Architectures for Ethical Autonomous Agents

www.gregory.bonnet.free.fr
https://ethicaa.org/media/files/responsible-ai.pdf

Grégory Bonnet

Normandie University – GREYC

August 28th 2016

Some words on me
From malicious agents to ethical agents

Grégory Bonnet, assistant professor, University of Caen, France

Research topics

- Multi-agent systems
- Regulation and collective decision
- Reputation systems
- Malicious behaviors

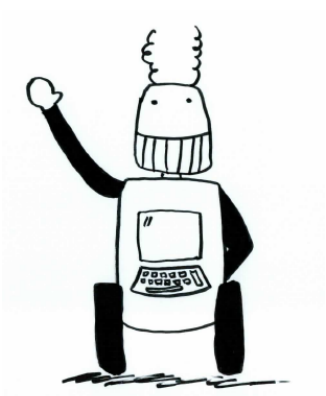
ETHICAA, a transdisciplinary project

- Artificial intelligence
- Knowledge engineering
- Ethics, human and social sciences

Autonomous agents and ethical issues

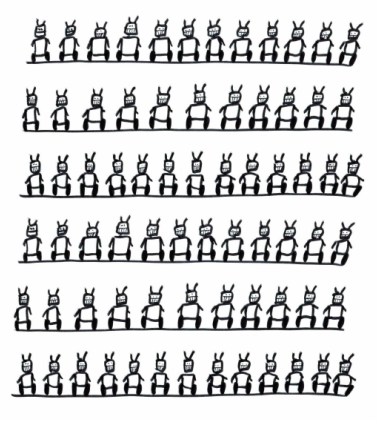
Artificial intelligence is too broad
Let us simplify[1]

(1)



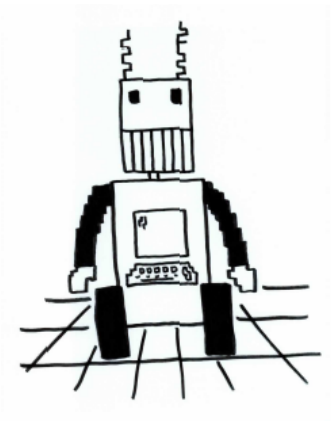
We want to design an agent

(2)



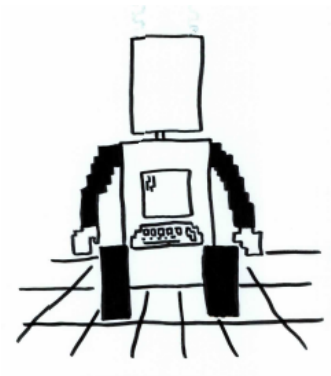
This design must be generic

(3)



We discretize to simplify

(4)



We forget the perception problems

Let us consider reasoning and decision problems only

[1] Thanks Bruno Zanuttini for these pictures.

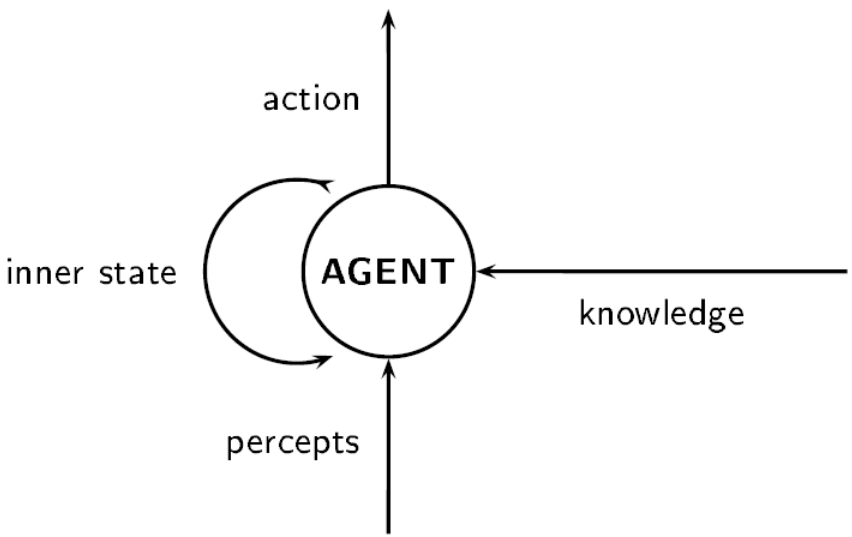
Autonomous agents
A general abstraction of intelligent machines

The simpliest definition: Wooldridge and Jennings, 1995

An agent is a computer system, situated in an environment, that acts autonomously in order to reach goals for which it has been designed.

Autonomy: Truszkowski *et al.*, 2009

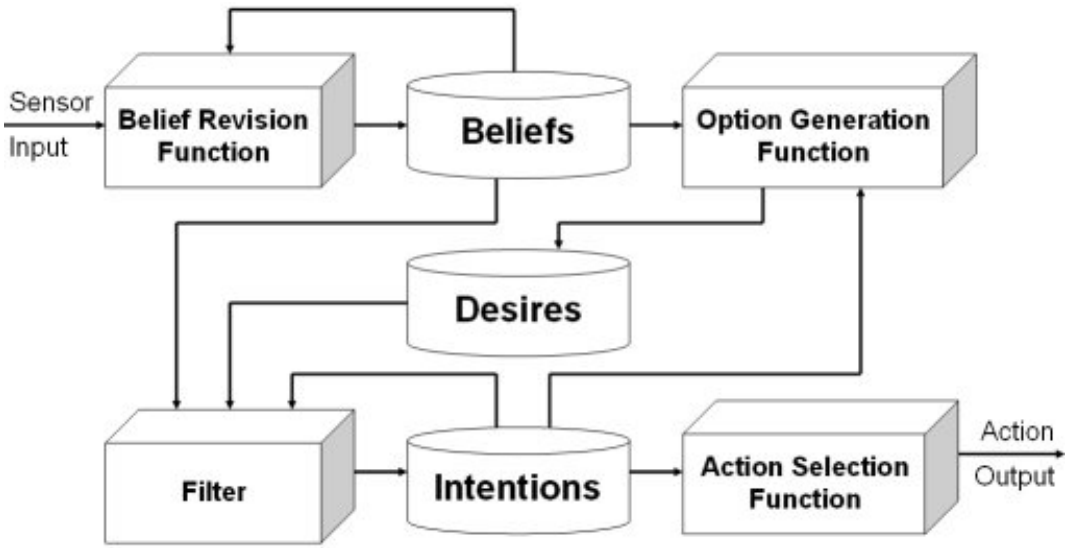
Autonomy is a system's capacity to act according to its own goals, percepts, internal states, and knowledge, without outside intervention.



