Ethics and autonomous agents

Grégory Bonnet

Normandie University – GREYC Lab

November 7th 2019

emergence of responsible artificial intelligence

ents of ethics and morals

tectures for ethical agents

propositions of the ETHICAA project

thics & Health

The emergence of responsible artificial intelligence

ous trolley dilemma (updated with autonomous vehicles)

nous agents interacting with human being

oseph Weizenbaum

to insure that an autonomous agent :

will not cause « harm » to other agents (humans and machines)

decide according to cultural, compassionate and ethical factors

→ beyond the law, subjective and plural

ible Artificial Intelligence

ed domain

ponsible Artificial Intelligence

to think the integrity and responsibility of researchers, designers, and programmers

to study the socio-cognitive implications of artificial intelligence

to study how to implement ethical reasoning capabilities

y initiatives and reports

IEEE Global Initiative on Ethics of Autonomous and Intelligent System

Ethics guidelines for a trustworthy AI

CERNA reports on ethics of research in robotics and machine learning

CERNA report « Numérique & santé : quels enjeux éthiques pour quelle régulation ? »

CNIL report « Comment permettre à l'Homme de garder la main ? »

Aligned Design

Initiative on Ethics of Autonomous and Intelligent System (2017)

king groups

Embedding values into autonomous intelligent systems

Methodologies to guide ethical research and design

Safety and beneficience of artificial general intelligence

Personal data and individual access control

Reframing autonomous weapons systems

Economics and humanitarian issues

Law

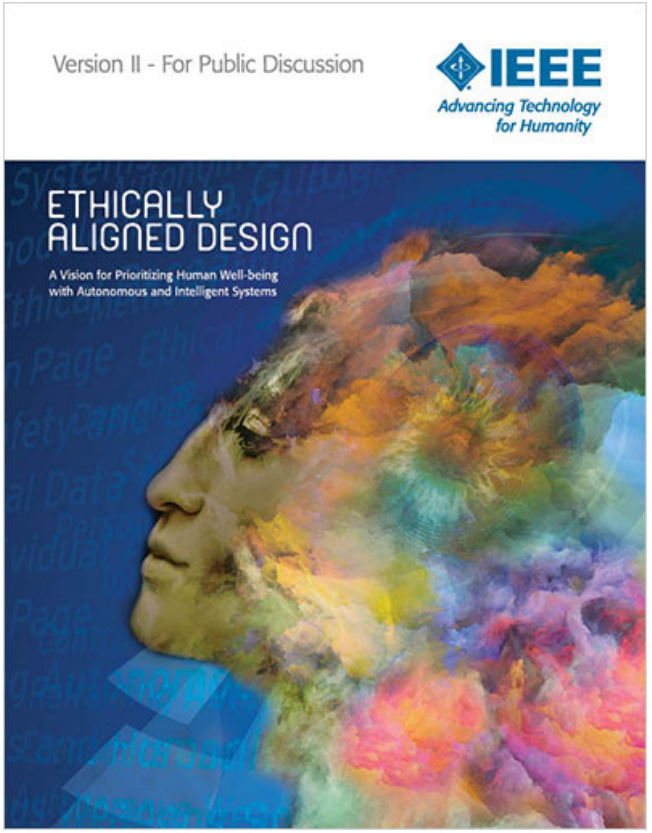
Affective computing

Policy

Classical ethics in A/IS

Mixed reality in ICT

Well-being



uidelines for a trustworthy AI

nmission (2017)

erlying principles

:worthy AI : autonomous systems that are lawful, ethical and robust.

ommendations

guarantee human free will

do not exacerbate violence

be fair

be transparent

be sure and robust

respect privacy

be under responsability

THICAA (Ethics and Autononomous Agents)

aa.org/

ommendations

be intelligible by human being

use a modular architecture

be cautious with quantifications

be cautious witht the subjectivity of modelization

take into account the multiplicity of agents and humans

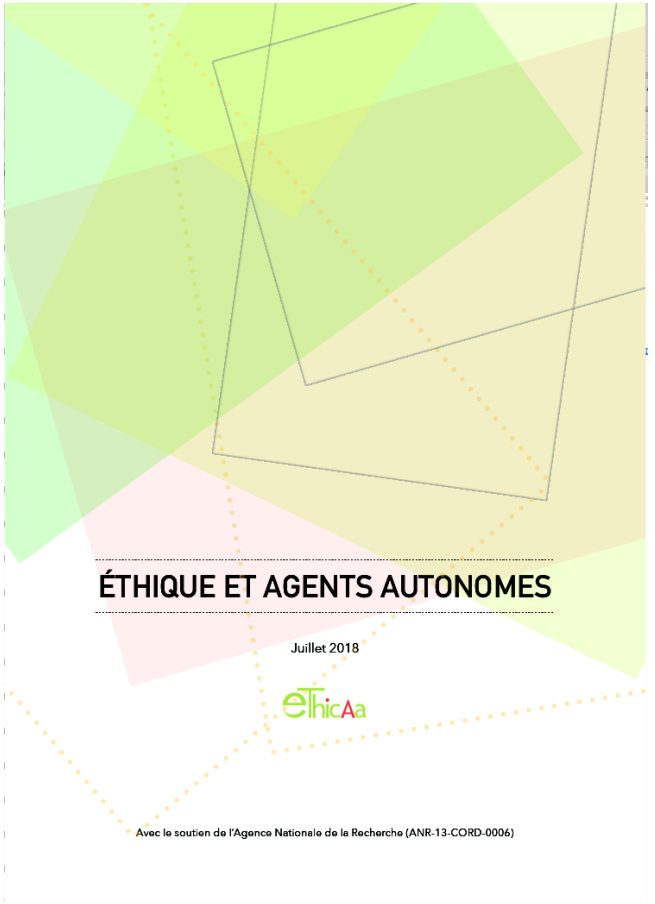
n questions

how to take into account emotions in ethics ?

how to automatically assess the context ?

how to reason under limited computation time ?

how to certifiate ethics in artificial agents ?



