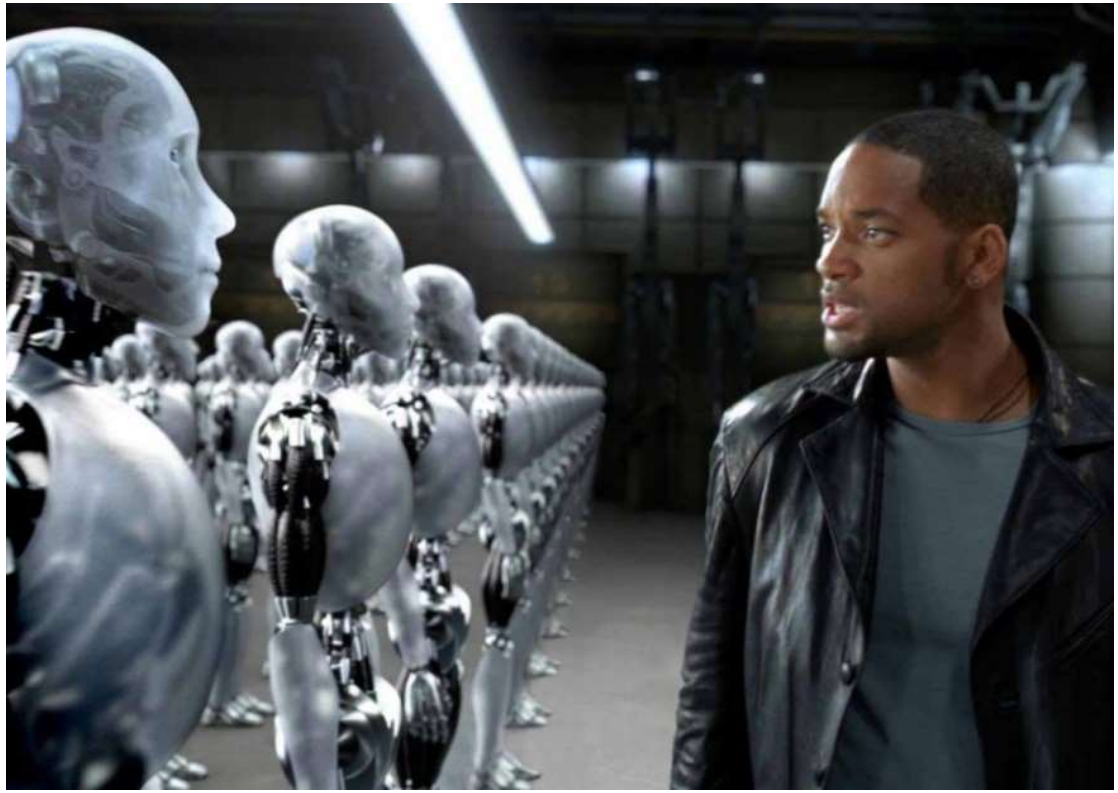


Teaching Ethics to Autonomous Systems:

Do androids dream of being good people?

AI4SE 2019
First Workshop on the application of Artificial Intelligence for Systems Engineering
Leganés, November 12-13, 2019

