Introduction to ethics in artificial agents

Grégory Bonnet

Normandie Université – GREYC

June 27th 2023

There are certain tasks which computers *ought* not be made to do, independant of whether computers *can* be made to do them.

Joseph Weizenbaum
Computer Power and Human Reason
From Judgement to Calculation
W. H. Freeman, 1976.

Outline

Introduction

Towards an European legal framework : the AI Act

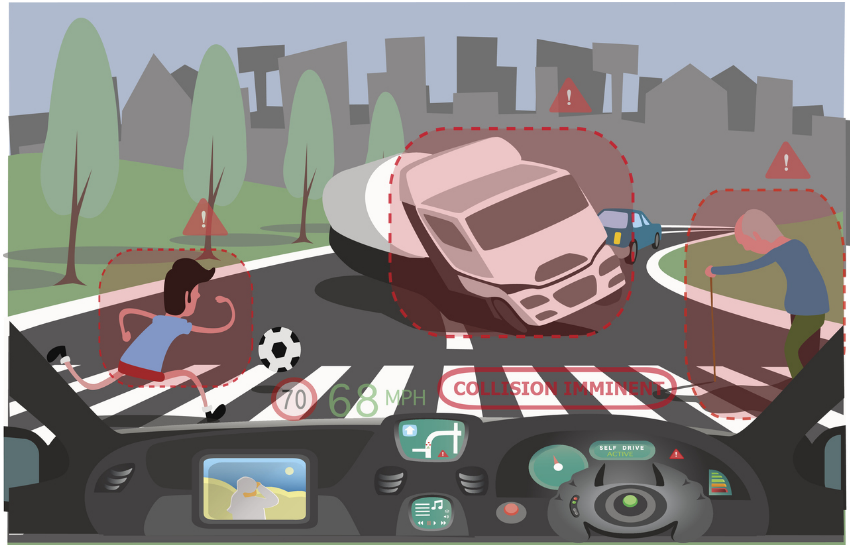
Can (and how) we program ethics ?

Ethical agent architectures – A review

Example – A BDI architecture for ethical judgment

The famous Trolley dilemma (for autonomous vehicles)

Image : https://medicalfuturist.com/



Responsible Artificial Intelligence

A pluridisciplinary domain

Lines of research

1. integrity of researchers, designers and developers
2. study of the socio-cognitive implications of artificial intelligence
3. implementation of ethical reasoning skills

Several initiatives, reports and legislative developments

- IEEE Global Initiative on Ethics of Autonomous and Intelligent System (2018)
- EU « Ethics guidelines for a trustworthy AI » (2019)
- EU Resolution on a civil liability regime for artificial intelligence (2020)
- EU Resolution on ethical aspects of artificial intelligence (2020)
- EU Resolution interpretation and application of international law for AI systems (2021)
- EU Artificial Intelligence Act (2021)
- → Voted and adopted by European Parliament on June 14th 2023

Artificial Intelligence Act

Definition of AI systems

Artificial Intelligence Systems (AIS)

AIS means software that :

- is developed with one or more of the techniques and approaches listed in Annex I
- can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with

Annex I

1. Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning
2. Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems
3. Statistical approaches, Bayesian estimation, search and optimization methods

Artificial Intelligence Act

Two kinds of AIS

- Forbidden AIS
  - Vulnerabilities exploitation (due to age, disability, subliminal techniques)
  - Physical person classification according to social or personnal criteria
  - Real-time biometric identification in public environment (outside « emergency »)
- High-risk AIS (identified in Annex II and III)
  - Authorized biometric systems, truth detection
  - Energy and grid management (road traffic, water, electricty, gaz, heat)
  - Education, credit, social prestation and public services access
  - Law and justice (risk management, predictive justice, law application and interpretation)
  - Migratory flows management

Requirements for high-risk AIS

- Continuous technical documentation and risk analysis
- Input and output data record-keeping
- Transparency and provision of information to users
- Mandatory effective human oversight
- Conformity assessment procedure and registration obligations
- → Fines between 250 000 and 30 000 000 euros

Artificial Intelligence Act

Grounded by ethical recommentations

UE Resolution on ethical aspects of artificial intelligence, robotics and related technologies

Any new regulatory framework for AI consisting of legal obligations and ethical principles for the development, deployment and use of artificial intelligence, robotics and related technologies should :

- fully respect the Charter and thereby
- respect human dignity, autonomy and self-determination of the individual, prevent harm, promote fairness, inclusion and transparency, eliminate biases and discrimination, including as regards minority groups,
- comply with the principles of limiting the negative externalities of technology used, of ensuring explainability of technologies, and of guaranteeing that the technologies are there to serve people and not replace or decide for them, with the ultimate aim of increasing every human being's well-being

Can (and how) we program ethics ?