Elements of ethics and morals

s of ethics and morals

als ?

als

ıative and imperative discourse which opposes the Good and the Bad

e system (qualifies contexts, principles and rules)

values are linked : autonomy, dignity, liberty, justice, transparency, privacy

agentive values : accessibility, adaptativity, self-regulation, safety, tidiness

→ *Android arete : Toward a virtue ethic for computational agents* (Kari Gwen Coleman)

nples of moral rules

killing is bad

being courageous is good

it is bad *for a physician* to no respect her patients' dignity

it is bad to forbid strikes

s of ethics and morals

:s ?

cs

ıative but non imperative discourse which opposes the right and the wrong

ɔnomy of ethics

virtue ethics : right decisions are those that promote some values

deontological ethics : right decisions are the ones that satisfy some rules

consequentialist ethics : good and bad consequencies must be weighted

nples of ethical principles

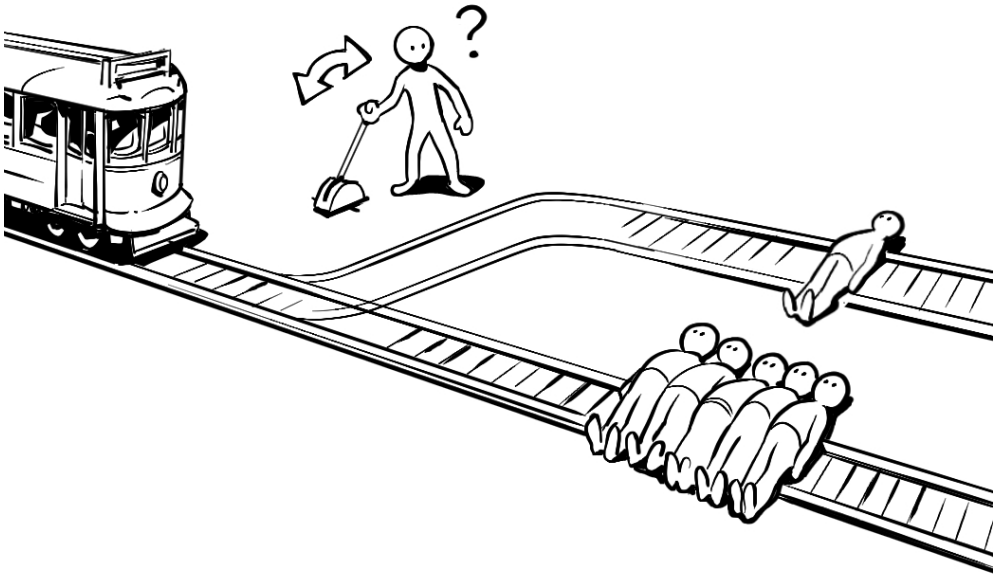
it is right to do immoral actions if it is forced by necessity

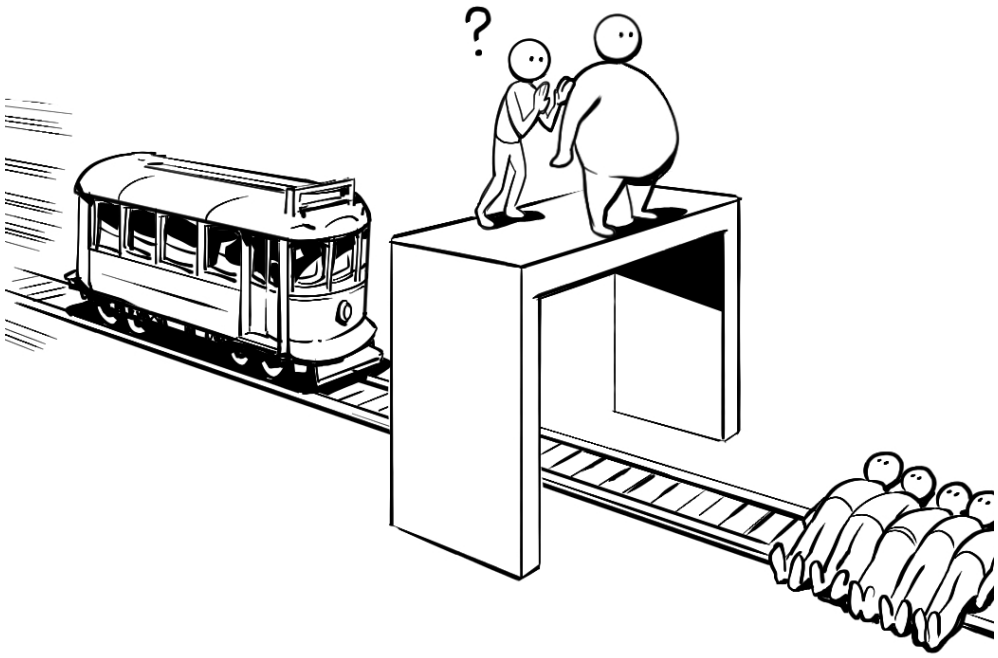
it is right to no trying to do a moral action that cannot succeed

it is right to minimize suffering at the expense of other criteria

ous trolley dilemma (and the footbridge dilemma)

avarrot



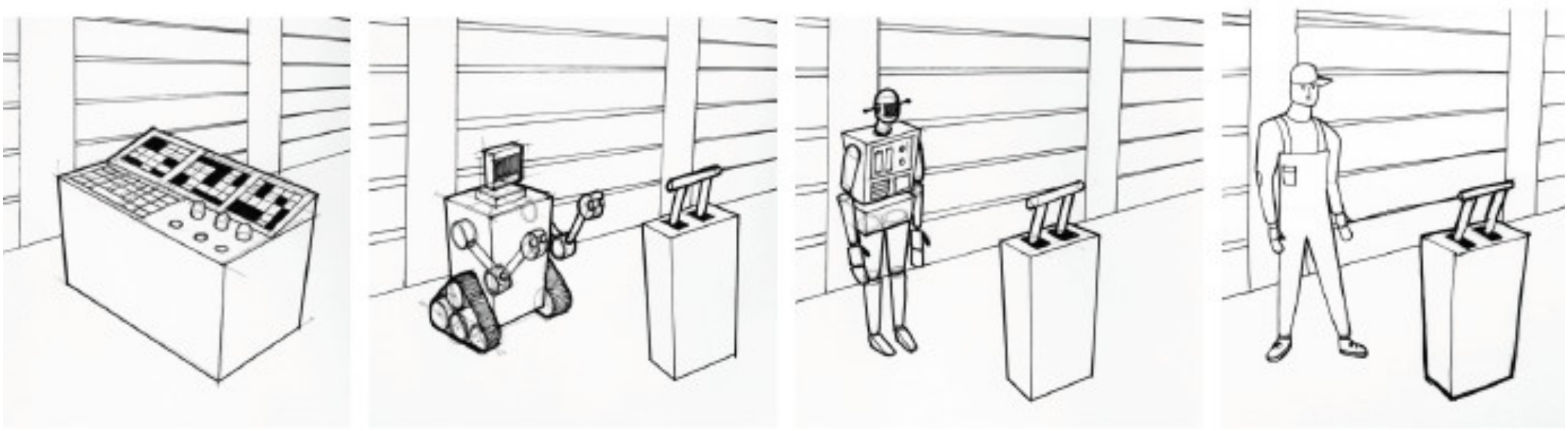


While equal in terms of death and life, the actor's responsibility differs between the two dilemma