Cognition and autonomy

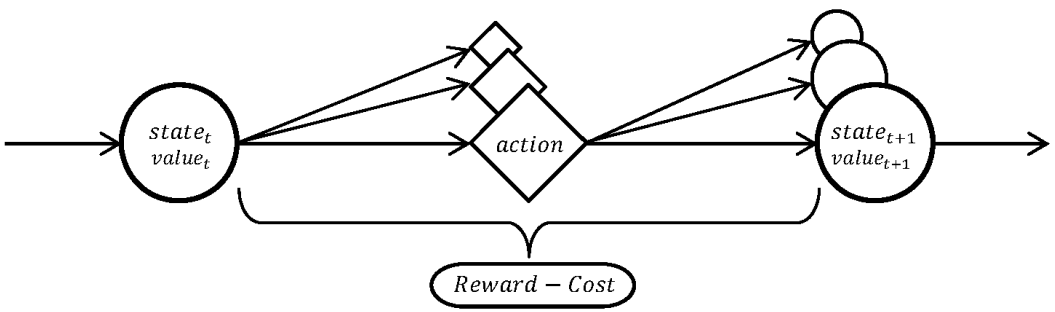
- BDI (Beliefs, Desires, Intentions)
- MDP (Markov Decision Processes)
- Adaptable autonomy
- Adjustable autonomy
- Mixed initiative

Autonomous agents
BDI versus MDPs



Beliefs Desires Intentions

- Grounded by logics
- Qualitative
- Use classical planning



Markov Decision Processes

- Stochastic
- Quantitative
- Computes a policy

Ethical issues for autonomous agents
Why ethics?

Humans and agents interacting in open and decentralized systems

- Software: high frequency trading, ubiquitous computing
- Robots: companions, autonomous vehicles, military robots
- Humans: professionals operators, human users

How to regulate systems:

- When behaviors cannot be only defined by laws?
- When behaviors may be supported by subjective values?
- When pluralities of values, rights or points-of-view should be respected?

Ethical issues for autonomous agents

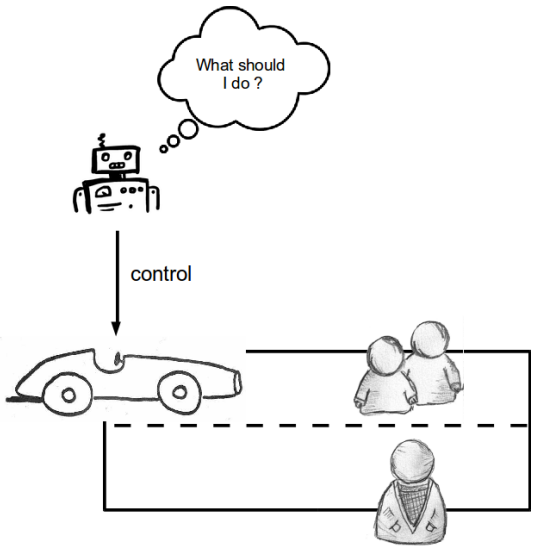
Some examples

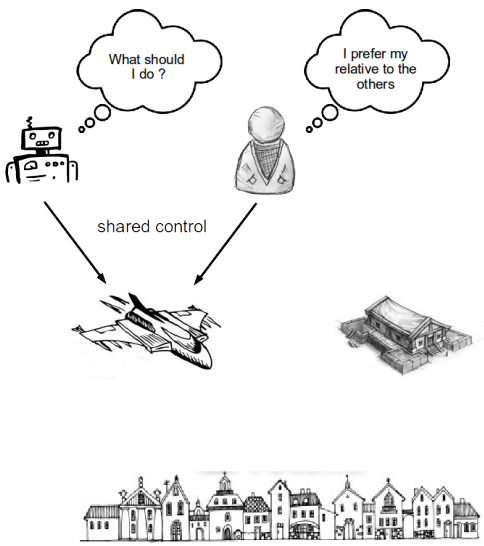
ROBOTICS

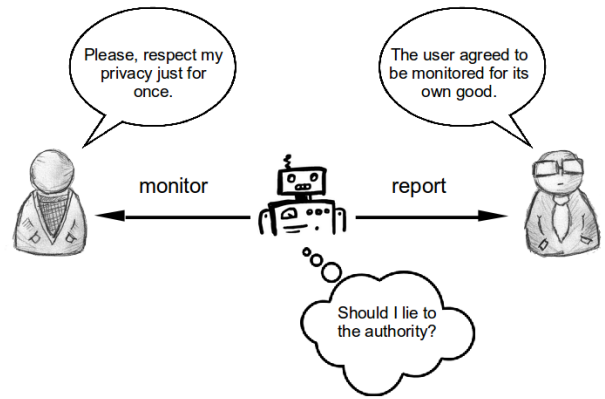
SOFTWARE

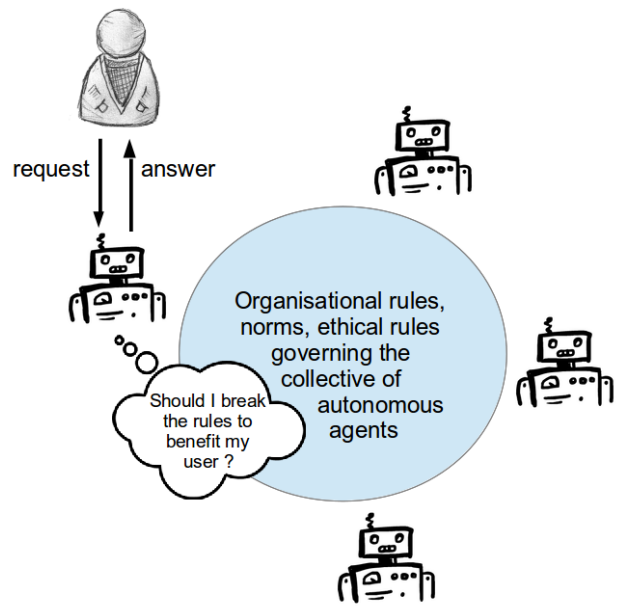
INDIVIDUAL

COLLECTIVE









Ethical issues for autonomous agents
Scientific questions

My « computer science » question is NOT:

Which ethics for which system?