Ethical artificial agent

Why ?

The emergence of questions about the responsibility and governance of algorithms raises ethical issues. It follows that providing tools for modeling, programming and integrating ethics into the decision-making mechanisms of artificial systems (agents) may be of interest.

Precision

Programming an artificial agent to be ethical does not mean that this agent is ethical, but that its decisions can be judged as ethical by an external human observer : it is therefore a simulation of ethics (just as we simulate emotions or decision making).

An ethical artificial agent should be able :

- to represent and reason on ethical factors
- to judge his actions and the actions of others in terms of morality and ethics
- to take into account the multi-agent dimension of ethics

Ethics and morality

Deleuze, 1990

Morality is a set of binding rules of a special kind, which consists in judging actions and intentions by relating them to transcendent values (it's right, it's wrong, etc.) ; ethics is a set of optional rules which evaluate what we do and say according to the mode of existence it implies.

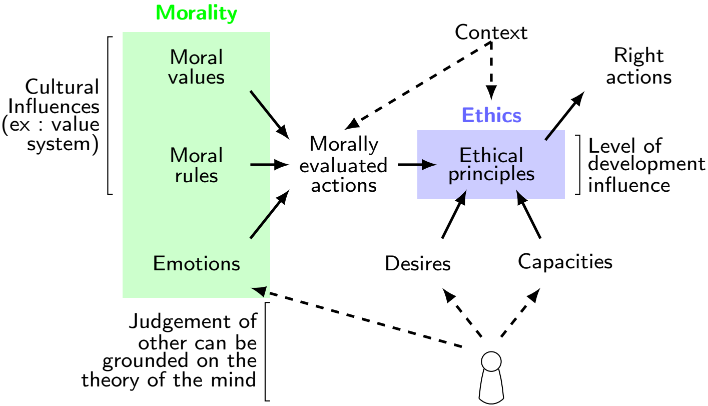
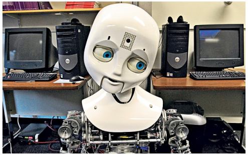
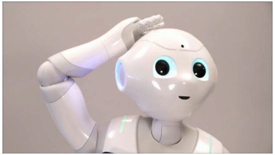

Morality
Cultural Influences (ex : value system)
Moral values
Moral rules
Emotions
Context
Morally evaluated actions
Ethics
Ethical principles
Right actions
Level of development influence
Desires
Capacities
Judgement of other can be grounded on the theory of the mind

Figure – N. Cointe, PhD. Thesis, 2017

Ethical issues in autonomous agents and muli-agent systems

We set aside machine learning ethical issues

1. Data bias and learning bias
2. Responsibility (designers, data provider, trainers, users, etc.)
3. Anchoring effects and minimization of personal situations

Autonomous agent issues

- value-based decision making
- trust in emotional agents
- causal responsibility

Multi-agent issues

- judging other agents
- other's harm avoidance, non discrimination
- fairness and equity in collective decision-making

Must machine ethics be human ethics ?