ggenovaf@gmail.com

References

https://gonzalogenova.wordpress.com/selected-publications/

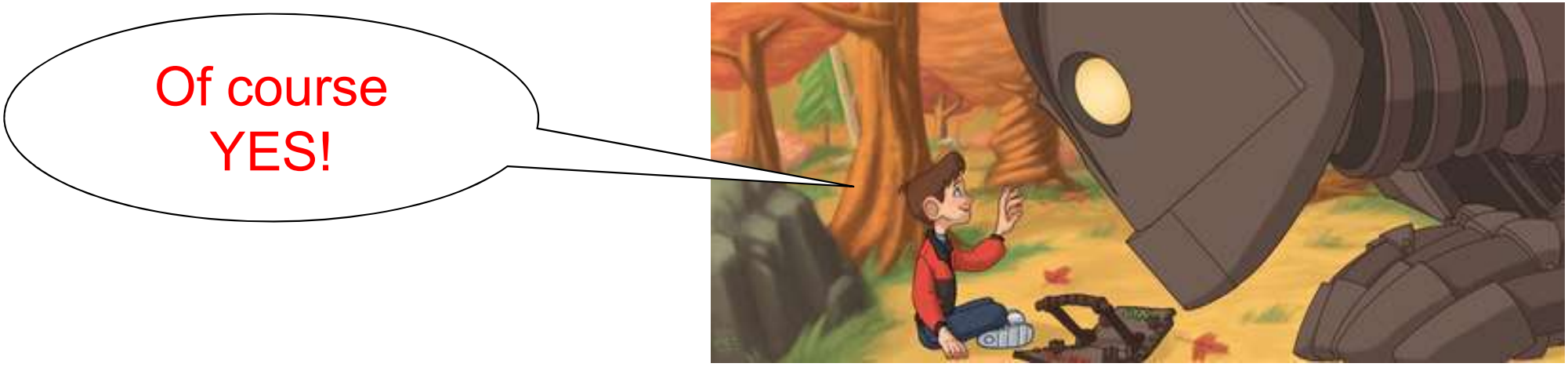
- Gonzalo Génova, Valentín Moreno. ***Teaching ethics to machines: Do androids dream of being good people?*** Science and Engineering Ethics (in preparation).
  - Gonzalo Génova. ***Can we teach ethics to machines?*** Ethics by Design Thematic day at PRIMA 2017: The 20th Int. Conf. on Principles and Practice of Multi-Agent Systems. Niza, Francia, Oct 30th 2017.
  - Gonzalo Génova. ***¿Podemos enseñar ética a las máquinas?*** Seminario Gregorio Peces-Barba en la Universidad Carlos III de Madrid (24/02/2017). https://www.youtube.com/watch?v=AFYP19ONDN4

- ***De máquinas e intenciones.*** https://demaquinaseintenciones.wordpress.com/
  - ¿Qué es una máquina? (13/07/2018)
  - ¿Puede ser libre una máquina computacional? (20/08/2018)

- Gonzalo Génova, Ignacio Quintanilla Navarro. ***Are Human Beings Humean Robots?*** Journal of Experimental & Theoretical Artificial Intelligence, 30(1):177–186, Jan 2018.
  - Gonzalo Génova. ***¿Ha llegado la era de las máquinas libres?*** Conferencia en Universidad Politécnica de Madrid. Madrid, 11 de abril de 2018. https://www.youtube.com/watch?v=hnZwrIdPwHw

- Gonzalo Génova, M. Rosario González. ***Educational Encounters of the Third Kind.*** Science and Engineering Ethics, 23(6):1791-1800, December 2017.

Can we teach ethics to machines?



The Iron Giant (Brad Bird, 1999)

The ethical algorithm and the ethics of algorithms

Charles Leonard

Seguir

One of the most challenging aspects of artificial intelligence (AI) is ***including ethics in the formula****. No matter how good machines become at mimicking and surpassing a human skill, there will be situations where* ***machines will have to make moral based decisions****.*

https://www.linkedin.com/pulse/teaching-ethics-machines-charles-leonard

The Moral Machine

MORAL MACHINE

Home Judge Design Browse About Feedback

Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. You can then see how your responses compare with those of other people.