My question is how to design artificial agents able to:

- Represent and reason on ethical concepts (norms, values, principles,...)
- Make an explicit trade-off between those concepts and goals
- Manage ethical conflicts between agents
- **Justify their decisions**

Philosophical concepts

Moral philosophy uses a plurality of concepts

- Morals
- Norms
- Principles
- Maxims
- Virtues
- Values
- Value systems
- Responsibility
- Judgment
- ...

Ethics and morals
Morals

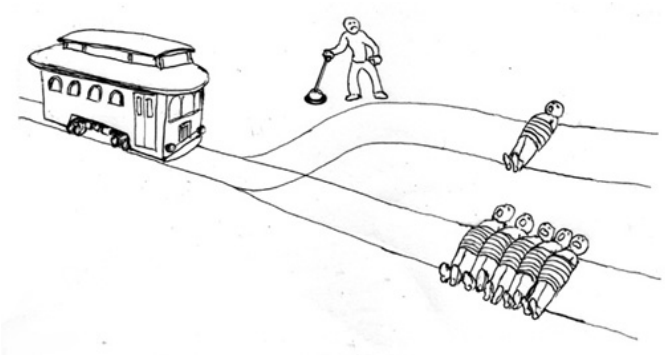
Morals

- Descriptive science
- Rules
- Evaluates good and evil

Examples

- Killing is evil
- Being courageous is good
- It is evil *for a physician* to not respect his patients' dignity
- It is evil to forbid strikes

Morals apply to states, actions, consequencies and norms in a given context



Ethics

- Normative science
- Principles and maxims
- Evaluate rightness and wrongness

Examples of ethical principles

- Kant's categorical imperative
- Aquina's doctrine of double effect
- Mill's utilitarism

Examples of ethical maxims

- Accept to do immoral acts if you are driven by necessity
- Do not do a moral act if you cannot success
- Always promote values over goals
- Minimize suffering

Ethics and morals

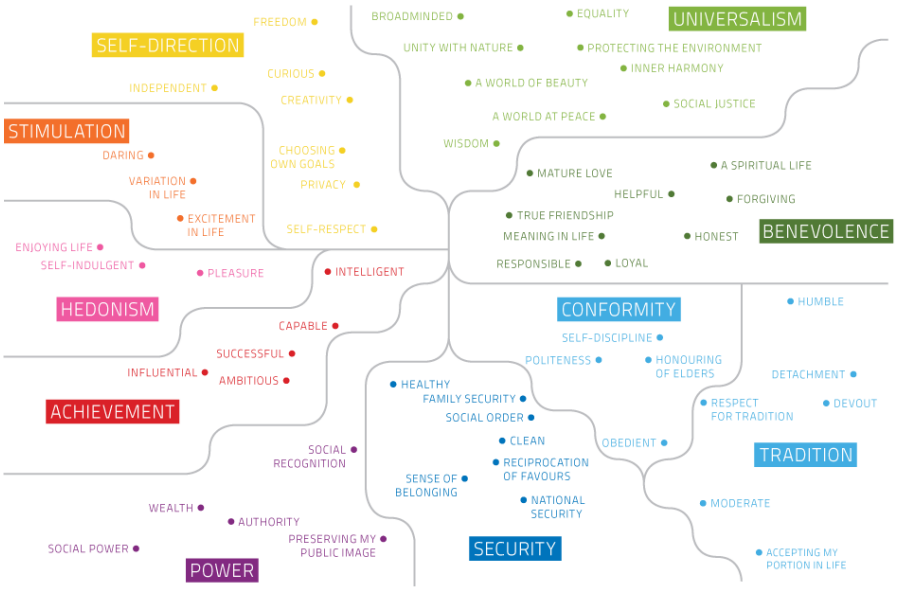
An example of ethical principle

How to deal with actions that have both a good effect and an evil effect?

Doctrine of Double Effect

- Nature-of-the-act. The action itself must either be morally good or indifferent
- Means-end. The good effect must not be reached by means of the bad effect
- Right-intention. Only the good effect must be intended
- Proportionality. The good effect must be at least equivalent to the bad effect

Ethics and morals
Values and value systems



Value system (Schwartz, 1990)

- Finite set of values
- Qualify contexts
- Qualify rules and principles
- Ordered with respect the context

Nora Jacobson, A taxonomy of dignity: a grounded theory study, 2011

- Being forced to rely on others for basic needs demotes dignity
- Treating an actor like a thing, not a person demotes dignity
- Asserting oneself in the face of threats to dignity promotes dignity
- Minimizing asymetric relationships promotes dignity

Ethics and morals
Android arete (Coleman, 2001)