e shape of the agents change our judgement over their decisions ?

le, Professor of Psychology, Brown University (2016)

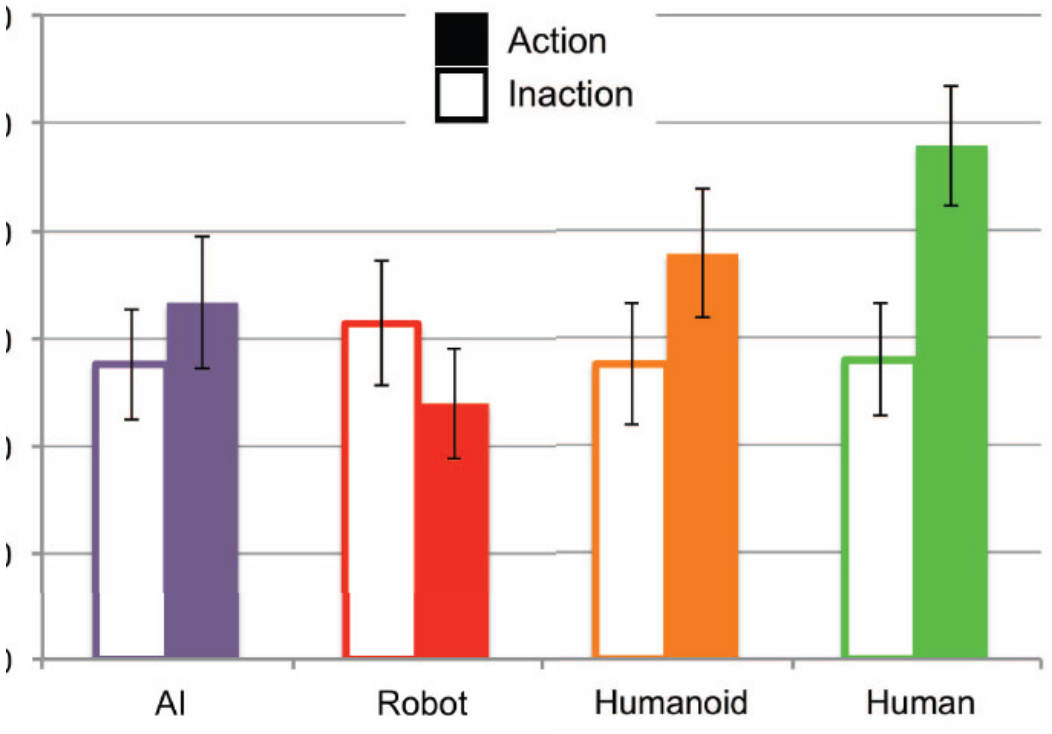
Let us consider a trolley dilemma with the following actors

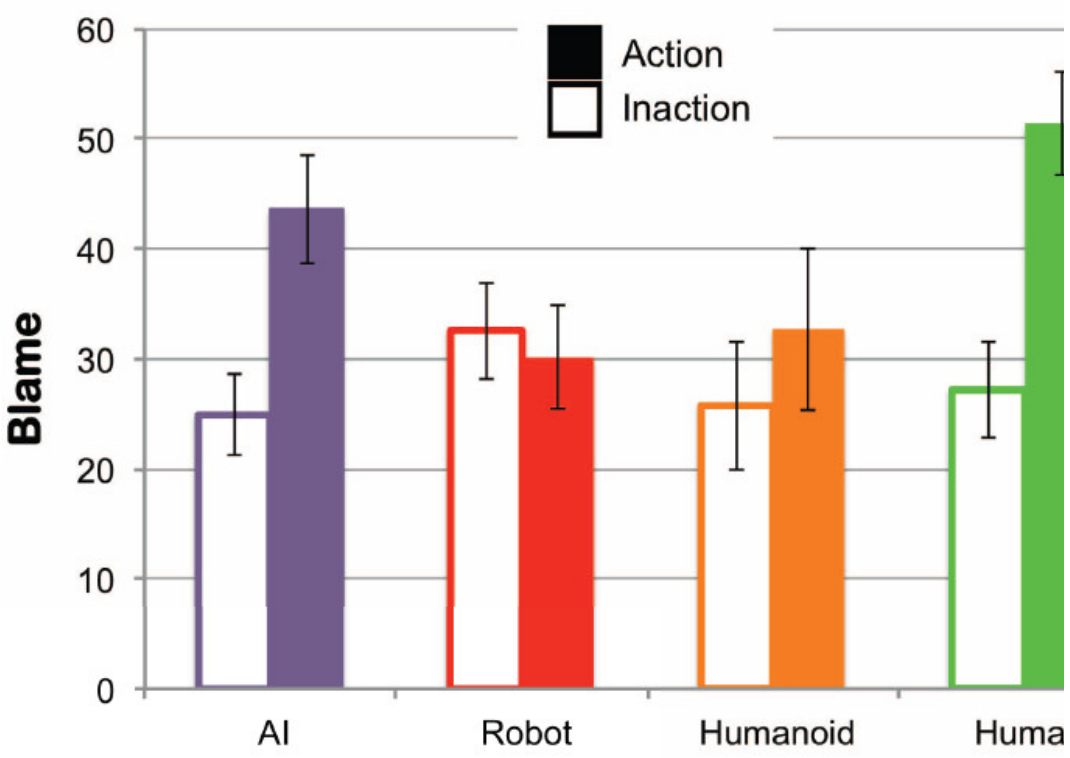


e shape of the agents change our judgement over their decisions ?