If you're feeling creative, you can also **design** your own scenarios, for you and other users to **browse**, share, and discuss.

http://moralmachine.mit.edu/

https://youtu.be/XCO8ET66xE4 (44s)

The Trolley Problem

Philippa Foot (1920–2010)

https://en.wikipedia.org/wiki/Trolley_problem
http://nyti.ms/1F3sCz7
http://naukas.com/2015/03/03/el-problema-del-tranvia-o-mato-al-gordo/

Artificial ethics: ethical algorithms

Student	Participation (10%)	Homework (50%)	Exam (40%)	Weighted average	Decision
Alicia	9	10	9	9,5	Pass
Isabel	5	5	2	3,8	Fail
Jaime	10	8	8	8,2	Pass
Laura	9	5	4	5,0	Pass
Nicolás	10	10	8	9,2	Pass
Pablo	5	4	6	4,9	Fail

Algorithms that **compute** a decision:

1. Who passes/fails.
2. Who receives a loan.
3. Who is the most suitable patient to receive a transplanted organ.

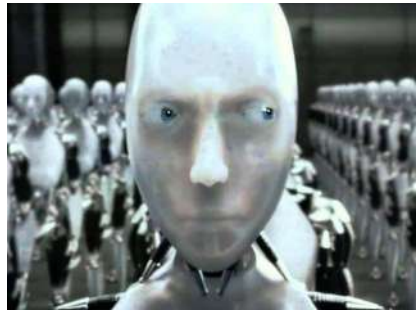
including ethics in the formula: there will be situations where machines will have to make ***moral based decisions***

We already use algorithms to make ethically-loaded decisions.

Explicit artificial ethics: The Three Laws of Robotics

Isaac Asimov
(1920–1992)

I, Robot
(Alex Proyas, 2004)



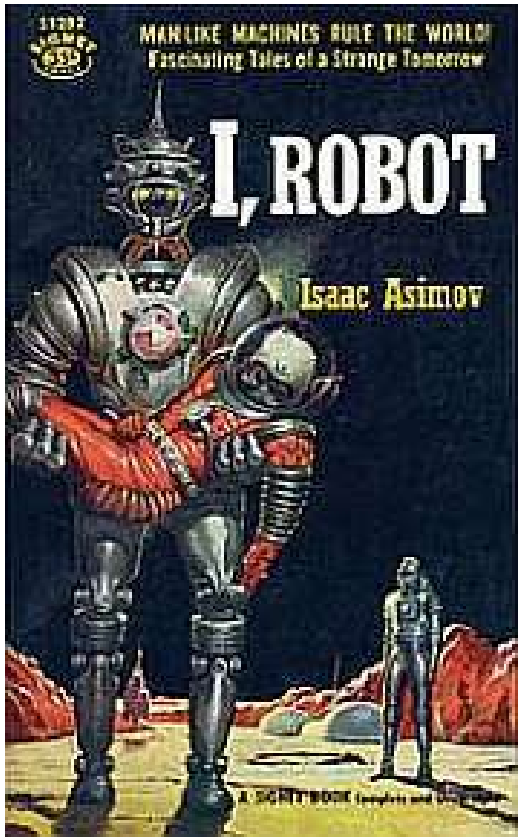
1. A robot may not injure a human being
or, through inaction, allow a human being to come to harm.

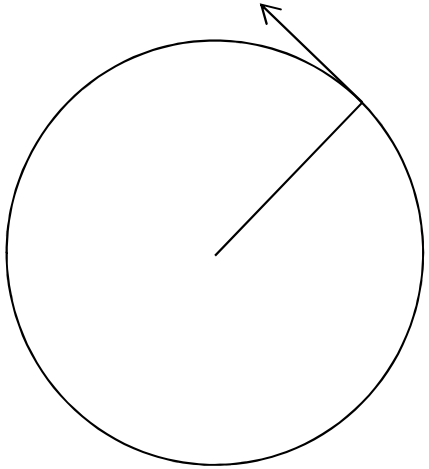
2. A robot must obey the orders given it by human beings,
except where such orders would conflict with the First Law.

3. A robot must protect its own existence
as long as such protection does not conflict with the First or Second Laws.

https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

Runaround / Buridan's Ass





http://www.singularitysymposium.com/laws-of-robotics.html
http://proyectokoan.com/671-2/

Are they implementable

(in a computational machine)

1. A robot may not injure a human being
or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings,
except where such orders would conflict with the First Law.