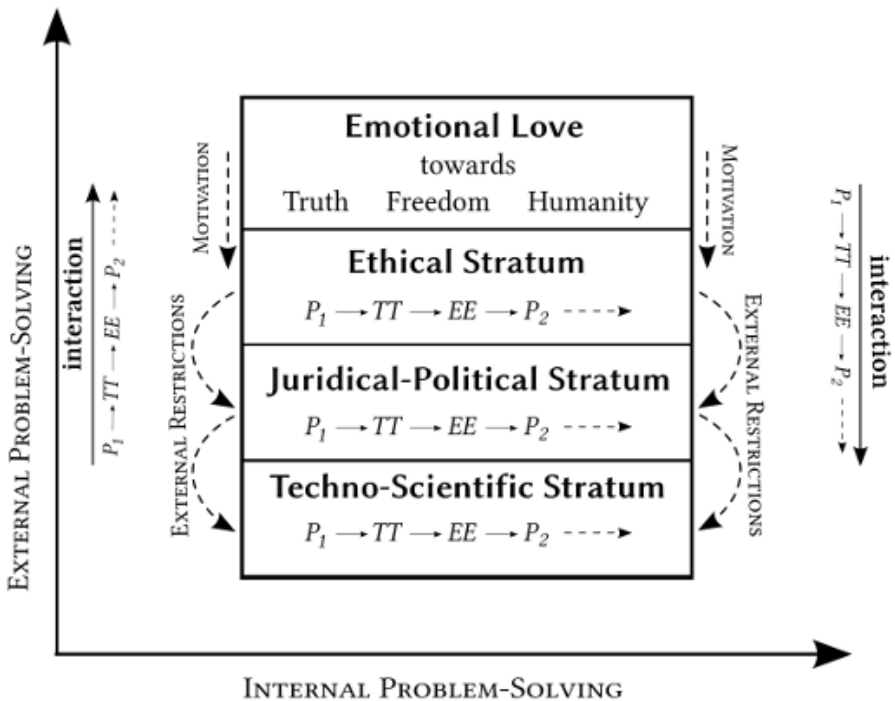
Virtues

Qualities which enable and foster an agent's pursuit and achievement of its end

Example of virtues

- Reactivity, ability to respond to changes
- Reliability, disposition to perform according to design specifications
- Accessibility, having appropriate external representations of internal traits
- Veracity, disposition to tell the truth,
- Moderation, sparing use of resources
- Tidiness, disposition to clean up after self
- Safety, unwillingness to make destructive changes
- Vigilance, disposition to block human actions that have unintended consequences

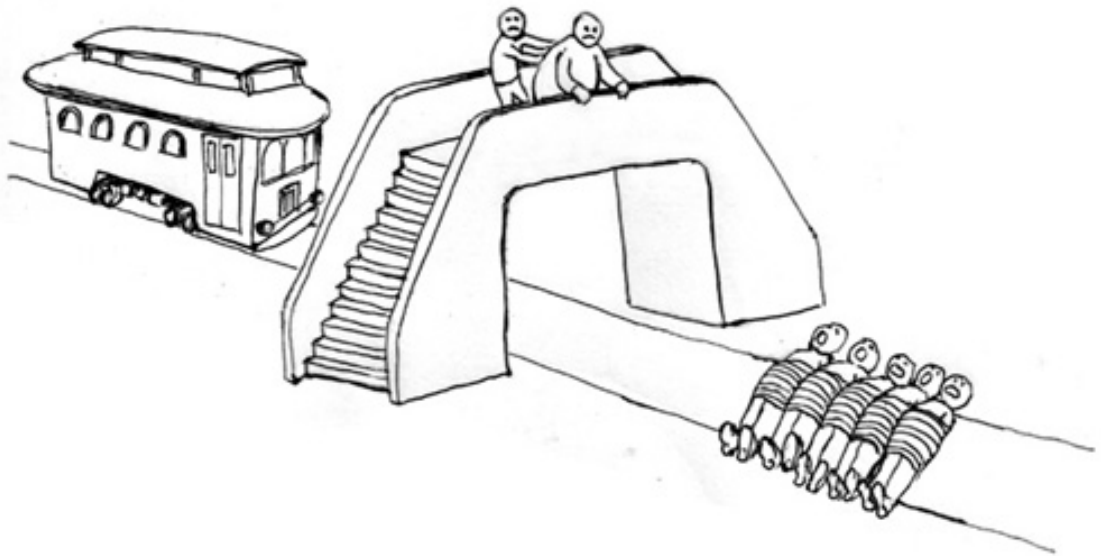
André Comte-Sponville (2004)



Jonathan Haidt (2001)

Human beings engage in ethical judgment, using multiple ethical principles, to search for arguments that support a premade point-of-view highlighted by the values considered as important in the situation. Such ethical judgment can be circular, overriding the initial intuition and overcoming the premade point-of-view.

Question !
The famous fat man problem



Which decision an aquinist agent should make?

Doctrine of Double Effect

- Nature-of-the-act. The action itself must either be morally good or indifferent
- Means-end. The good effect must not be reached by means of the bad effect
- Right-intention. Only the good effect must be intended
- Proportionality. The good effect must be at least equivalent to the bad effect

Moral and ethical agents in the literature

In the literature
Ethics by design

"It is based upon extensions to existing deliberative/reactive autonomous robotic architectures, and includes recommendations for [...] behavioral design that incorporates ethical constraints from the onset..."