> *"By providing a framework for identifying and critiquing assumptions about what a 'good' computer is, a study of android arete provides focus and direction to the developmentof future computational agents."*

Kary G. Coleman. Android arete : Toward a virtue ethic for computational agents. Ethics and Information Technology 3(4), pages 247-265, October 2001

Agentive	**Social**	**Environmental**	**Moral**
Self-movement	Communicativity	Thirft / Moderation	Non-maleficience
Self-regulation	→ Explicit responsiveness	Tidiness	→ Freedom from biais
Autonomy	→ Implicit responsiveness	Obedience	→ Safety
Goal-orientation	Veracity	Safety	→ Vigilance
Intelligence	Accessibility	Identifiability	Beneficence
→ Reliability	→ Character	Openness	Obedience
→ Efficiency	Respectfulness	Proper inquisitiveness	Accessibility
→ Accuracy	Reliability		Self-protection
Flexibility	Flexibility		Vulnerability
→ Reactivity	Adaptativity		
→ Adaptativity	Reactivity		
Autopoiesis			

- (M) Accessibility : having external representation of moral qualities
- (M) Vigilance : disposition to block human actions that have unintended consequences
- (E) Thrift : sparing used of resources
- (E) Tidiness : disposition to clean up after self

How to program ethics ?

1. Which definition of ethics to consider ?

   There are choices and implicits within a given definition to be aware of.

2. Which modelling and resolution approach to chose ?

   Quantification versus qualification ; specification versus machine learning.

3. Which ethical concept to model and to make explicit ?

   Values, rules, emotions, causal responsibility, etc.

4. How to evaluate an ethical artificial agent ?

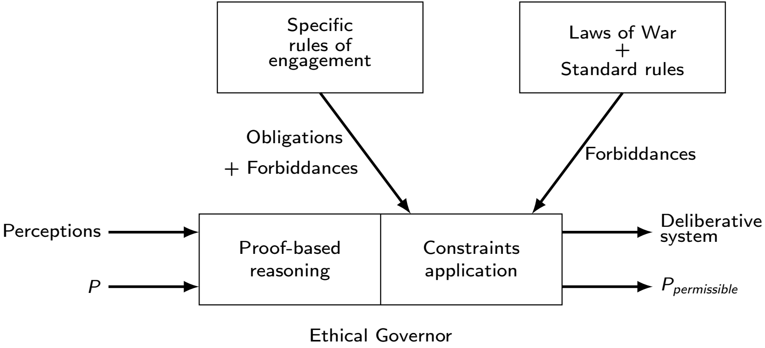
   Do we have the same ethical requirement for a machine than for ourselve ?

Ethical agent architectures – A review

Ethical agent architectures – Procedural approaches

> *"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

- Lack of genericity
- No distinction between ethics and operational procedures

Ethical agent architectures – Procedural approaches

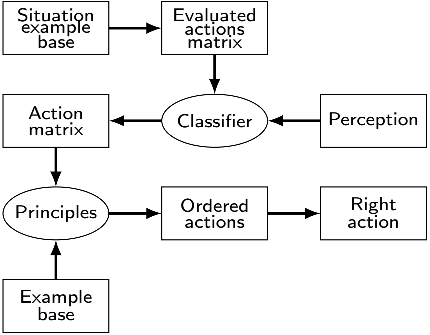
Exemple of procedure from (Arkin, 2009)

```
while lethal response authorized, military necessity exists, responsibility assumed do
    if target is sufficiently discriminated then
        if $C_{forbidden}$ satisfied then {no violation of LOW exists}
            if $C_{obligate}$ is true then {lethal response required by ROE}
                optimize proportionality using principle of double intention
                engage target
            else {no obligation/requirement to fire}
                do not engage target
                continue mission
            end if
        else {permission denied by LOW}
            if previously identified target surrendered or wounded then
                notify friendly forces to take prisoner
            else
                do not engage target, report and replan
                continue mission
            end if
        end if
    end if
    report status
end while
```

Ethical agent architectures – Learning approaches

"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, 42(4) :324-331, 2015.



Merits

- Generic approach
- Explicit representation of certain ethical principles

Drawbacks

- No representation of all concepts
  (e.g. responsibility, reasoning, elicitation)
- No policy evaluation
- Tied to machine learning limits

Ethical agent architectures – Learning approaches

Exemple of ethical principle from (Anderson and Anderson, 2015)

Action intrinsic evaluation

An action is associated to a set a promotion measure according to a set of duties (values) d_i.

General form of an ethical principle

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

Ethical agent architectures – Deep learning approaches

"Systems need the ability to anticipate and understand the norms of the different communities in which they operate [by] focusing on [...] descriptive ethics."