3. A robot must protect its own existence
as long as such protection does not conflict with the First or Second Laws.

Laws?
Asimov?
I don't get it…

Dificulties: all consequences?

1. A robot may not injure a human being
or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings,
except where such orders would conflict with the First Law.

3. A robot must protect its own existence
as long as such protection does not conflict with the First or Second Laws.

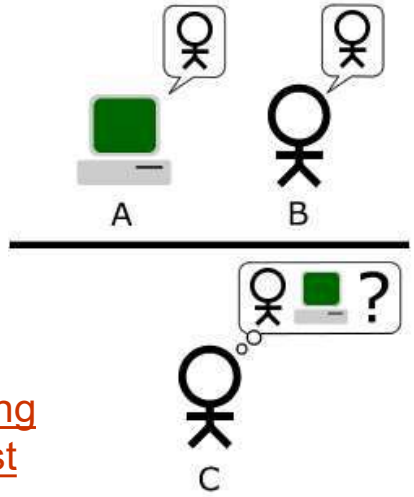
https://www.linkedin.com/pulse/getting-results-all-creating-domino-effect-your-life-aldo-moller

Dificulties: what is a human being?

1. A robot may not injure a human being
or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings,
except where such orders would conflict with the First Law.

3. A robot must protect its own existence
as long as such protection does not conflict with the First or Second Laws.



Turing Test

Duck Dodgers in the 24½th Century
https://www.youtube.com/watch?v=OEXvsGe657Q (57s)

This is serious, man!

Dificulties: are *deontic premises* "true"?

1. A robot may not injure a human being
or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings,
except where such orders would conflict with the First Law.

3. A robot must protect its own existence
as long as such protection does not conflict with the First or Second Laws.

How do we distinguish the Good from the Evil "on the face of it"?

Is the difference empirically verifiable?

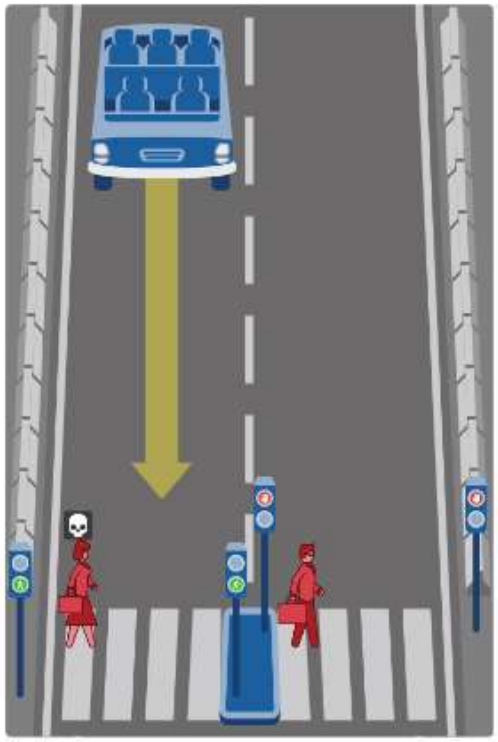
Self-programed artificial ethics: learning by imitation