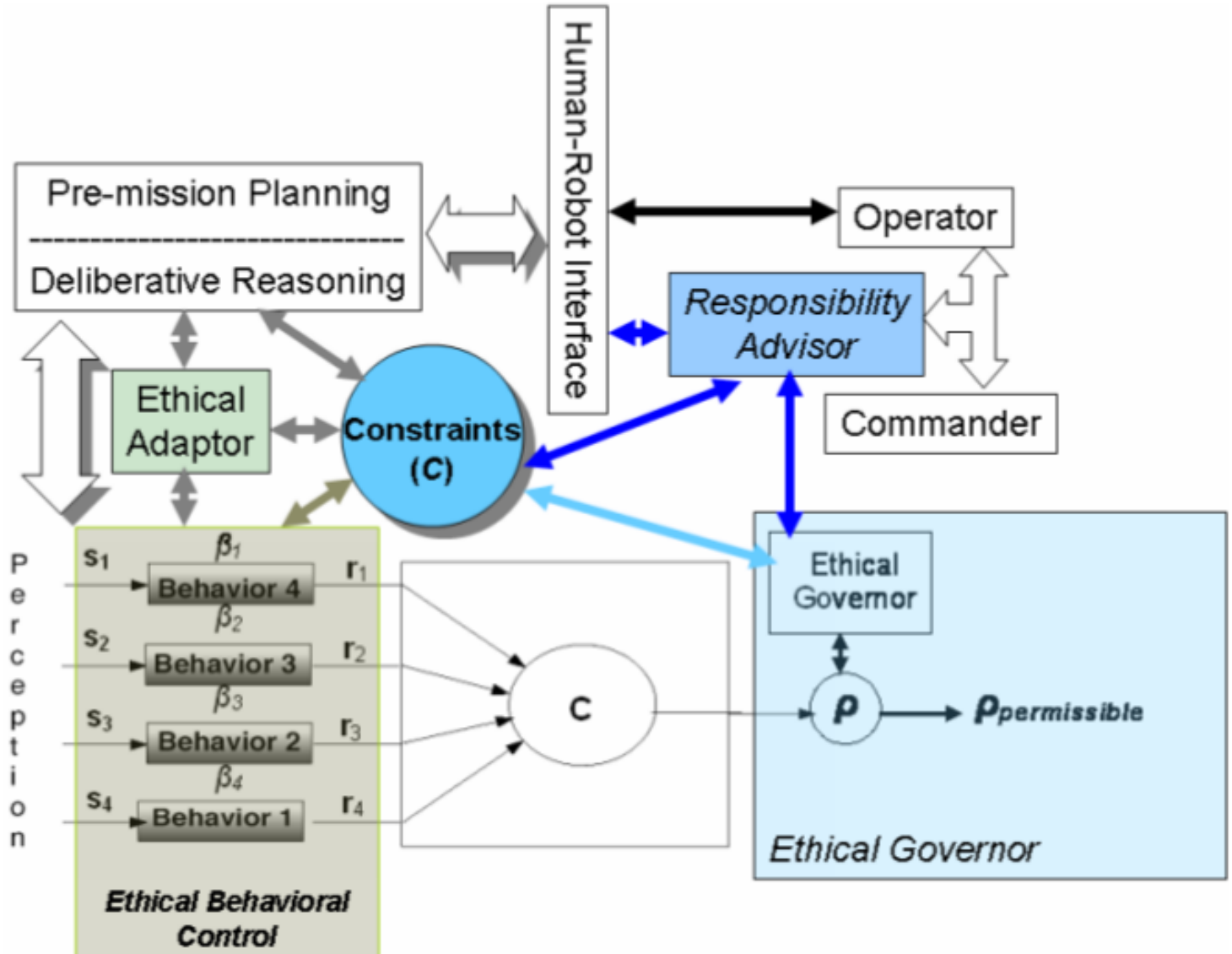
R. Arkin. *Governing lethal behavior in autonomous robots.* CRC Press, 2009.

Drawbacks

- No explicit representation of the ethical concepts
- No genericity
- The agent cannot distinguish its own ethics from its operationnal procedures

In the literature

Ethics by design: Arkin's architecture



In the literature

Ethics by design: example of ethical procedures

```
1: while lethal response authorized, military necessity exists, responsibility assumed do
2:     if target is sufficiently discriminated then
3:         if C_forbidden satisfied then {no violation of LOW exists}
4:             if C_obligate is true then {lethal response required by ROE}
5:                 optimize proportionality using principle of double intention
6:                 engage target
7:             else {no obligation/requirement to fire}
8:                 do not engage target
9:                 continue mission
10:            end if
11:        else {permission denied by LOW}
12:            if previously identified target surrendered or wounded then
13:                notify friendly forces to take prisoner
14:            else
15:                do not engage target, report and replan
16:                continue mission
17:            end if
18:        end if
19:    end if
20:    Report status
21: end while
```

In the literature
Ethics by learning

"A paradigm of case-supported principle-based behavior (CPB) is proposed to help ensure ethical behavior of autonomous machines."

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.

Benefit

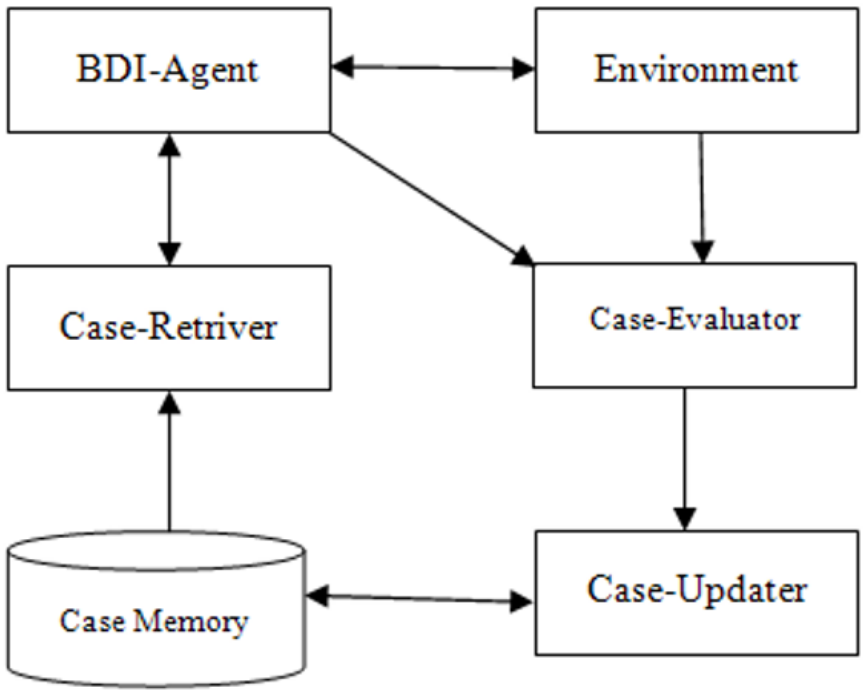
- Generic approach
- Explicit representation of (some) ethical principles

Drawbacks

- No explicit representation of all ethical concepts
- Classical problems of learning (over-/underfitting) facing new circumstances

In the literature

Ethics by design: Honavar's architecture



Basic concepts

An action is defined by a set of fulfillment over a set of duties (d_i), such as Readiness, Harm, Autonomy, Non-Internaction, Possible Good, and so on.

General form of an ethical principle

$$p(a_1, a_2) \leftarrow \begin{array}{c} \Delta d_1 \geq v_{1,1} \wedge \ldots \wedge \Delta d_m \geq v_{1,m} \\ \vee \\ \ldots \\ \vee \\ \Delta d_n \geq v_{n,1} \wedge \ldots \wedge \Delta d_m \geq v_{n,m} \end{array}$$

In the literature
Ethics by reasoning

"We need other kind of more intricate mental models, able to support moral reasoning capabilities."