N. Lourie, R. Le Bras and Y. Choi. SCRUPLES : A corpus of community ethical judgments on 32,000 real-life anecdotes. 35th AAAI Conference on Artificial Intelligence, pp. 13470-13479, 2021.

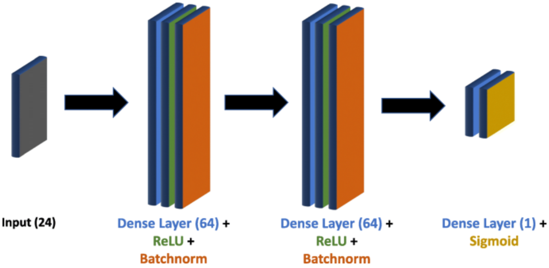

Input (24)
Dense Layer (64) + ReLU + Batchnorm
Dense Layer (64) + ReLU + Batchnorm
Dense Layer (1) + Sigmoid

Figure – Deep learning for MIT moral test (Wiedeman, 2020)

Merits

- Generic approach
- Context assessment
- Several corpus of moral situations

Drawbacks

- No explicit representation of ethical concepts
- No reasoning → statistical correlations
- Corpus do not talk about sequential strategies

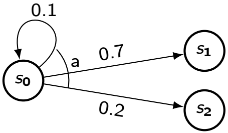
Ethical agent architecutre – Decision theoretic approaches

> *"Formally, this is expressed as an optimization problem with a set of constraints for the task and a constraint for the ethical framework."*

J. Svegliato, S. Nashed and S. Zilberstein. Ethically compliant sequential decision making. AAAI 2021.

Moarkov Decision Processes + Constraints

- Typology : moral, amoral, imoral, optimal policies
- Evaluation based on the price of morality
- Captures : Divine Command Theory, Prima Facie Duties, éthique des vertus



Merits

- Generic approach
- Convergence and optimality proofs
- Constraint axiomatics

Drawbacks

- Difficulty to express non-linear constraints
- Implicit notion of causality (classical limits of MDPs)
- No distinction between morality and ethics

Ethical agent architecutre – Decision theoretic approaches

Example of *prima facie duties*

- Δ a set of duties
- $\phi : \Delta \times S \to \mathbb{R}^+$ a penalty function
- $\tau \in \mathbb{R}^+$ a tolerance threshold

Ethical principle

$$\rho_\Delta(\pi) = \sum_{s \in S} d(s) J^\pi(s) \leq \tau$$

$$J^\pi(s) = \sum_{s' \in S} T(s, \pi(s), s')[\sum_{\delta \in \Delta_{s'}} \phi(\delta, s') + J^\pi(s')]$$

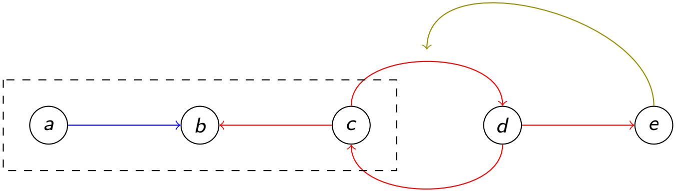
Informally

A policy π is moral if the sum of the cumulative expected penalty $J^\pi(s)$ starting from the state s is less than the tolerance τ.

Ethical agent architectures : argumentative approaches

Basic concept in formal argumentation

- Arguments $\mathcal{A} = \{a, b, c, d, e\}$
- Attack relationship $R_i = \{(a, b), (c, b), (c, d), (d, c), (d, e)\}$
- Admissible arguments (conflict-free and defending themselves)
- Acceptability semantics (special set of admissible arguments)
- Preference (ex. $a \succ b \succ c \succ d \succ e$) constraining the attack relationship
- Dialectical frameworks that express both attacks and supports
- Meta-argumentation expressing attacks on attacks



Ethical agent architectures – Argumentative approaches

Value-based argumentation frameworks

“[...] 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 frameworks (VAF)

- "In the context C, the plan P achieves the goal G which promotes the value V"
- A function $v : \mathcal{A} \rightarrow \mathcal{V}$ that associates to each argument a value
- Admissible arguments are characterized base on preferences (credulous or sceptical acceptance)

Merits

- High-level mode which is understandable by non-experts
- Extension to deal with multiple values, demoted values, probabilities, etc.

