Ethics and AI:
State of the art and perspectives

Nadia Abchiche-Mimouni

nadia.abchichemimouni@univ-evry.fr

IBISC Lab. Univ. Evry, Université Paris-Saclay

Plan

- General ethics and definitions
- Why is Artificial Intelligence concerned with ethical issues?
- Ethics in Artificial Intelligence
  - Institutionnal initiatives
  - Research works
- Perspectives and open discussion

General ethics (1/6)

Definitions and visions from moral philosophy

- Ethics
  - Set of values whose purpose generates the good, is looking for the Good (Aristote)
  - Relativize the notion of good/bad action
  - Reflection argued for the well-act / be beautiful
- Moral
  - Good/bad action
  - Set of standards specific to an individual, a social group or a people
  - Concept of rights, duties and prohibitions
- Deontology
  - Set of Rules and Duties Governing a Profession

General ethics (2/6)

Definitions and visions from moral philosophy

General definition

Ethics proposes to question moral values and moral principles that should guide our actions, in different situations, in order to act in accordance with them.

([Mill, John Stuart 1998], Kant)

1. Normative ethics (substantial ethics) → allow to determine, between two actions, which is morally better
2. Meta ethics → philosophical analysis of ethical discourse and its epistemological and metaphysical presuppositions
3. Applied ethics → analysis of concrete situations raising ethical issues for decision-making

General ethics (3/6)

Definitions and visions from moral philosophy

1. Normative ethics:

a) Ethical or moral nihilism: asserts that there is nothing morally right or wrong. It argues that:
   1. A judgments such as "The killing of innocent people is always morally wrong" can not be true or false to the extent that moral judgments have no logical value.
   2. A statement like "This stone weighs 30 kg" has a logical value to the extent that this can be true or false.
b) Ethical relativism: ethical constructivism, ethical subjectivism
c) Ethical realism: ethical naturalism, ethical non-naturalism

General ethics (4/6)

Definitions and visions from moral philosophy

2. Meta ethics:

- Ethical or moral nihilism: asserts that there is nothing morally right or wrong. It argues that judgments such as "The killing of innocent people is always morally wrong" can not be true or false to the extent that moral judgments have no logical value. A statement like "This stone weighs 30 kg" has a logical value to the extent that this can be true or false.
    - Ethical relativism: ethical constructivism, ethical subjectivism
    - Ethical realism: ethical naturalism, ethical non-naturalism

General ethics (5/6)

3. Applied ethics

- Bioethics
  - human procreation (assisted procreation, abortion, gamete donation, prenatal diagnosis, cloning ...);
  - end of life (palliative care, therapeutic relent lessness, euthanasia ...);
  - genomics;
  - public health;
  - neuroscience and neuropsychiatry;
- The ethics of the environment
  - sustainable development / responsibility towards future generations;
  - management of natural resources (water, forests, subsoils ...);
  - waste management;
  - industrial and agricultural pollution;
  - animal rights;
  - genetically modified organisms (GMOs);
  - biodiversity / ecosystem conservation;
  - etc.

General ethics (6/6)

Ethical dilemmas

Definition

- Situations in which any available choice leads to transgressing some accepted ethical principle and yet a decision has to be made, [Kirkpatrick, 2015].
- An ethical principle is unable to give a different valuation (a preference) between two options: each option is supported by ethical reasons, given that the execution of both is not possible [McConnell 2014].

Example

In Book I of Plato's Republic, Cephalus defines 'justice' as "speaking the truth" and "paying one's debts".
Socrates quickly refutes this by suggesting that it would be wrong to repay certain debts—for example, to return a borrowed weapon to a friend who is not in his right mind.
Conflict between two moral norms:

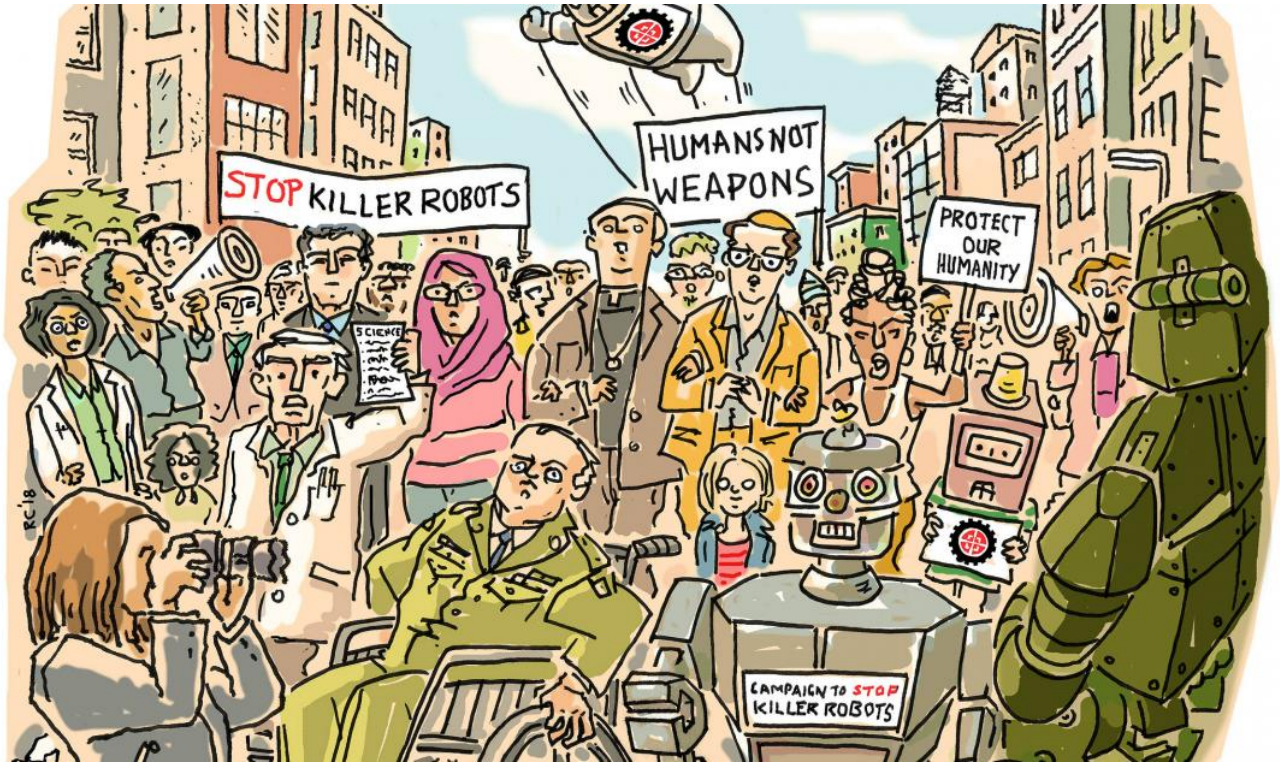
1. Repaying one's debts
2. Protecting others from harm

Why and how is AI concerned with ethical issues?

- Self driving cars
- Algorithmic discrimination
- Data ownership
- …

chombosan/Alamy Stock Photo

https://twitter.com/technicolorr/status/74782369681



Why and how is AI concerned with ethical issues?

Jobs replacement

...