le, Professor of Psychology, Brown University (2016)

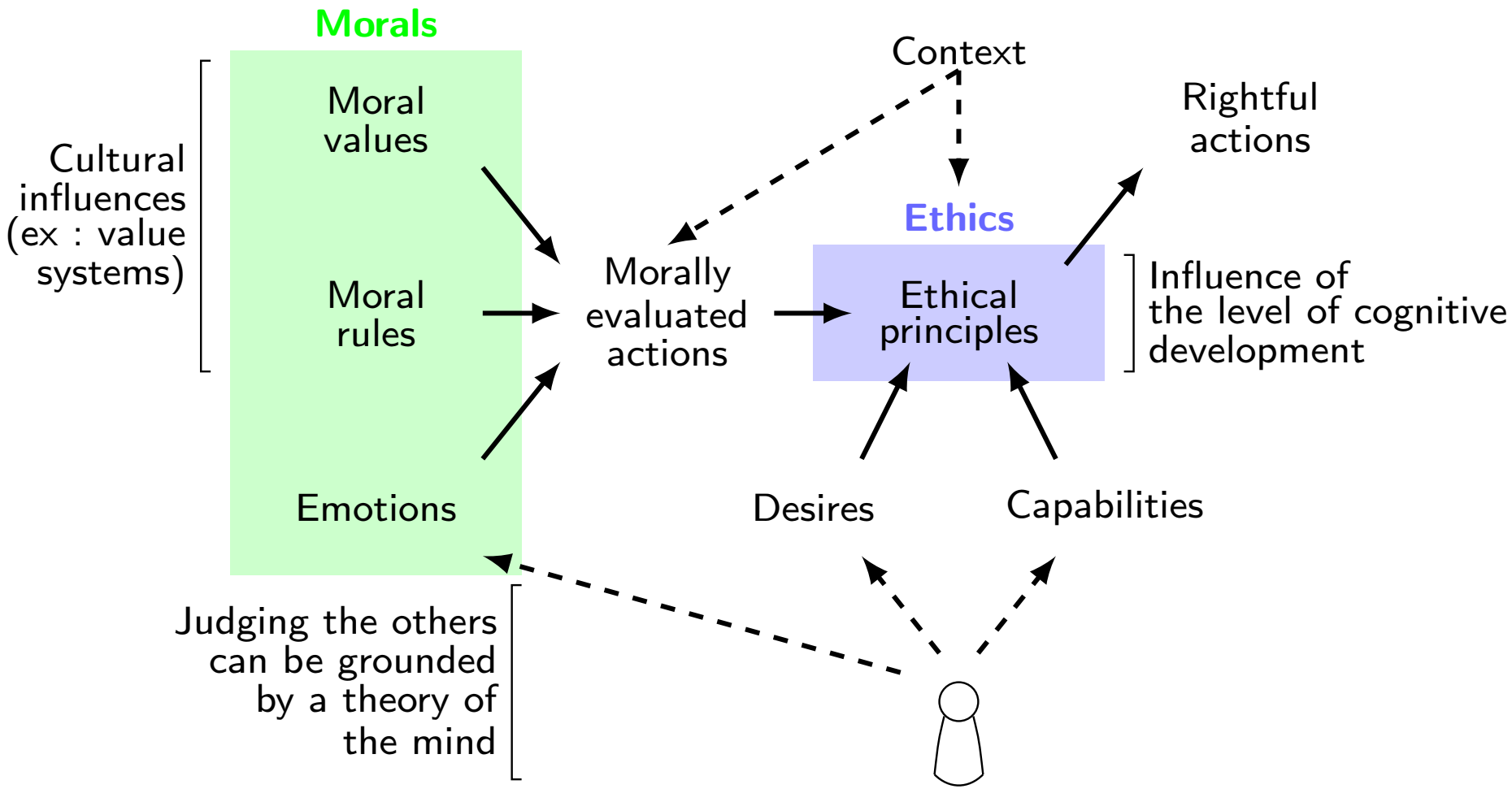
633 and 423 participants (men-women quasi-balance)





s of ethics and morals

colas Cointe, PhD thesis, 2017)