Drawbacks

- No grounded logics behind the arguments
- No ethical principles which structures the attacks

Ethical agent architectures – Declarative approaches

> *"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

Some references

Works from Berreby, Bringsjord, Cointe, Ganascia, Lorini, Peireira, Sarmiento . . .

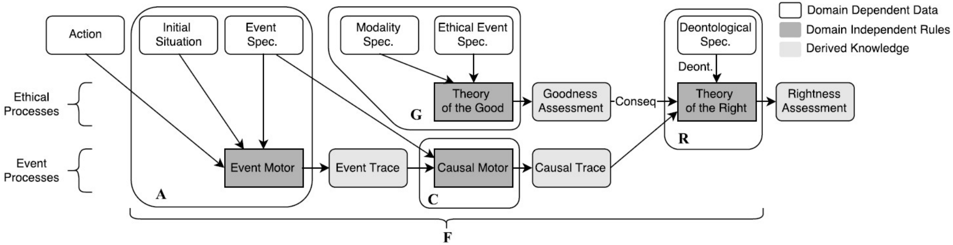


Figure – An ethical modular framework (Berreby, 2018)

Merits

- Generic approach
- Specification « easy » to read for non-expert
- Decisions are interpretable (i.e. proofs)

Drawbacks

- Complexity tied to the grounding logics
- Difficulties to express uncertainty

Ethical agent architectures – Declarative approaches

Exemple of ethical principles in Prolog and ASP

Modeling morality

Associating valuations to actions and states.

Aristotelian ethics (Ganascia, 2007)

```
act(P, G, A)        :- action(A), person(P), goal(P, G), solve(P, G, A), not unjust(A).
                    :- action(P, G, A), action(P, G, AA), A ≠ AA.
just(A)             :- worstcons(A, C), worstcons(AA, CC), worse(C, CC), not unjust(A).
unjust(A)           :- worstcons(A, C), worstcons(AA, CC), worse(CC, C), not just(A).
notworstcons(A, C) :- cons(A, C), cons(A, CC), worse(CC, C), not worse(C, CC).
worstcons(A, C)     :- cons(A, C), not notworstcons(A, C).
```

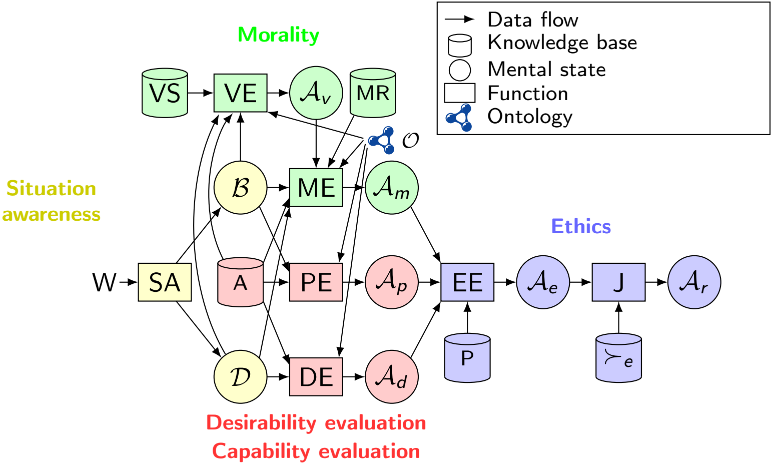
Doctrine of double effect (Berreby, 2018)

```
imp(dde1,A):- act(A), bad(A,X,M).
imp(dde2,A):- act(A), cons(S,A,T1,E1), cons(S,E1,T2,E2),
              bad(E1,X1,M1), good(E2,X2,M2).
imp(dde3,A):- imp(benefitsCosts,A).
per(dde,A) :- act(A), not imp(dde1,A), not imp(dde2,A), not imp(dde3,A).
```

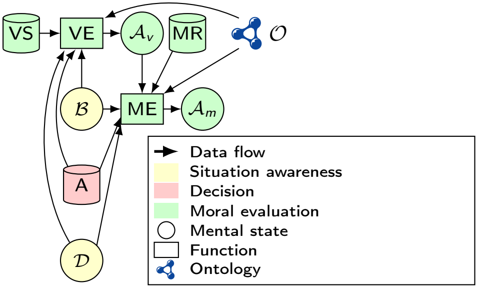
Example – A BDI architecture for ethical judgment