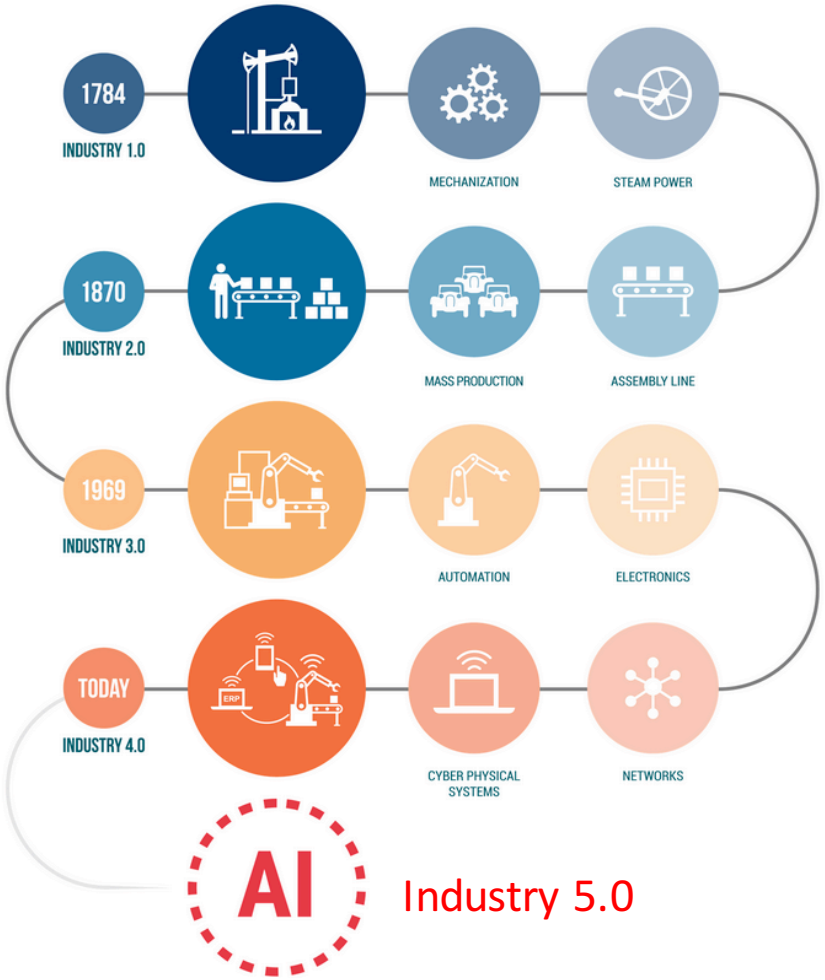
Catalogue of fears

Probability of computerisation of different occupations, 2013
(1 = certain)

Job	Probability
Recreational therapists	0.003
Dentists	0.004
Athletic trainers	0.007
Clergy	0.008
Chemical engineers	0.02
Editors	0.06
Firefighters	0.17
Actors	0.37
Health technologists	0.40
Economists	0.43
Commercial pilots	0.55
Machinists	0.65
Word processors and typists	0.81
Real-estate sales agents	0.86
Technical writers	0.89
Retail salespeople	0.92
Accountants and auditors	0.94
Telemarketers	0.99

Source: "The Future of Employment: How Susceptible are Jobs to Computerisation?", by C. Frey and M. Osborne (2013)

Economist.com

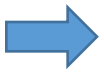


Industry 5.0

Why and how is AI concerned with ethical issues?

Ethical questions

- Algorithms transparency
- Ethics of personalization technologies
- Encoding legislation into machine
- Responsibility
- Value alignment
- Sustainable AI systems
- …



Open letter for:

- *Research Priorities for Robust and Beneficial Artificial Intelligence* ([Stuart Russell & al., 2015])

