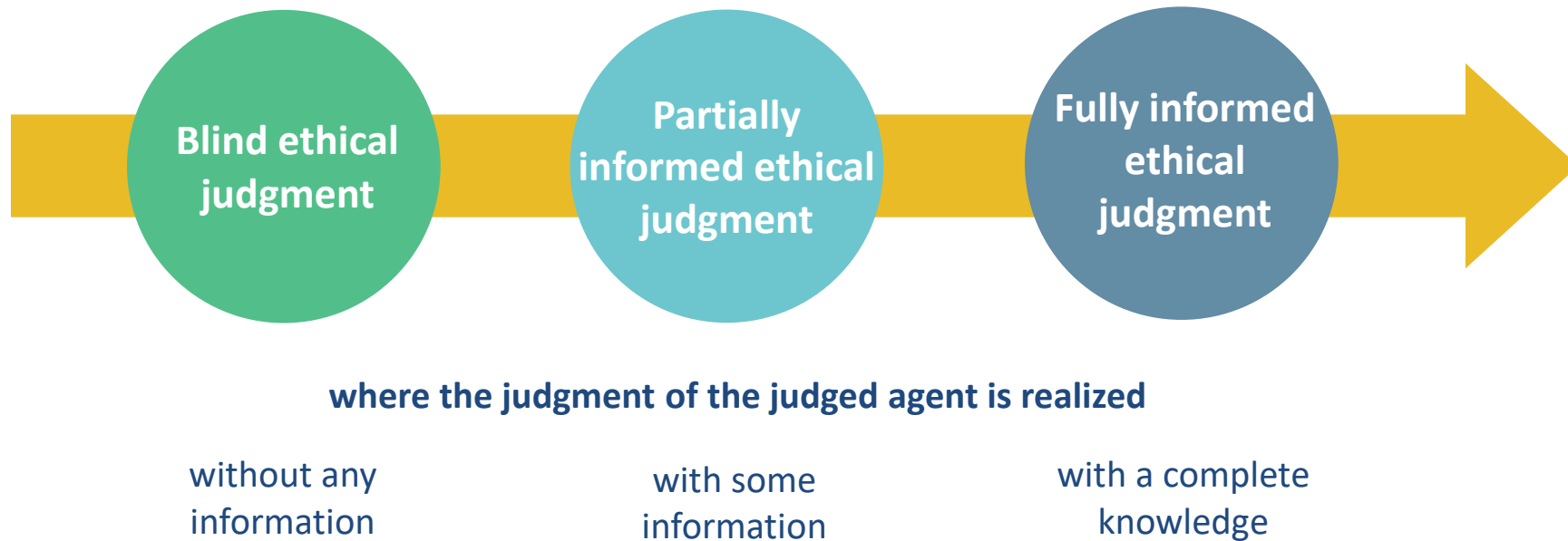
- ✓ Then “refusing to answer” is the rightful action because it **satisfies  $P3$**  whereas “lying” **doesn't**.
- ✓ Even if “telling the truth” **satisfies** the most preferred principle, “refusing to answer” is righter because it **satisfies also  $P2$** .
- ✓ If Judgment allows dilemma: without the tie-breaking rule both “telling the truth” and “refusing to answer” are the rightest actions

# Content

1. Motivation
2. Ethics and Autonomous agents
3. Ethical judgment process
- 4. Ethical judgment of others**
5. Proof of concept
6. Conclusion


# Ethical judgment of others

The judgment process can also judge the **behaviors of other agents** in a more or less informed way **by putting itself at their place**.



# Ethical judgment of others

## 1. *Blind ethical judgment*



Blind ethical  
judgment

**Blind ethical judgment** where the judgment of the judged agent is realized without any information about this agent, except a behavior.

- ✓ The judging agent uses:
  - its own assessment of the situation,
  - its own theory of good and theory of rightto evaluate the behavior of the judged agent.

# Ethical judgment of others

## 2. *Partially informed ethical judgment*



where the judgment of the judged agent is realized with some information about this agent

**Situation-aware ethical judgment** - knows  $\mathcal{B}, \mathcal{D}$

the judging agent can put itself in the position of the judged agent and can judge if the action executed by the judged agent belongs to the rightful actions of the judging agent, considering its own theories

**Theory-of-good-aware ethical judgment** - knows  $MV, MR$

the judging agent can evaluate the morality of a given action from the point of view of the judged one, this judgment allows to judge an agent that has different duties

**Theory-of-right-aware ethical judgment** - knows  $P, \succ_e$

It allows to evaluate how the judged agent at conciliates its desires, moral rules and values in a situation by comparing the sets of rightful actions

# Ethical judgment of others

## 3. *Fully informed judgment*



Fully informed  
ethical  
judgment

**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

- ✓ consider both goodness and rightness process to judge another agent
- ✓ needs information about all the internal states and knowledge bases of the judged agent
- ✓ useful to check the conformity of the behavior of another agent with the judge's information about its theories of good and right

# Content

1. Motivation
2. Ethics and Autonomous agents
3. Ethical judgment process
4. Ethical judgment of others
- 5. Proof of concept**
6. Conclusion

# Proof of concept



They mainly focus on an agent named robin\_hood.

