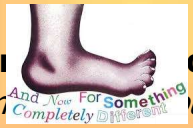
Aspetos Profissionais e Sociais da Engenharia Informática

This is a philosophy lesson...

but will take some thinking for you to understand the title



Rui L Aguiar, UA/IT

2

Objective of this class

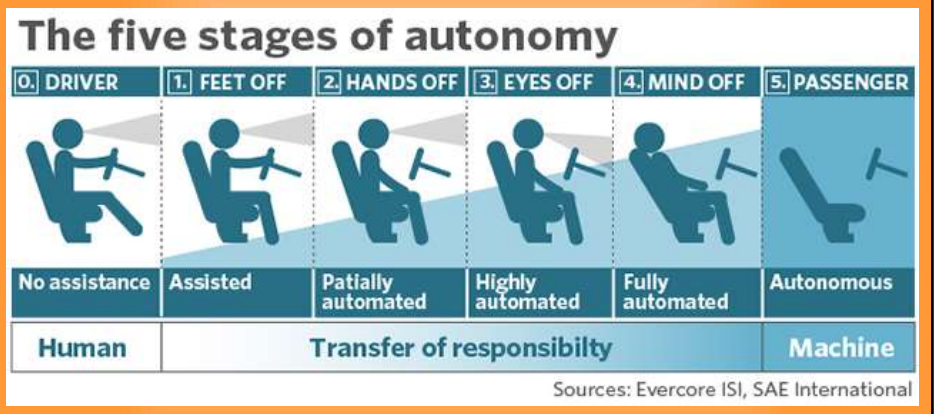
- Notions on several concepts
  - Ethics
  - Identity
  - Impact in Informatics

3

Lets have fun - I

4

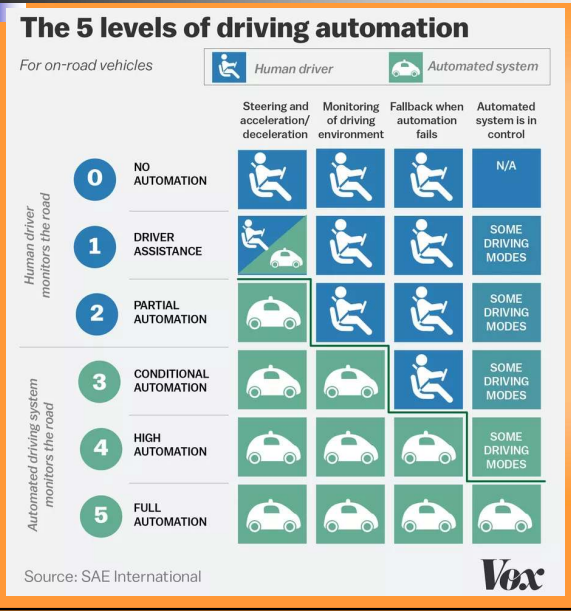
Context – why not Automated Driving?



5

5

NHTSA/SAE Levels of Automation



National Highway Traffic Safety Administration, U.S. federal government that is responsible for regulating and enforcing safety standards for motor vehicles and related equipment.
Society of Automotive Engineers, which is an international professional organization that develops standards and guidelines for the automotive industry.

6

6

The original "trolley problem" (1905)

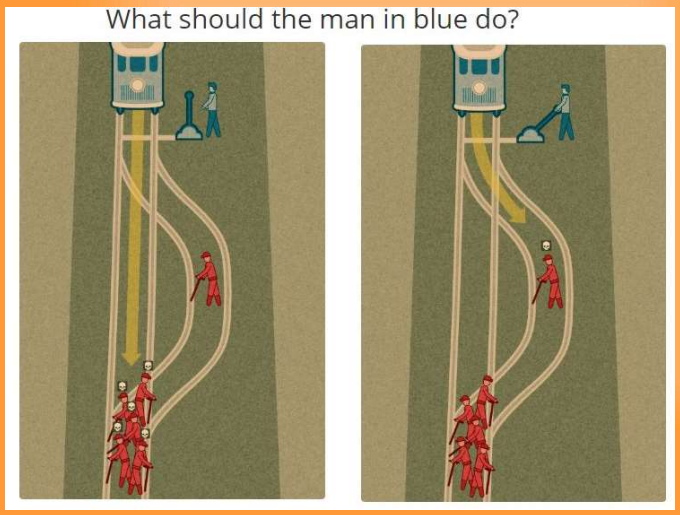


Image credit: moralmachine.mit.edu

7

Lets have fun - I

VOTE

8

Reasons – class built!

- Move lever
  - Principle of maximum hapiness
  - Minor certainty on the outcome
  - Making people responsible per actions (sleeping person)
- Do not move lever
  - Unable to take decision
  - Not accepting responsibility
  - Fate

9

9

Reasons – class built in 2023!