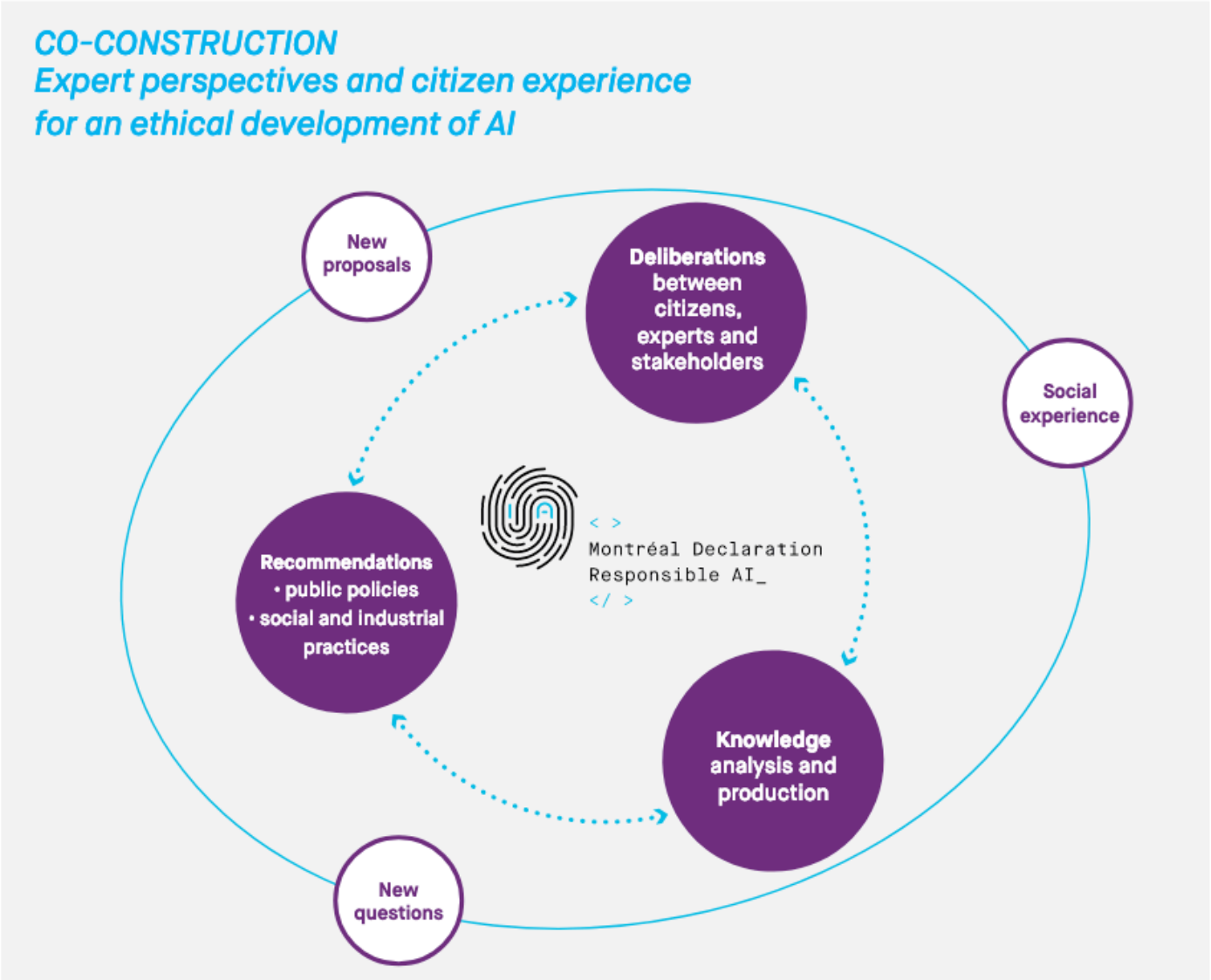
Why and how is AI concerned with ethical issues?

Societal recommendations

Declaration of Montreal for a responsible development of AI

Yoshua Bengio, and Yann LeCun

awarded the Turing prize, 27th mars 2019



Why and how is AI concerned with ethical issues?

Societal recommendations
Declaration of Montreal for a responsible development of AI

The Declaration sparked public debate and encouraged a progressive and inclusive orientation to the development of AI.

https://www.montrealdeclaration-responsibleai.com/

Ethics in Artificial Intelligence

Institutional initiatives (1/5)

Ethics and Governance of AI Initiative (Launched in 2017)

- Hybrid research effort and philanthropic fund that seeks to ensure that technologies of automation and machine learning are researched, developed, and deployed in a way which vindicate social values of fairness, human autonomy, and justice.
- A joint project of the MIT Media Lab and the Harvard Berkman-Klein Center for Internet and Society.
- Projects on three domains:
  1. AI and Justice
  2. Information Quality
  3. Autonomy and Interaction

Ethics in Artificial Intelligence

Institutional initiatives (2/5)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Moving "From Principles to Practice" with standards projects, certification programs, and global consensus building to inspire the *Ethically Aligned Design* of autonomous and intelligent technologies

Ethically Aligned Design, https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

1. Human Rights
2. Well-being
3. Data Agency
4. Effectiveness
5. Transparency
6. Accountability
7. Awareness
8. Competence

Ethics in Artificial Intelligence

Institutional initiatives (3/5)

World economic forum
https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

Top 9 ethical issues in artificial intelligence

1. Unemployment. What happens after the end of jobs?
2. Inequality. How do we distribute the wealth created by machines?
3. Humanity. How do machines affect our behaviour and interaction?
4. Artificial stupidity. How can we guard against mistakes?
5. Racist robots. How do we eliminate AI bias?
6. Security. How do we keep AI safe from adversaries?
7. Evil genies. How do we protect against unintended consequences?
8. Singularity. How do we stay in control of a complex intelligent system?
9. Robot rights. How do we define the humane treatment of AI?