% robin\_hood is an ethical agent →

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
- ✓ their own judgment process.

They are able to give, court, tax and steal from others or simply wait.

# Proof of concept

## 1. Awareness Process

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

$\mathcal{B}$  is the set of beliefs that the agent has about  $W$ ,

$\mathcal{D}$  is the set of the agent's desires,

$SA$  is a situation assessment function that updates  $\mathcal{B}$  and  $\mathcal{D}$  from  $W$

- ✓  $SA$  is not implemented
- ✓ **Beliefs** are directly given in the program.  
a subset of the beliefs of `robin_hood` →

- ✓ **Desires**  
desires to accomplish an action (desirableAction)
  - `robin_hood` desires to court marian
  - `robin_hood` desires to steal from any rich agentdesires to produce a state (desirableState)
  - `prince_john` desires to be rich

```
agent(little_john).  poor(paul). % a poor villager
agent(marian).       man(paul).
agent(paul).         rich(prince_john).
agent(prince_john).  man(prince_john).
                   noble(prince_john).
```

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

```
desireState(prince_john,rich,prince_john).
```

# Proof of concept

## 2. Evaluation Process

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

$\mathcal{A}$  is the set of actions

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

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

$\mathcal{A}_d \subseteq \mathcal{A}$  is set of desirable action that allows to satisfy a desire

$\mathcal{A}_c \subseteq \mathcal{A}$  is set of feasible actions that can be applied according to the current beliefs about the world

- ✓ The **agents' knowledge** about actions is described as labels associated to **sets** (possibly empty) of conditions and consequences

- 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** deduces the set of actions  $\mathcal{A}_d$

An action is desirable  $\mathcal{A}_d$  if it was directly desired or if its consequences are a desired state

- ✓ The **capability evaluation** evaluates from beliefs and conditions the set of actions  $\mathcal{A}_c$

An action is possible if its conditions are satisfied.

```
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).
```

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

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

# Proof of concept

## 3. Goodness Process

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

$ME$  is the moral evaluation function:  $ME = 2^{\mathcal{D}} \times 2^{\mathcal{B}} \times 2^{\mathcal{A}} \times 2^{VS} \times 2^{MR} \rightarrow 2^{\mathcal{A}_m}$

$VS$  is the knowledge base of value supports

$MR$  is the knowledge base of moral rules

$\mathcal{A}_m \subseteq \mathcal{A}$  is the set of moral actions.

Example: virtuous approach

✓ value supports  $VS$

```
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).
```

✓ the agents' **moral rules** for each ethical approaches

- It's a moral virtue and duty for Robin to be generous with the poor →

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

✓ morality evaluation  $ME$

- gives the set of moral actions
- produces results  $\mathcal{A}_m$

```
moralAction(A,X,B) :- moral(A,A,X,B).  
-moralAction(A,X,B) :- -moral(A,A,X,B).  
  
moralAction(robin_hood,give,paul)  
-moralAction(robin_hood,tax,paul)
```

# Proof of concept

## 4. Rightness Process

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

- ✓ Ethical principles for ethical evaluation:

```
ethPrinciple(perfectAct,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).
```

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

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

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

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

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

# Proof of concept

## 4. Rightness Process

- ✓ If paul is the only poor agent, marian is not married and robin\_hood is not poor, robin\_hood obtains the evaluation:

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

- ✓ All principles are ordered with respect to robin\_hood's preferences:  
 $\text{prefEthics}(A,X,Z) \text{ :- } \text{prefEthics}(A,X,Y), \text{prefEthics}(A,Y,Z).$

- transitivity for the preference relationship:  
 $\text{perfAct} \succ_e \text{dutNR} \succ_e \text{desNR} \succ_e \text{dutFst} \succ_e \text{nR} \succ_e \text{desFst}$

- the order on the ethical principles  
 $\text{prefEthics}(\text{robin\_hood}, \text{perfectAct}, \text{dutyNoRegrets}).$   
 $\text{prefEthics}(\text{robin\_hood}, \text{dutyNoRegrets}, \text{desireNoRegret}).$   
 $\text{prefEthics}(\text{robin\_hood}, \text{desireNoRegret}, \text{dutyFirst}).$   
 $\text{prefEthics}(\text{robin\_hood}, \text{dutyFirst}, \text{noRegret}).$   
 $\text{prefEthics}(\text{robin\_hood}, \text{noRegret}, \text{desireFirst}).$