Architecture overview

Joint work which Nicolas Cointe and Olivier Boissier



Value model



Value support $\langle a, w, w', v, sv \rangle \in VS$

- $a \subseteq A$: a set of actions
- w : initial situation
- w' : hypothetic situation (consequencies)
- $v \in \mathcal{O}$: value
- $sv \in \mathcal{O}$: evaluation support

Examples

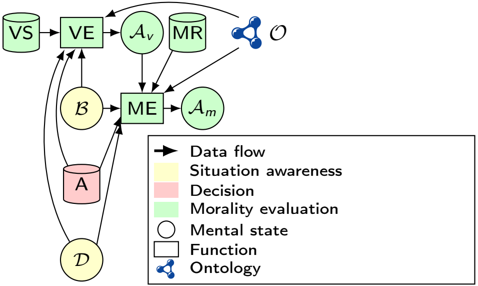
- Making an action which makes a poor agent a non-poor agent promotes the value generosity

$$\langle any, poor(\mathrm{a}), \neg poor(\mathrm{a}), generosity, promote \rangle$$

- Generosity is a subvalue of benevolence

$$subvalue(generosity, benevolence)$$

Morality evaluation



→ Data flow
🟨 Situation awareness
🟥 Decision
🟩 Morality evaluation
○ Mental state
□ Function
Ontology

Moral rules $\langle a, w, w', vc, mv\rangle \in MR$

- $a \subseteq A$: a set of actions
- w : initial situation
- w' : hypothetic situation (consequencies)
- vc : promoted and demoted values
- $mv \in \mathcal{O}$: morality evaluation

Examples

- Virtue : "Making a generous action is highly moral"

$$\langle any, \top, \top, \{\langle generosity, promote, min\rangle\}, highly\ moral\rangle$$

- Deontology : "Giving something to a poor agent is moral"

$$\langle \{give(\mathrm{a})\}, poor(\mathrm{a}), \top, \emptyset, moral\rangle$$

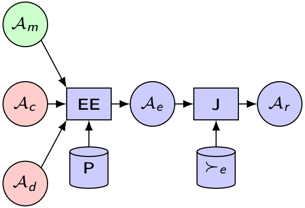
- Consequentialism : "Making an action which makes possible other highly moral action is moral"

$$\langle any, impossible(a'), possible(a') \wedge goodness(a', s', mr_x, highly\_moral), \emptyset, moral\rangle$$

Ethical evaluation

Judging an action

An action est permissible (or not) with respect to a principle and a tuple $\langle \mathcal{A}_m, \mathcal{A}_c, \mathcal{A}_d \rangle$. Judgment allow to build the set $\mathcal{A}_r$ of the right actions, i.e. which satisfy the best the ordered set of principles.



Examples

- P1 If an action is possible, desirable and moral with respect to least one moral rule, then it is a right action.
- P2 If an action is not immoral with respect to all moral rules, then it is a right action.
- P3 If an action satisfies the doctrine of double effect, then it is a right actions.

$$P1 \succ_e P3 \succ_e P2$$

Judging other agents

From one-shoot judgment to continuous judgment

To judge

- Evaluating a behavior (a set of actions)
- With respect to a set of beliefs
- Producing a belief to quality an observed behavior

Behavior

A behavior $b_{a_j,[t_0,t]}$ of agent a_j on timesteps $[t_0, t]$ is the set of actions α_k that a_j made between t_0 and t such that $0 \leqslant t_0 \leqslant t$.