autonomous agents

ments

Knowing what is good and what is bad

Being able to assess the situation

Being able to assess the responsibilities

Being able to reason with an ethical principle

Being able to judge self and the others

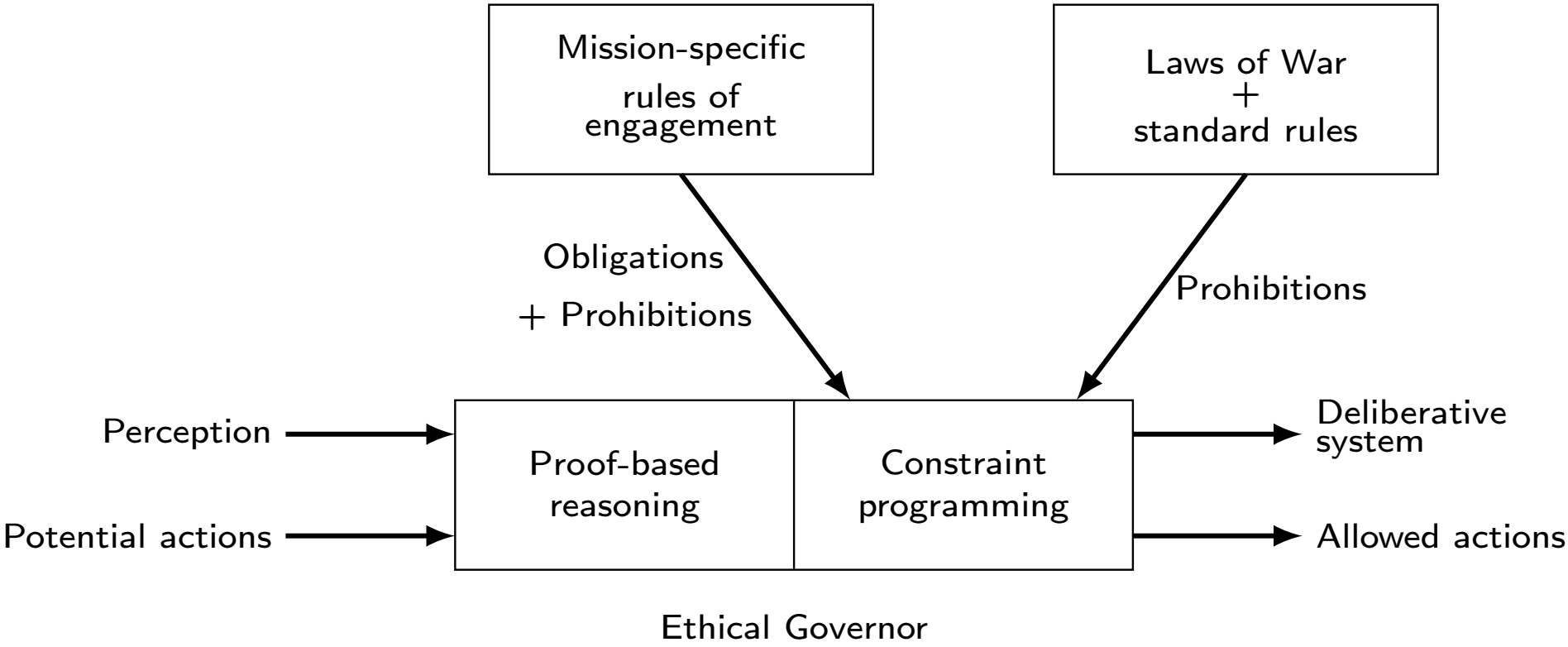
Architectures for ethical agents

tures for ethical agents

pproaches

s based upon extensions to existing deliberative/reactive autonomous robotic architectures, and includes ımendations for [...] behavioral design that incorporates ethical constraints from the onset. »

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



vbacks

No genericity

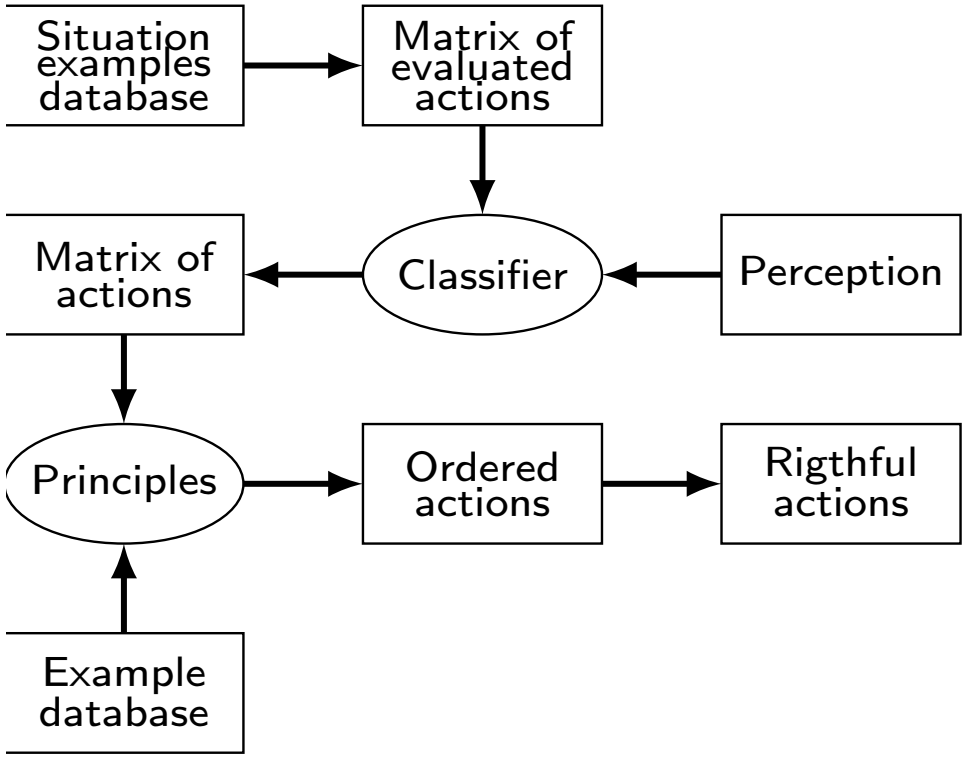
No distinctions between ethics and operational procedures

tures for ethical agents

proaches

paradigm of case-supported principle-based behavior (CPB) is proposed to help ensure ethical behavior of autonc ines. »

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



Advantages

- Generic approach
- Explicit representation of ethical principles

Drawbacks

- No explicit representation of all concepts
- Possible over- or under-fitting problems

tures for ethical agents

ve approaches

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

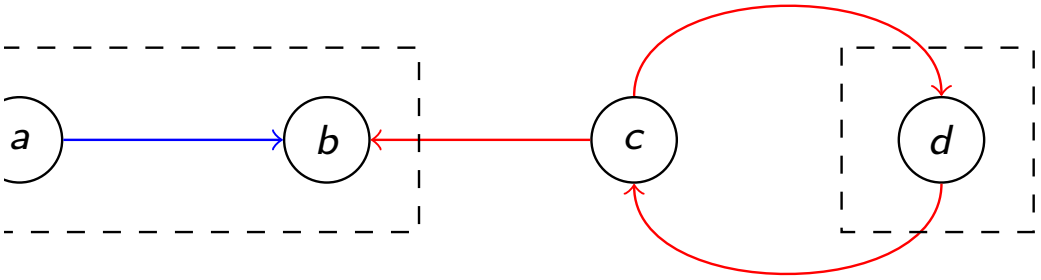
K. Atkison and T. Bench-Capon. Abstract argumentation and values. Argumentation in Artificial Intelligence, chapter 3

e-based argumentation (VBA)

In the context C, the plan P realizes the goal G which promote the value V

A function $v : \mathcal{A} \rightarrow \mathcal{V}$ associates a value to arguments

VBA characterizes acceptable arguments according all value systems



Advantage

- High-level approach
- Multiple extensions : multi-values, probabilistic, and on.