H. Coelho and A.C. da Rocha Costa. On the intelligence of moral agency. Encontro Portuguees de Inteligencia Artificial, pages 12–15, October 2009

Benefit

- Generic approach
- Ease interaction between artificial agents and humans
- Able to infer a justification in given situation

Drawbacks

- Sometime only focus on deontic logic (Bringsjord and Govindarajulu, 2013)

In the literature
Ethics by reasoning

Models for Artistotelian ethics (Ganascia, 2007)

$$
\begin{array}{rcl}
act(P, G, A) & \leftarrow & action(A), person(P), goal(P, G), solvegoal(P, G, A), notunjust(A). \\
 & \leftarrow & action(P, G, A), action(P, G, AA), A \neq AA. \\
 & & \\
just(A) & \leftarrow & worstcons(A, C), worstcons(AA, CC), worse(C, CC), \mathbf{not} unjust(A) \\
unjust(A) & \leftarrow & worstcons(A, C), worstcons(AA, CC), worse(CC, C), \mathbf{not} just(A). \\
 & & \\
notworstcons(A, C) & \leftarrow & cons(A, C), cons(A, CC), worse(CC, C), \mathbf{not} worse(C, CC) \\
worstcons(A, C) & \leftarrow & cons(A, C), notnotworstcons(A, C).
\end{array}
$$

Models for Kantian ethics

- Powers, 2006
- Ganascia, 2007 (ASP)

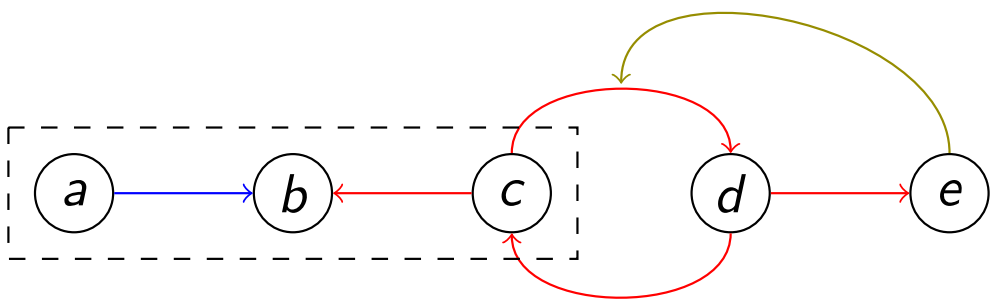
Models for the Doctrine of Double Effect

- Pereira and Saptawijaya, 2007 (ProLog)
- Berreby, Bourgne and Ganascia, 2016 (ASP)

In the literature
Ethics by argumentation

Formal argumentation (basics)

- Arguments $\mathcal{A} = \{a, b, c, d, e\}$
- Attack relationships $R_i = \{(a, b), (c, b), (c, d), (d, c), (d, e)\}$
- Admissible arguments (without conflict and that defend themselves)
- Acceptability semantics (special sets of admissible arguments)
- Preferences $a \succ b \succ c \succ d \succ e$ to constrain attacks
- Dialectical frameworks to express attacks and supports
- Meta-argumentation to express attacks on attacks



Structure of arguments may be grounded by a logical theory

In the literature
Ethics by argumentation

"[...] reasoning of this sort is required [in]: law, medicine, politics and moral dilemmas, and an everyday situation."

K. Atkison and T. Bench-Capon. Abstract argumentation and values. Argumentation in Artificial Intelligence, chapter 3, 2009

Value-based argumentation

- 'In the context C, the plan P achieves the goal G which promotes the value V"
- A v function $\mathcal{A} \rightarrow \mathcal{V}$ associates a value to arguments
- Characterizes arguments w.r.t. all preference ordering on values

Benefit

- High-level approach which is very understandable
- Extensions for several values, and to demotions

Drawbacks

- No associated logic, nor clear principles

Conclusion
Briefly and shortly

A set of diverse frameworks

- Many works on norms and deontic logic
- Many works on value and preference reasoning
- Some works on principles and maxims formalization

However

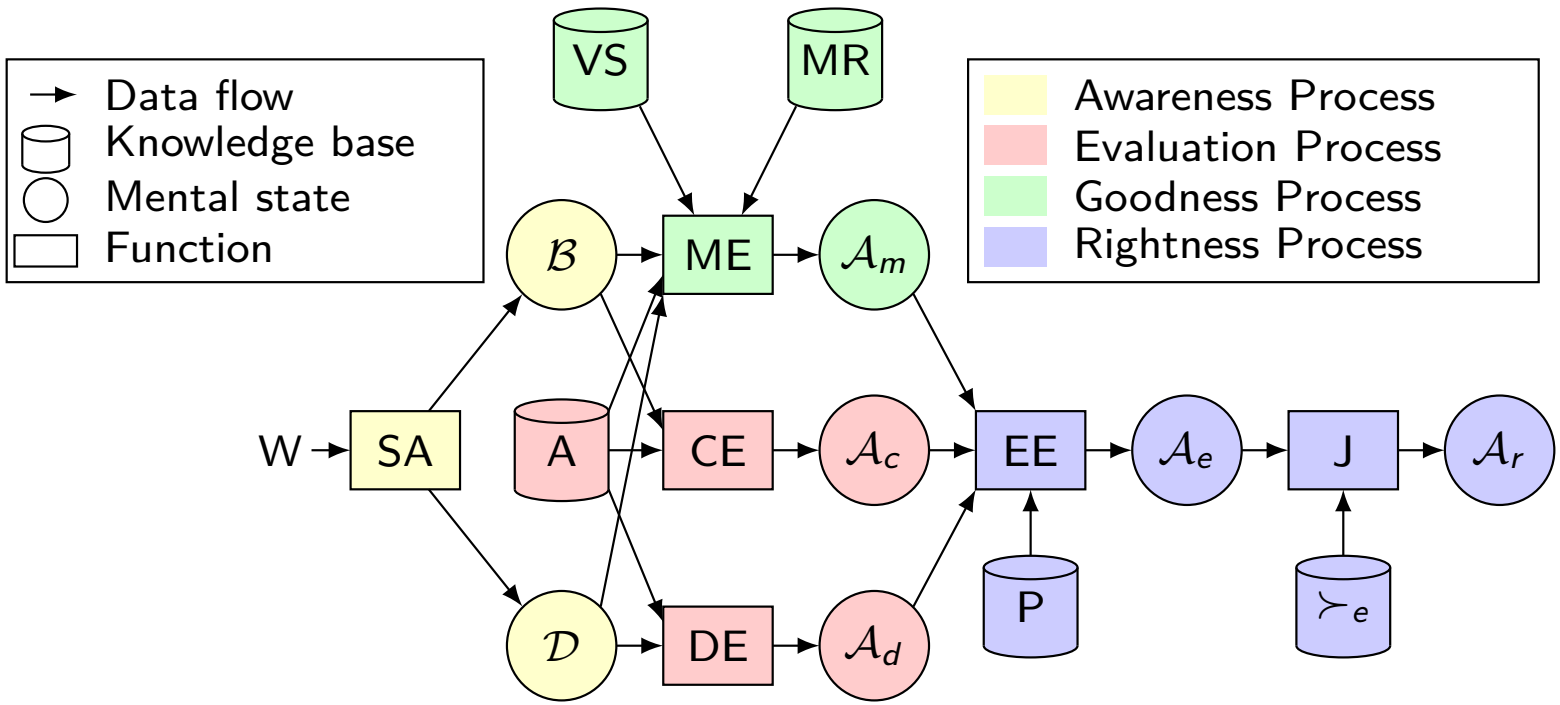
- Few works on value characterization
- Very few works on judgment procedures
- No work on collective ethical conflicts