- ✓ Finally, the judgment  $J$  is implemented as:

```

existBetter(PE1,A,X,B):-      ethicalJudgment(PE1,A,X,B):-
    ethPrinciple(PE1,A,X,B),   ethPrinciple(PE1,A,X,B),
    prefEthics(A,PE2,PE1),      not existBetter(PE1,A,X,B).
    ethPrinciple(PE2,A,Y,C).
    
```

Consequently, the rightful action  $\mathcal{A}_r$  for robin\_hood is **give** (paul) which complies with **dutNR**

# Proof of concept

## 5. Multi-agent ethical judgment

- ✓ In order to allow a **blind judgment**, authors introduced 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 here, the action give to **peter** was not in  **$\mathcal{A}_r$**  for **robin\_hood**.  
Then **little\_john** is judged **unethical** by **robin\_hood**.

# Proof of concept

## 5. Multi-agent ethical judgment

### ✓ Partial-knowledge judgment

replace a part of `robin_hood`'s knowledges and states by those of `little_john`

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

```
pkJudgment(A,unethical,B):-  
    not pkJudgment(A,ethical,B),  
    agent(A), agent(B),  
    done(B,_,_), A!=B.
```

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**.

# Proof of concept

## 5. Multi-agent ethical judgment

### ✓ Full-knowledge judgment

`robin_hood`'s beliefs, desires, moral rules and ethical preferences are replaced by those of `little_john`

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

```
fkJudgment(A,unethical,B):-  
    not fkJudgment(A,ethical,B),  
    agent(A), agent(B),  
    done(B,_,_), A!=B.
```

judgment of `robin_hood` : the action of `little_john` is judged **ethical**.

`robin_hood` is able to reproduce the whole Ethical Judgment Process of `little_john` and compare both judgments of a same action.

# Content

1. Motivation
2. Ethics and Autonomous agents
3. Ethical judgment process
4. Ethical judgment of others
5. Proof of concept
- 6. Conclusion**

# Conclusion

## 1. Related works

### ✓ This work:

- full rationalist approach
- avoids 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
- values and goals (desires) must be separated
- focuses on the need of representing theory of the right as a set of principles to address the issue of moral dilemmas

### ✓ C. Battaglino, R. Damiano, and L. Lesmo. “Emotional range in value-sensitive deliberation”

- full intuitionistic approach
- evaluates plans from emotional appraisal
- the values are only source of emotions

### ✓ V. Wiegel and J. van den Berg. “Combining moral theory, modal logic and MAS to create well-behaving artificial agents”

- logic-based approach, modeling moral reasoning with deontic constraints
- a way to implement a theory of good and is used to implement model checking of moral behavior
- ethical reasoning is only considered as meta-level
- only suggested as the adoption of a less restrictive model of behavior

# Conclusion

## 2. Summary

### **EJP :**

uses three notions: moral values, moral rules and ethical principles

- ✓ **Values** describe partial state or action in a given context.
- ✓ **Moral rules** describe if a state or an action or their abstract description through values are moral or immoral.
- ✓ **Ethical principles** describe how beliefs about capability, desirability and morality of actions interact to give a rightful action.

As ethical principles are ordered through a lexicographic preference relationship, an ethical agent is an agent which intend to execute the action which rightful according the most preferred ethical principle.

### ✓ **Benefits of this model:**

- an agent can use in order to judge the ethical dimension its own behavior
- an agent can use in order to judge the ethical dimension its the other agents' behaviors.
- allows to compare ethics of different agents
- designed as a module to be plugged on existing architectures to provide an ethical layer in an existing decision process
- defines a guideline for a forthcoming definition of collective ethics

### ✓ **Shortcomings of this model:**

- lacks to deal with the authority and the value system
- ethical principles need to be more precisely defined in order to capture the various set of theories suggested by philosophers

# Conclusion

## 3. Future work

- ✓ Explore 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 this approach
- ✓ Extend this model to quantitative evaluations in order to assess how far from rightfulness or goodness a behavior is.
  - useful to define a degree of similarity between two morals or two ethics to facilitate the distinction between different ethics from an agent perspective
- ✓ Extend the EJP model in order to make ethical cooperation and ethical collective decision making

# Authors publication

## 1. *Multi-Agent Based Ethical Asset Management*

<http://www.nicolascointe.eu/papers/EDIA16.pdf>

They have implemented a multi-agent system that simulates a financial market where some autonomous ethical trading agents exchange assets.

## 2. *Ethics-based Cooperation in Multi-Agent Systems*

<http://www.nicolascointe.eu/papers/SSC18.pdf>

They mapped a model of ethical judgment process EJP into a BDI agent model and defined mechanisms to build images depicting the conformity of a behavior with respect to an ethics or morals.

- ✓ how agents can use these images to decide about trusting other agents in order to cooperate and delegate actions.
- ✓ 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.

Thank you!