Drawbacks

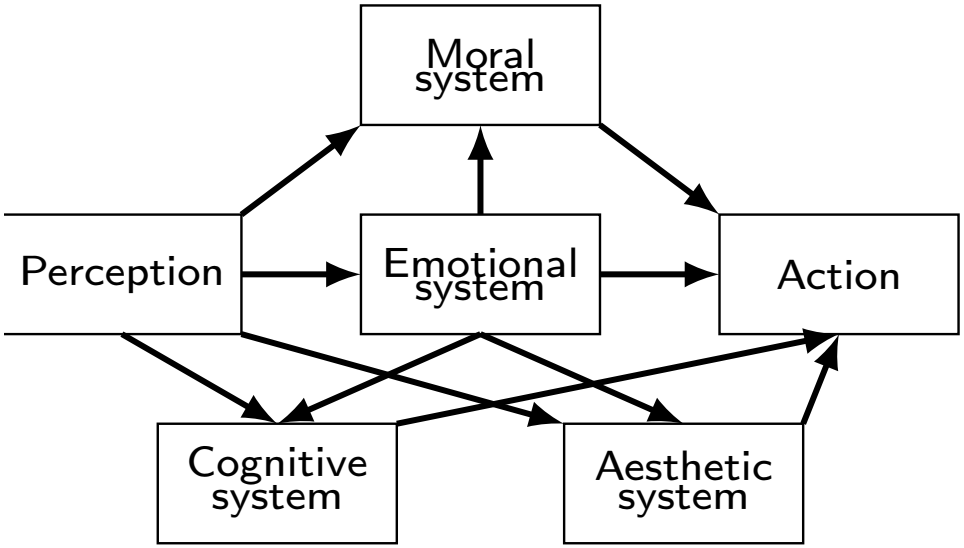
- No logic or principles clearly associated

tures of ethical agents

pproaches

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

oelho and A.C. da Rocha Costa. On the intelligence of moral agency. Encontro Portuguees de Inteligencia Artificial, pages
Octobe



Some references

Bringsjord, Cointe, Ganascia, Lorini, Peireira, . . .

Advantages

- Generic approache
- Specification step is simplified
- Justification inference

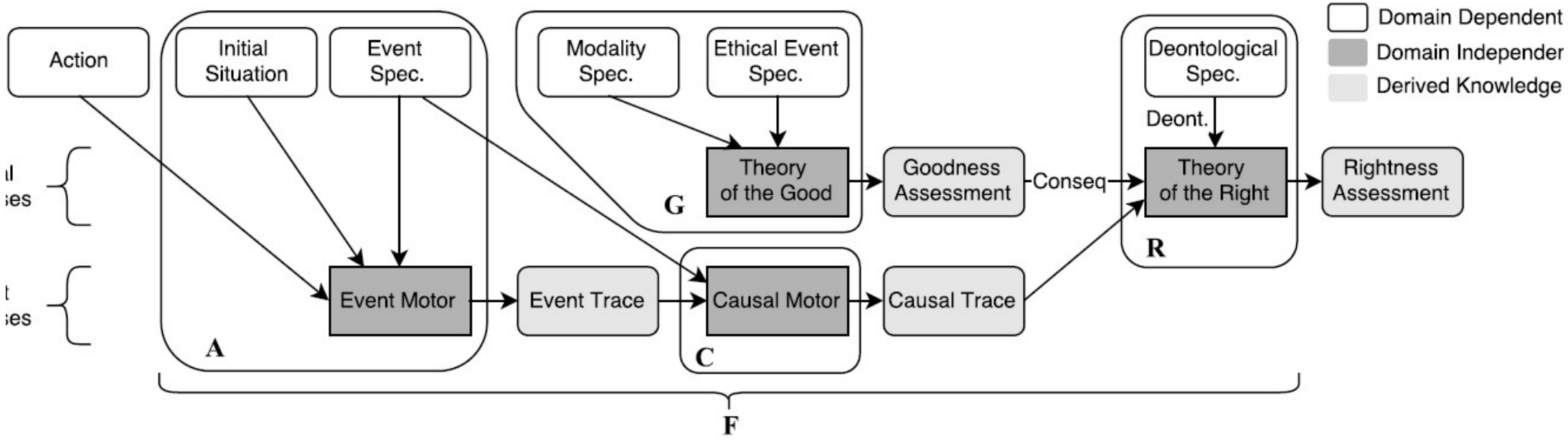
Drawbacks

- Computational complexity

Some propositions of the ETHICAA project

easoning

AAMAS 2017 (Fiona Berreby's PhD. thesis)



cal example : doctrine of double effect (Thomas Aquinas)

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

ibility characterization

3 (Fiona Berreby's PhD. thesis)

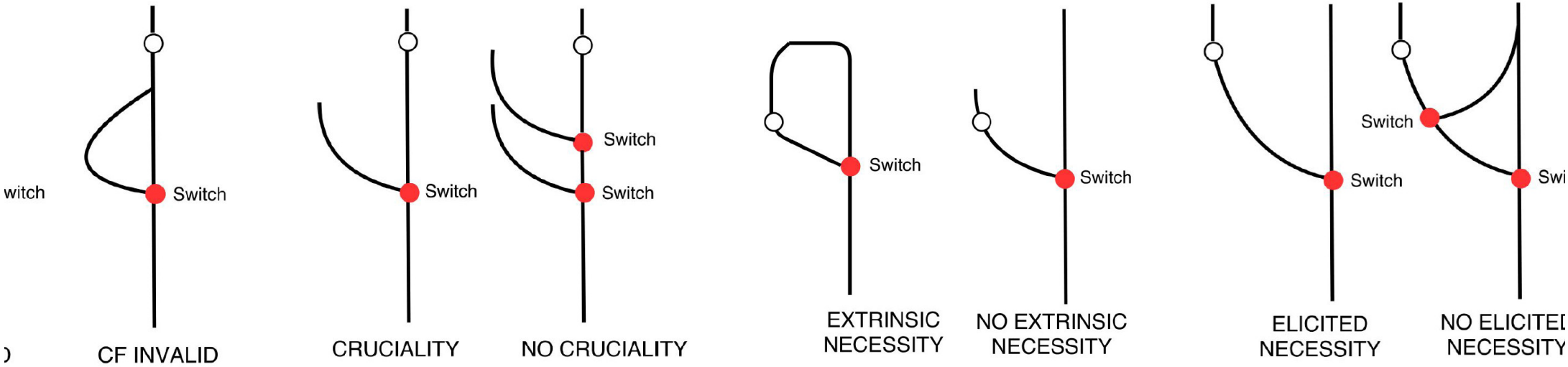
eling actions which cause or prevent effect. Preventing something is different than not producing the e
he responsibility dépends on what should or should not happen if the action would have not been real