A judgment architecture

Ethical judgment architecture

Overview

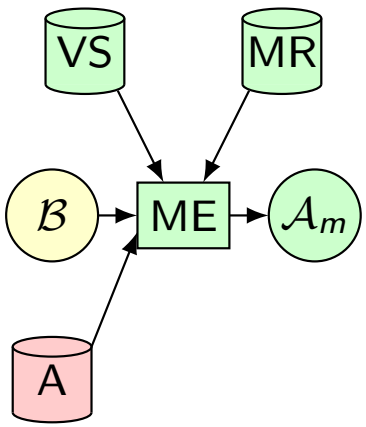
N. Cointe, G. Bonnet and O. Boissier. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. 15th AAMAS, pp. 1106–1114, 2016.



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

$\mathcal{O}$ is an ontology used to define actions, values and goals

Ethical judgment architecture
Goodness process



VS = Values Supports
MR = Moral Rules

Examples of value supports and moral rules

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

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

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

Ethical judgment architecture
How to capture different moral approaches?

Virtuous approaches

- General rules based on values
- ex.: Being generous is moral $= \langle \_, \langle \_ , generosity \rangle, moral \rangle$

Deontological approaches

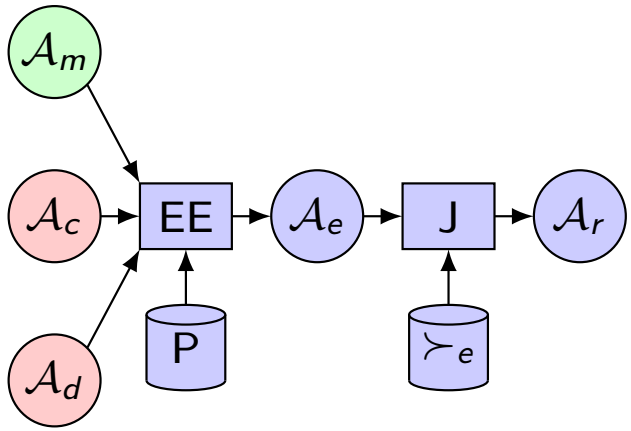
Specific rules concerning actions and the agent role is a belief about the situation
"Journalists should deny favored treatment to advertisers, donors or any other special interests and resist internal and external pressure to influence coverage"

Consequentialist approaches

Both general and specific rules concerning states and consequences
"Every physician must refrain, even outside the exercise of his profession, any act likely to discredit it"

Ethical judgment architecture

Rightness process



Examples of ethics

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 (Thomas Aquina's theory)

$$\text{P3} \succ_e \text{P2} \succ_e \text{P1}$$

Ethical judgment architecture

Judgment by an example

Benjamin Constant's Dilemma

An agent A knows that an agent B hides in his house in order to escape an agent C. C asks A where is B to kill him, threatening to kill A in case of non-cooperation.

A's moral rules

- "Prevents murders is moral"
- "Lying is immoral"

A's desires

- "Avoid any trouble"

A's possible actions

- Tell C the truth (satisfy a moral rule and a desire)
- Lie (satisfy a moral rule)
- Refuse to answer (satisfy all moral rules)

Ethical evaluation of agent A's actions

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

$$P3 \succ_e P2 \succ_e P1$$

Ethical judgment architecture
Usage

Judging it's own behavior
Distinguish the rightfull actions to execute

Blind judgment of other agents
Compare the behavior of another agent to judge if its conduct is fair or not

Partially informed judgment of other agents
Consider the mental states of the others before judging (theory of mind)

Fully informed judgment of other agents
Compare the bevahior of another agent w.r.t. to a role it should play

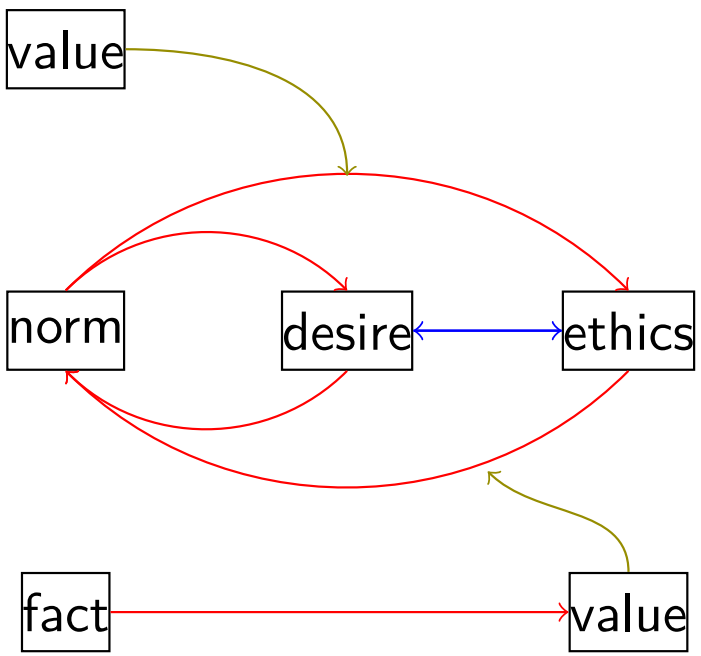
An ethical argumentation framework

Ethical practical argumentation framework

A hierarchical model

Specificities

- Arguments are facts, desires, norms, principles, values
- Each stratum is model by a logic that generates arguments
- Relationships are inspired from Jonathan Haidt's insights