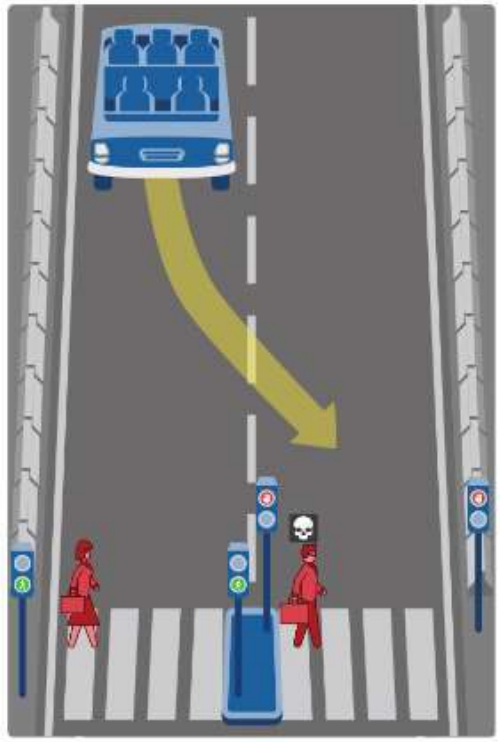
Why these percentages 10-50-40?
Let's ask the teachers…

Teacher	Participation	Homework	Exam	Sum
Prof. Castells	10%	50%	40%	100%
Prof. Guijarro	40%	20%	40%	100%
Prof. Jiménez	10%	10%	80%	100%
Prof. Muiño	20%	20%	60%	100%
Prof. Trujillo	50%	25%	25%	100%
Prof. Zaldívar	15%	55%	30%	100%
Average A	24,2%	30,0%	45,8%	100%
Average B	27,0%	34,0%	39,0%	100%

"Customize the ethical decision to the maximum common denominator…"

Even if the driving skills of the car can be standardized, its "ethics" would have to be ***customized to the maximum common denominator of the territory*** *where it will be used.*





We can make a machine learn, that is, imitate, ethical decisions.

http://moralmachine.mit.edu/

Artificial ethics: problems raised

The limits of the imitation game

Mechanical, programmed behavior

- Prophesying the future
- Recreate unjust historical **biases**
- Accountability (responsibility)
- Transparency, explanability of decisions
- Corruption of AI systems
- Unwanted, unintentional effects

Biases…

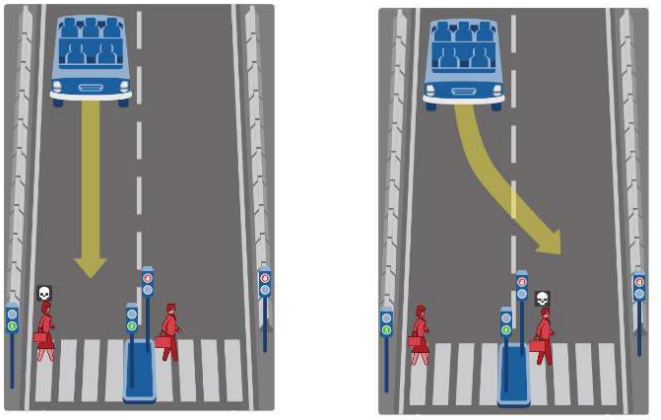
But, what is a bias?

The underlying problem: what is teaching ethics?

Estudiante	Participación (10%)	Trabajo (50%)	Examen (40%)	Promedio ponderado	Decisión
Alicia	9	10	9	9,5	Aprueba
Isabel	5	5	2	3,8	Suspende
Jaime	10	8	8	8,2	Aprueba
Laura	9	5	4	5,0	Aprueba
Nicolás	10	10	8	9,2	Aprueba
Pablo	5	4	6	4,9	Suspende

Computable decisions
(explicit rules)

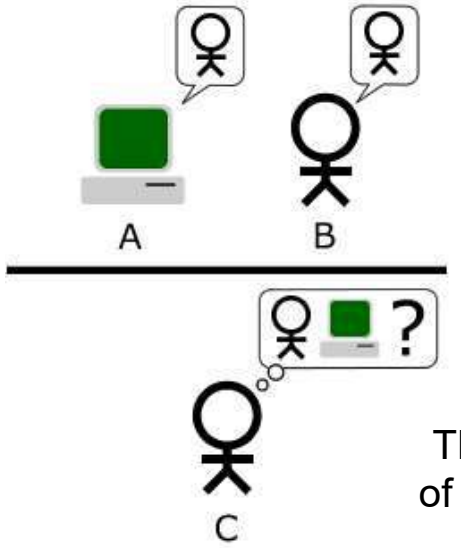
Imitable behavior
(implicit rules)



***Ethics* is not only, nor principally, following a *code of conduct* or *imitating* the behavior of others.**

- Recognize human beings.
- Recognize human dignity
- Recognize ethical values.
- Recognize good and evil on oneself's own.