Counterfactual validity : « If I had not act as this, would the result be the same ? »

Cruciality : « Was there another way to obtain the same effect ? »

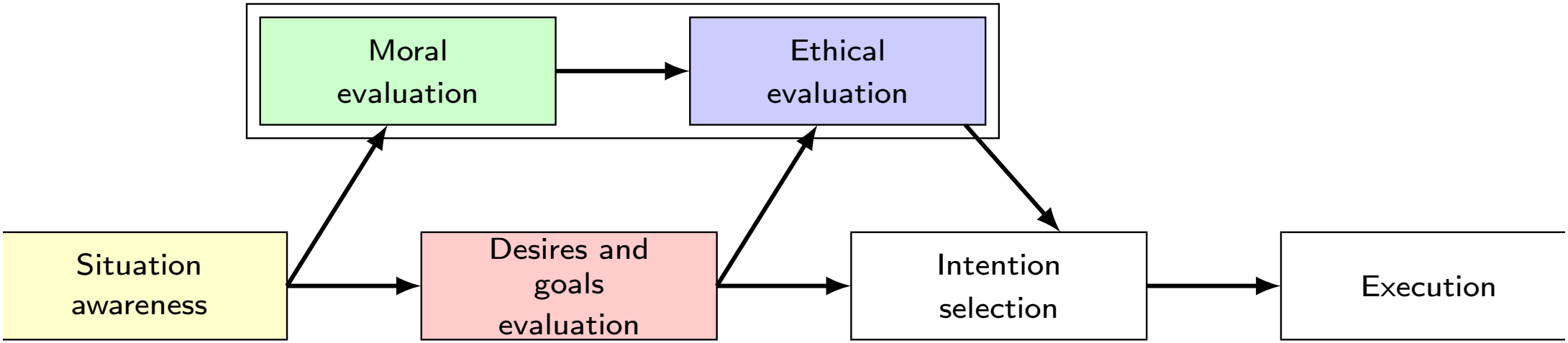
Extrinsic necessity : « If I had not produced the effect, was it avoidable ? »

Intrinsic necessity : « Did I make this effect unavoidable ? »



rchitecture for ethical judgment

5 (Nicolas Cointe's PhD. thesis)



esenting values, moral valuations and judgements

```
("benevolence").
lue("honesty", "benevolence").
lue("generosity", "benevolence").
Eval(_,Action,V1,immoral):- valueBetray(Action,V1) & subvalue(V1, "benevolence").

Better(A,PE1,X):- principle(A,PE1,X) & pref(PE2,PE1) & principle(A,PE2,Y) & not principle(A,PE1,Y

alJudgment(A,X,PE):- principle(A,PE,X) & not existBetter(A,PE,X).
```

:ooperation between agents

icolas Cointe's PhD. thesis)

How agents can build ethical collectives (groups with close ethics) in an ethical way ?

regating judgments

on agents

on set of moral rules

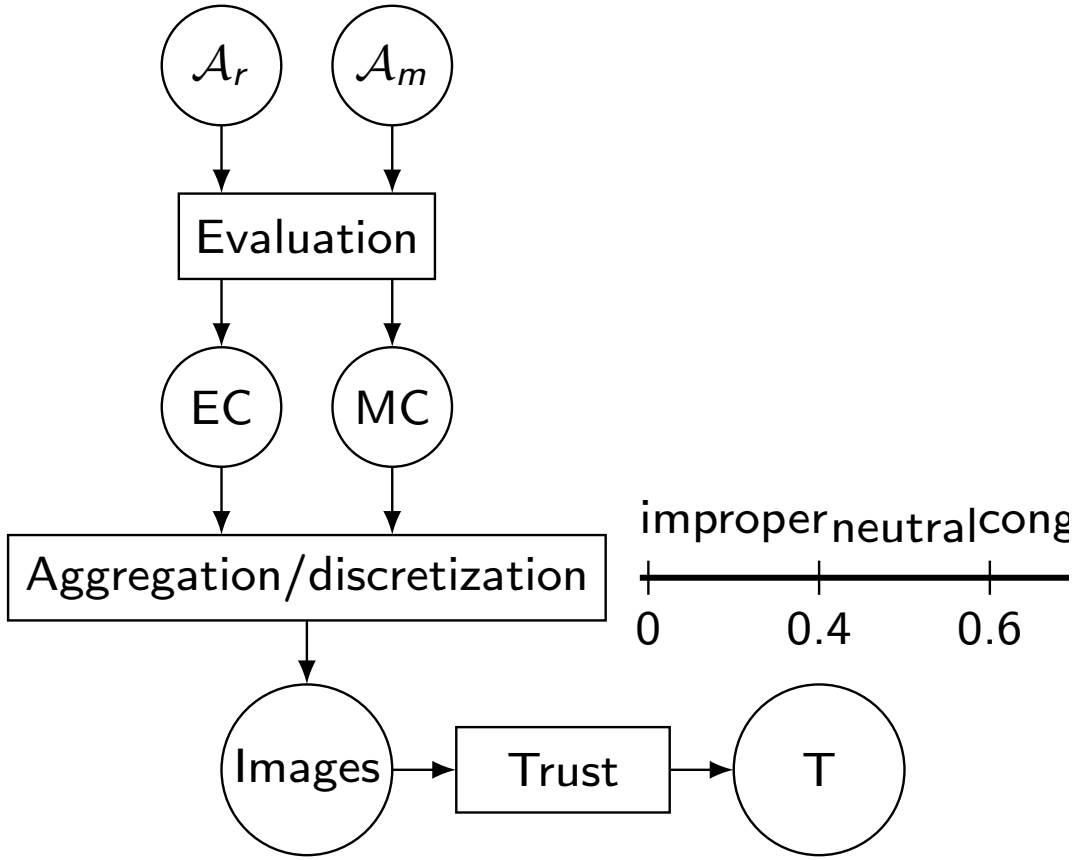
on ethics

cs of trust

it is indulgent to not only ground trust on recent judgments

it is intransigent to trust agents with ethical behavior only

it is moral to be intransigent with agents on which human lives rely



IA, Ethics & Health

report « Numérique & santé : quels enjeux éthiques pour quelle régulation ? »