Ethics in Artificial Intelligence

Institutional initiatives (4/5)

Future of life institute, https://futureoflife.org/

What we really need to do is make sure that life continues into the future. […] It's best to try to prevent a negative circumstance from occurring than to wait for it to occur and then be reactive."

-Elon Musk on keeping AI safe and beneficial

Ethics in Artificial Intelligence

Institutional initiatives (5/5)

(*) Institute for Ethics in Artificial Intelligence (Facebook and Technical University of Munich)
(*) https://newsroom.fb.com/news/2019/01/tum-institute-for-ethics-in-ai/
(*) Date Science Central
(*) https://www.datasciencecentral.com/profiles/blogs/ethics-in-artificial-intelligence
(*) https://www.information-age.com/ethics-artificial-intelligence-123475134/
(*) https://www.information-age.com/ethics-artificial-intelligence-123475134/
(*) http://www.ethicsandai.com/
(*) European AI Alliance
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top
(*) The Ethics Of Algorithms, ethicsofalgorithms.com, Drexel university, Philadelphia, Pennsylvania, USA
(*) CIFAR Pan-Canadian Artificial Intelligence Strategy
https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy
(*) Online ethics center for engineering and science, https://www.onlineethics.org/

Current research works (1/)

A taxonomy of integration AI and Ethics ([Yu & al., 2018], IJCAI)
4 axes:

1. Exploring Ethical Dilemmas
2. Individual Ethical Decision Frameworks
3. Collective Ethical Decision Frameworks
4. Ethics in Human-AI Interactions

Current research works

Exploring ethical dilemmas

- Two examples:
  1. **GenEth** ethical dilemma analyzer ([Anderson and Anderson, 2014], AAAI)
     - Through a dialog with ethicists, helps codify ethical principles in any given domain by proposing a set of representation schemas for framing the discussions on AI ethics which includes: Features, duties, actions, cases and principles.
  2. **Moral Machine project** (http://moralmachine.mit.edu/)
     - Allows participants to judge various ethical dilemmas facing Autonomous Vehicules which have malfunctioned, and select which outcomes they prefer.
     - The decisions are analyzed according to different considerations including: 1) saving more lives, 2) protecting passengers, 3)upholding the law, 4) avoiding intervention, 5) gender preference,6) species preference, 7) age preference, and 8) social value preference.
     - The project provides a user interface for participants to design their own ethical dilemmas to elicit opinions from others.

Current research works

Individual Ethical Decision Frameworks

- Formal verification of moral values in MAS ([BONNET & al. 2017])
  - Consider whether moral rules shared by many of us are followed by the agents.
  - Hard problem because most of the moral rules are often not compatible. In such cases, humans usually follow ethical rules to promote one moral rule or another.
  - Using formal verification to ensure that an agent follows a given ethical rule could help in increasing the confidence in artificial agents.
  - A set of formal properties through an ethical rule ordering conflicting moral rules with respect to a value system.
  - If the behaviour of an agent verifies these properties (which can be proven using an existing proof framework), it means that this agent follows this ethical rule.

Current research works

Individual Ethical Decision Frameworks

- MoralDM: Moral Decision Making, ([Blass and Forbus, 2015])
  - Moral decision-making by humans not only involves utilitarian considerations, but also moral rules
  - Rules are acquired from past example cases and are often culturally sensitive
  - Agents resolve ethical dilemmas by leveraging on two mechanisms:
    1. First-principles reasoning
    2. Analogical reasoning