This is not computable.



The Test of Dignity

http://en.wikipedia.org/wiki/Turing_test

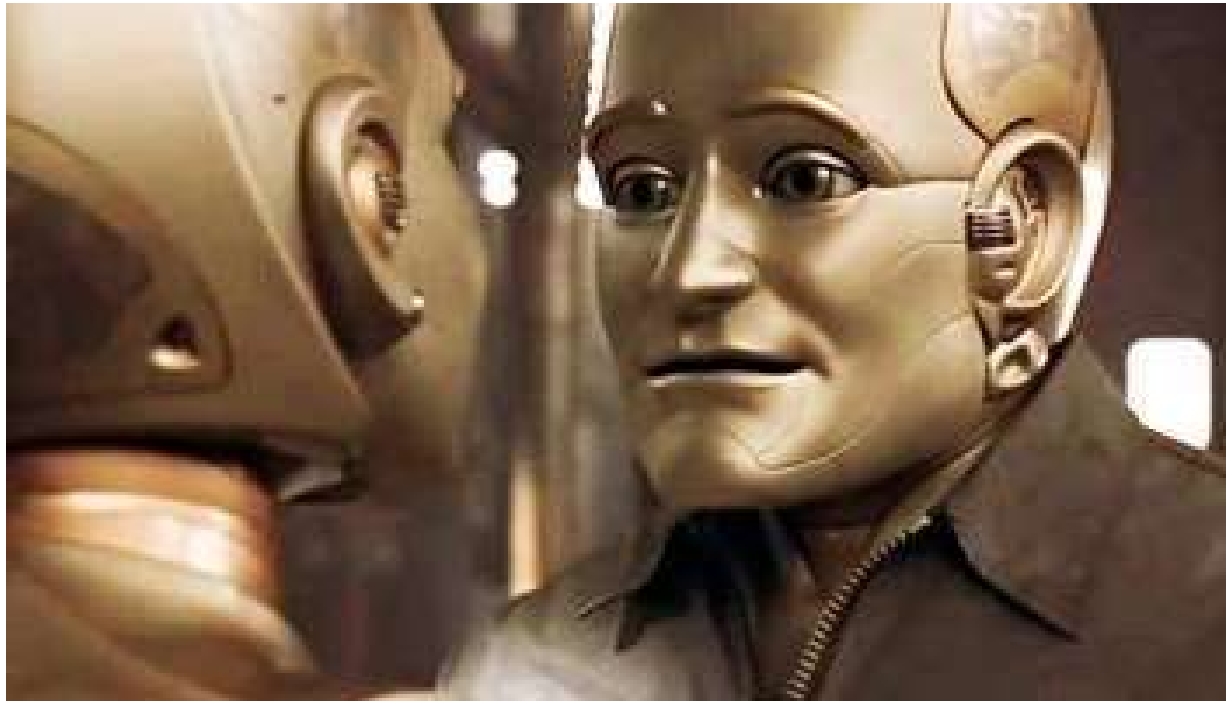
The human and the robot before the mirror

Outsourced intelligence

The humanization of the robot

The robotization of the human

The **rationality of the action** (of the ends) can not be solved algorithmically: the ends are **preconditions** of the algorithms.



Bicentennial Man (Chris Columbus, 1999)

Who makes the decision?

We invent artifacts to solve problems.

But ***we*** are more than ***problem solvers***.

We are capable of self-proposing objectives.

Or not?

The abolition of man?