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

Judging other agents

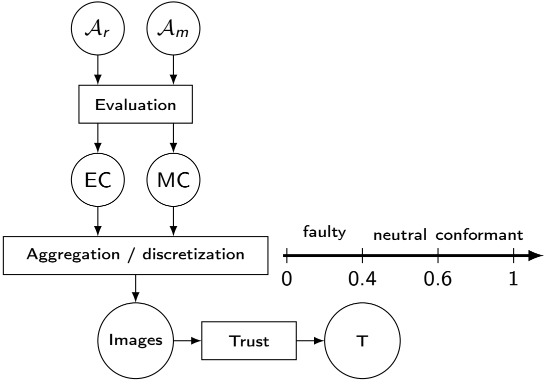
To produce an image

Kinds of judgments

- Blind judgment (only based on the judge agents beliefs, values, moral rules and principles)
- Partly informed judgment (based on beliefs about the judged agent beliefs, values, moral rules or principles)
- Fully informed judgment

Kinds of aggregations

- on a set of agents
- on a subset of moral rules
- on ethics



Build trust in the ethics or morality

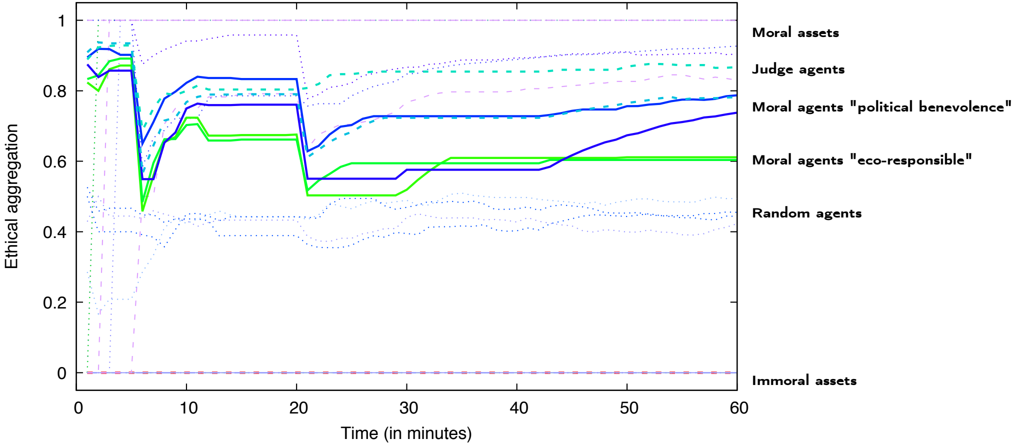
We can define epistemic actions (which produce beliefs instead of world's changes)

$$ethical\_trust(a_j, a_i) \text{ or } moral\_trust(a_j, a_i, ms, mt)$$

Ethics of trust

- forgiving : building trust only based on recent judgments
- intransigent : trust only the agents which behavior is judged as ethical
- Is is moral to be intransigent with agents on which human lives rely

Experiments – Evaluating the judgment process



Conclusion

Conclusion

AI Act adoption

- Towards an European regulation framwork
- Ethical issues for autonomous agents are still important to deal with :
  - Mono-agent – Value-based decision making, causal responsibility, epistemic responsibility, trust
  - Multi-agent – Judging others, non-discrimination, fairness, equity

Ethical architectures

- Be intelligible and readable by humans
- Use modular architectures
- Emphasize qualitative rather than quantitative models
- Take into account the subjectivity of models
- Take into account the multiplicity of agents and humans

Last words

In the final analysis, it is the human being, by observing these models and the decisions made, who can say whether or not they are ethically sound. However, we must remain vigilant about our own subjectivity.

Some references

Kary G. Coleman
Android arete : Toward a virtue ethic for computational agents
Ethics and Information Technology 3(4), 2001.

Fiona Berreby
Models of Ethical Reasoning
PhD. Thesis, Sorbonne Université, 2018.

Nicolas Cointe, Grégory Bonnet and Olivier Boissier
Ethical Judgment of Agents' Behaviors in Multi-Agent Systems
15th International Conference on Autonomous Agents and Multiagent Systems, 2016.

Luís Moniz Pereira and Ari Saptawijaya
Programming Machine Ethics
Studies in Applied Philosophy, Epistemology and Rational Ethics 26, 2016.

Justin Svegliato, Samer Nashed and Shlomo Zilberstein
Ethically Compliant Sequential Decision Making
AAAI Conference on Artificial Intelligence, pp. 11657–11665, 2021.

Robert Trappl (editor)
A Construction Manual for Robots' Ethical Systems : Requirements, Methods, Implementations
Cognitive Technologies, Springer, 2015.