Current research works

Individual Ethical Decision Frameworks

- *Ethical Judgment of Agents' Behaviors in Multi-Agent Systems: (*[Cointe & al. 2016])
  - Explicit representation of ethics
  - Explicit process of ethical judgment
  - Consider both individual and collective reasoning on various theories of good and right
  - Ethics consists in conciliating desires, morals and abilities
  - Generic Ethical Judgment Process (EJP) use evaluation, moral and ethical knowledge. It is structured along Awareness, Evaluation, Goodness and Rightness processes
  - Context of a BDI model, using mental states such as beliefs and desires
  - The judgment process is useful for an agent to judge it's own behavior and to judge the behaviors of other agents

Current research works

Individual Ethical Decision Frameworks

- Combining 2 ways of moral decision making frameworks for artificial intelligence, ([Conitzer et al., 2017])
  - Machine learning based ethical decision-making, the key approach is to classify whether a given action under a given scenario is morally right or wrong.
  - Well-labeled training data, possibly from human judgements is acquired.
  - Game theory is combined with machine learning: game theoretic analysis of ethics is used as a feature to train machine learning approaches,
  - While machine learning helps game theory identify ethical aspects which are overlooked.

Current research works

Collective Ethical Decision Frameworks

- Event-Based and Scenario-Based Causality for Computational Ethics ([Fiona & al., 2018])
  - Causal model based approach
  - Moral responsibility
  - Modelling actions and omissions
  - Matrix of causal relations (supporting and opposing relations)
  - Scenario-based trace (context and the cause of the cause)

	Strong	**Weak**
Supporting	Causes	Enables
Opposing	Prevents	Excludes

Current research works

Collective Ethical Decision Frameworks

Responsible Autonomy ([Dignum, 2017])

- No optimal solution to solve ethical dilemma
- Ensure that AI is developed responsibly incorporating social and ethical values
- Societal concerns about the ethics of AI must be reflected in design
- Three principles (ART):
  1. **A**ccountability: answerability, blameworthiness and liability;
  2. **R**esponsibility: being in charge, or being the cause behind whether something succeeds or fails;
  3. **T**ransparency: openness of data, processes and results for inspection and monitoring.

Current research works

Collective Ethical Decision Frameworks

Responsible Autonomy ([Dignum, 2017])

- Who is responsible to the decision: four possible levels of autonomy and regulation:
  1. **Human control:** a person or group of persons are responsible for the decision;
  2. **Regulation:** the decision is incorporated in the systemic infrastructure of the environment;
  3. **Artificial Moral Agents (AMA):** systems incorporate moral reasoning in their deliberation and to explain their behavior in terms of moral concepts;
  4. **Random:** the autonomous system randomly chooses its action when faced with a (moral) decision
- How decisions are dependent on different moral and societal values:
  - Determine which moral values to aim for and which ethical principles to adhere to in a given circumstance;
  - Basic values refer to desirable goals that motivate action and transcend specific actions and situations.
  - Values are quite consistent across cultures. Classified into 4 dimensions: (i) Openness to change, (ii) Self-enhancement, (iii) Conservation, (iv) Self-transcendence.
  - Prioritize different moral and societal values, depending on individual and socio-cultural environment;
  - Values serve as criteria to guide the selection or evaluation of actions, taking into account the relative priority of values.

Current research works

Collective Ethical Decision Frameworks

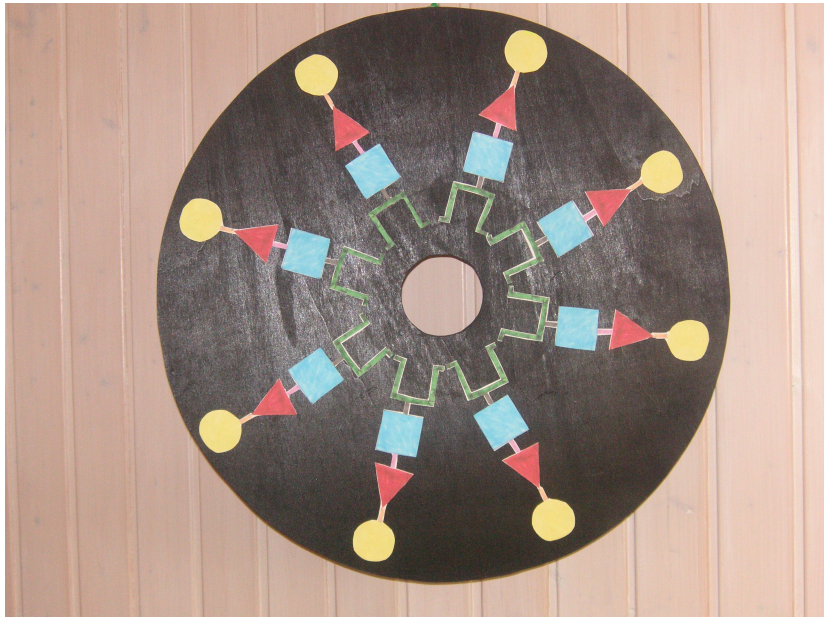
Responsible Autonomy ([Dignum, 2017])