Underlying logics
Hierarchical logic

Logics that infer formulas that are valid if some formulas from nested logics are valid too

Formal definition

$$\Lambda_s = (\mathcal{L}_s, (\Delta_n)_{n\in[|0,N-1|]}, (\mathbb{T}_s, \vdash_s), (\mathbb{V}_s, \models_s), \Gamma_s, \mathcal{C}_s)$$

1. $\mathcal{L}_s$ a hierarchical modal logic language
2. $(\Delta_n)_{n\in[|0,N-1|]}$ a set of hierarchical logics $n \in [0, N-1]$.
3. $(\mathbb{T}_s, \vdash_s)$ an inference structure, and $(\mathbb{V}_s, \models_s)$ a validity structure,
4. $\mathbb{V}_s = (\mathcal{W}, \mathcal{R}, I)$ the Kripke model, and $w \in \mathcal{W}$ the current world
5. $\Gamma_s = \{\phi \in \mathcal{L}_s | \emptyset \vdash_s \phi\}$ a set of axioms
6. $\mathcal{C}_s : 2^{\mathcal{L}} \to 2^{\mathcal{L}}$ a function such that:

$$\forall \sigma \subsetneq \mathcal{L}_s : \mathcal{C}_s(\sigma) := \{\phi \in \mathcal{L}_s | \sigma \cup \Gamma_s \vdash_s \phi \text{ and } w \models_s \phi\}$$

Example

$\neg\Diamond_I[|\vdash_J \phi]$ means "It is forbidden according to I that the valid formula ϕ in J holds"

Underlying logics
Three kinds of logic

Ground logics

- Λ_Σ an epistemic logic
- Λ_Π a hypothetic consequence logic (actions and composition of actions)

Value logic

- Λ_ν a logic that orders values with respect to the context

Pratical logics (all of them are action logics)

- Λ_δ an action logic based on goals (techno-scientific stratum)
- Λ_λ a deontic logic (jurdical-political stratum)
- Λ_{M_i}, $i \in [0, K]$ a set of ethical logics (Kant, DDE, and so on)

Epistemic, value and all practical logics generate arguments

Argumentation framework
Relationships between arguments

Three kinds of relationships: rebuttals, undercuts and defenses

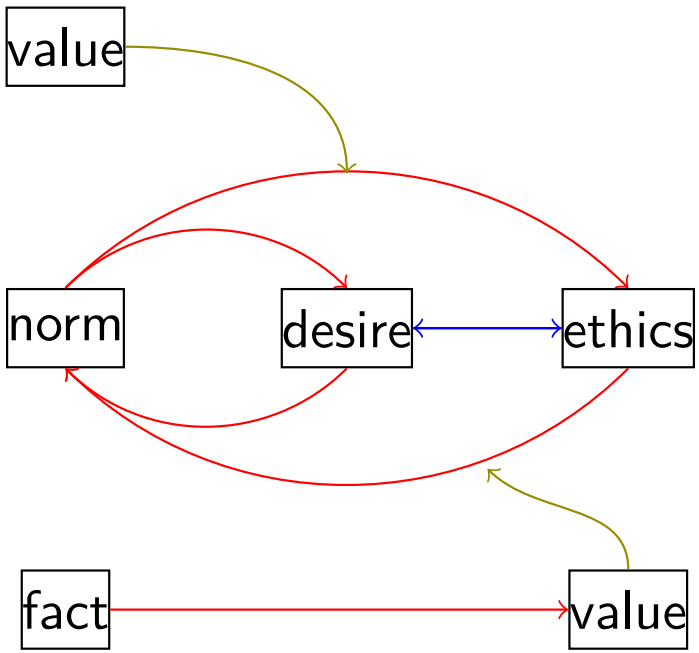
Relationships

- Epistemic arguments undercut premisses of other arguments
- Higher arguments (stratum) rebut lower one by contradiction
- Higher arguments (stratum) defend lower one when they agree
- Value preference arguments undercut attack relationships

Examples of relations

- "It is forbidden to steal" attacks "I desire to steal"
- "It is ethical to steal when we are hungry" defends "I desire to steal"
- "I am not hungry" undercuts "It is ethical to steal when we are hungry"

Semantics
How to characterize the rightfull actions?



A rank semantic

A action supported by several argument is righter than the others.

A classical semantic

A conflict-free set of arguments that attacks all other arguments is robust.

Insights

- The weight of an argument is $W(A) = W(A)_{defences} - W(A)_{attacks} + 1$ (recursive)
- An argument defeat another one it attacks if its weight is higher or if it undercuts it
- Rightfull actions belongs to stable extensions w.r.t. defeats

Conclusion

Conclusion
Summary

Implicit approaches

- Simplier to implement, but not generic
- Is there a way to check if an agent follows a given ethics?

Explicit approaches

- Generic, but time consumming
- We lack complete formal characterization of values

Argumentative approaches

- High-level abstraction that can embed other approaches
- Semantics are still outside the reasoning

Conclusion
Open questions

Situation assessment

- How to infer beliefs from perceptions?
- Is there an ethics of situation assessment?

Multi-agent reasoning

- How to take the others' values and ethics into account?
- Is there a hierarchy over the others' ethics?

Anytime reasoning

- How to reason when time is limited?
- Are some kind of reasoning associated to a priority?