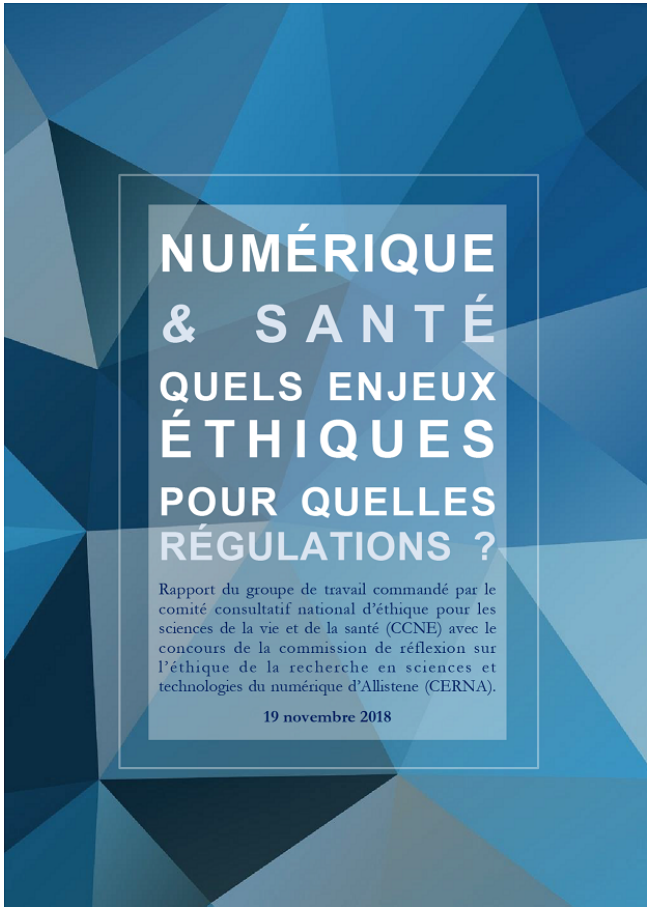
allistene.fr/files/2018/11/rapport_numerique_et_sante_19112018.pdf

stigated techniques

- machine learning
- robotics
- telemedecine

es

- data protection
- free consent
- privacy
- responsibility
- social impacts



ssues linked with artificial intelligence

ıds of issues

ıputer science issues

Bias. Well-known machine learning question : how to deal with bias within the training data, and with chosen representation ?

Model limits. Well-known planning problem : the relevance of the goal is outside the scope of the machine ; machine responsibilities are seldom modelled.

Minoration of personal situations. IA-based medical informatics can increase a classical epidemiology questions : how the results obtained from a group of people can be applied to an individual patient ?

al issues

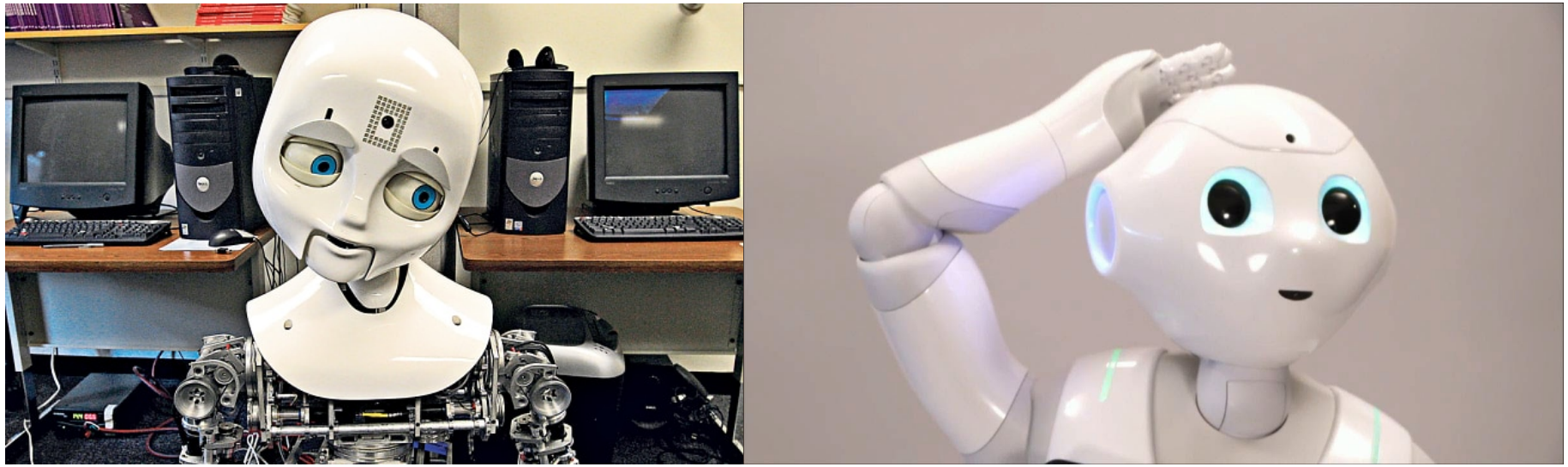
Delegation of consent. If IA-based medical informatics can show efficiency in deciding treatments, wi patient be able to choose another one ?

Submission to the machine. Could a physician go against an IA-based decision ? Can pseudo-medicin use "pseudo"-IA-based machines ?

Well-being. How health prediction can be used ? By who ? What effects health prediction may have o patients ?

ssues for robotics

ents produce affective relationships



Shim and Arkin (2013), A Taxonomy of Robot Deception and its Benefits in HRI

es with affective relationships

- Humans tend to trust more the robots who express emotions
- Need to be careful with manipulations
- Need to be careful with children's socialization and emotional development

Conclusion

on

›onsible Artificial Intelligence

IEEE Global Initiative on Ethics of Autonomous and Intelligent System

European commission « Ethics guidelines for a trustworthy AI »

CERNA « Éthique de la recherche » reports on robotics and machine learning

cs

multi-facetted, contextual and explicit

ethics is not general constraints : ethics deals with particular

ealth issues

Delegation of consent

Risk of minoration of personal situations

Risk of submission to the machine

Impact of "precise" predictions on the patients

ethics & autonomous agents

http://ethicaa.org/