Ethical Deliberation	Computational reqs	ART
User Control	• Realtime reasoning • Ensure situational awareness to user • Explanation capabilities • Output internal state in user understable way	• Delegated to user
Regulation	• Formal link from values to norms to behaviour • Define institutions for monitoring and control • Moral reasoning can be done off-line	• A: institutional • R: institutional • T: system (by requirement)
AMA	• Formal link from values to norms to behaviour • Define reasoning rules • Supervised learning of morality • Realtime reasoning	• A: system (by explanation) • R: system (by deliberation) • T: system (by requirement)

Computational and ART consequences of ethical deliberation mechanisms

Perspectives and open discussion

- Ethics in AI raise many scientific challenges
- All domains of computer science will be concerned (software engineering, modelling, optimization, formal methods, security, machine learning…)
- This is no less a revolution in the science of automation
- We are moving towards new perspectives to re-enchant our research profession



THANK YOU FOR YOUR ATTENTION!

Bibliography

- [Cointe & al., 2016] Nicolas Cointe, Grégory Bonnet and Olivier Boissier, *Ethical Judgment of Agents' Behaviors in Multi-Agent Systems*, Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.
- [Conitzer et al., 2017] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In AAAI, pages 4831–4835, 2017.
- Emmanuel Kant (1724-1804) Königsberg Allemagne.
- [Fiona & al., 2018] Fiona Berreby, Gauvain Bourgne, Jean-Gabriel Ganascia: Event-Based and Scenario-Based Causality for Computational Ethics. AAMAS 2018: 147-155
- [Mill, John Stuart 1998] Mill, John Stuart (1998) *Utilitarianism* [archive] Oxford University Press (ISBN 0-19-875163-X)
- [McConnell, 2014] T. McConnell. Moral dilemmas. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Fall edition, 2014., https://plato.stanford.edu/entries/moral-dilemmas/#ConMorDil

Bibliography

1. [Anderson and Anderson, 2014] Michael Anderson and Susan Leigh Anderson. GenEth: A general ethical dilemma analyzer. In AAAI, pages 253–261, 2014.
2. [Berreby & al. 2018] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2018. Event-Based and Scenario-Based Causality for Computational Ethics. InProc. of the17th International Conference on Autonomous Agents and Multiagent Systems(AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,IFAAMAS, 9 pages.
3. [Dignum, 2017] Virginia Dignum, Responsible Autonomy. In IJCAI, pp. 4698-4704, 2017.
4. [StuartRussell & al., 2015] StuartRussell, DanielDewey. And MaxTegmark, **Research Priorities for Robust and Beneficial Artificial Intelligence**, Copyright © 2015, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602. https://futureoflife.org/data/documents/research_priorities.pdf?x60419
5. [BONNET & al., 2017] Gaël BONNET, Bruno MERMET and Gaëlle SIMON *Formal verification of moral values in MAS,* ARTICLE VOL 31/4 - 2017 - pp. 449-470- doi:10.3166/ria.31.449-470
6. [Blass and Forbus, 2015] Joseph A. Blass and Kenneth D. Forbus. Moral decision-making by analogy: Generalizations versus exemplars. In AAAI, pp. 501–507, 2015.
7. [Yu & al., 2018] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser and Qiang Yang, Building Ethics into Artificial Intelligence, In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18) pp. 5527-5533, 2018.