- Move lever
  - It is the smallest evil
  - It is the rationale for utility function (a-la Stuart Mill)
  - Assuming:
    - All persons are equal
    - I do not know anyone in the scenario
- Do not move lever
  - It is a crime
  - I cannot make judgments on people I do not know
    - It is not my right to to decide who lives and who dies

10

10

The original "trolley problem" (1905)

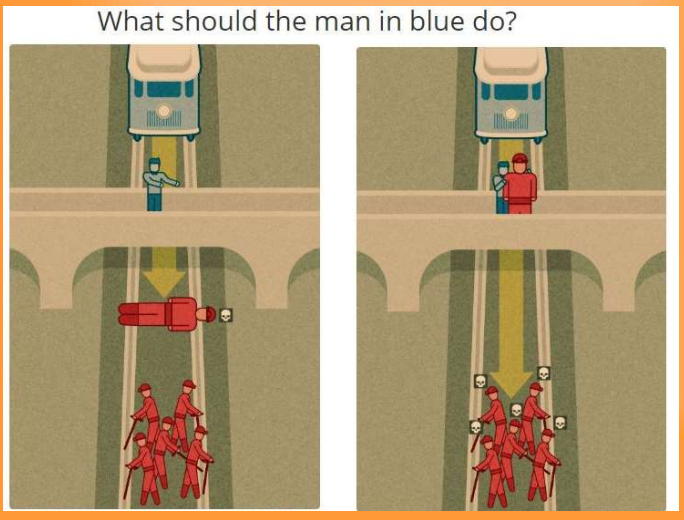


Image credit: moralmachine.mit.edu

11

Lets have fun - I

VOTE

12

Reasons – class built!

- Move lever
  - Principle of maximum hapiness
  - Minor certainty on the outcome
  - Making people responsible per actions (sleeping person)
- Do not move lever
  - Unable to take decision
  - Not accepting responsibility
  - Fate
  - Making people responsible
  - Personal perception associated in action
  - The tool (the person) is innocent.

13

13

Reasons – class built in 2023!

- Push person
  - It is the smallest evil
  - It is the rationale for utility function (a-la Stuart Mill)
  - Assuming:
    - All persons are equal
    - I do not know anyone in the scenario
- Do not push person
  - It is a even larger crime
  - I cannot make judgments on people I do not know
    - It is not my right to to decide who lives and who dies
    - The person on the bridge has no fault on the situation

14

14

Had fun?

- Have you heard of police and the courts?
  - In general…
  - You will be held strongly accountable of ACTING
  - You will be less chargeable (if at all) for FAILING to ACT.

Details will be important: are you driving the trolley? Are you watching? Were you able to think it over? Are you related with any person in question?

Side note: there are different philosophy theories associated to these cases (*utilitarism, intention, etc..)*

15

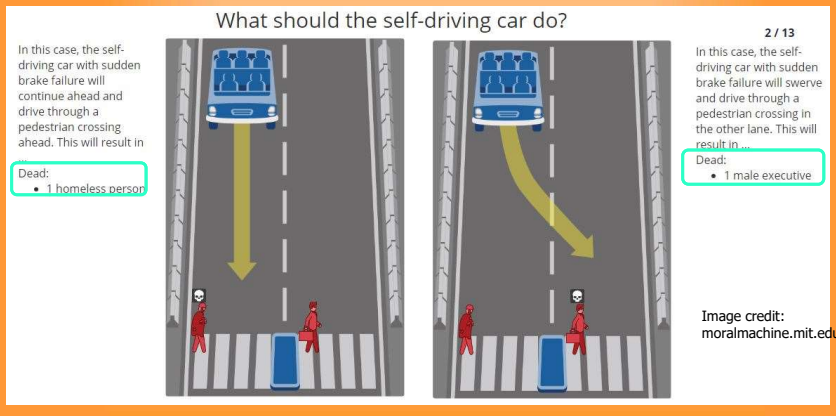
15

MIT's Moral Machine

- MIT moral experiment for automative drive
- Multiple scenarios, multiple personas
- 2 million people, 10 languages, 233 countries = 40 million decisions
- In the main interface of the Moral Machine, users are shown unavoidable accident scenarios with two possible outcomes, depending on whether the autonomous vehicle swerves or stays on course
- **Results were VERY VARIANT of country and culture!**

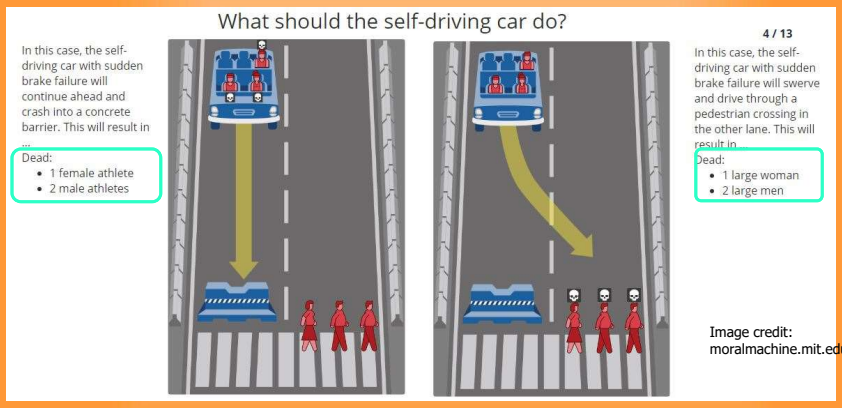
16

The Moral Machine



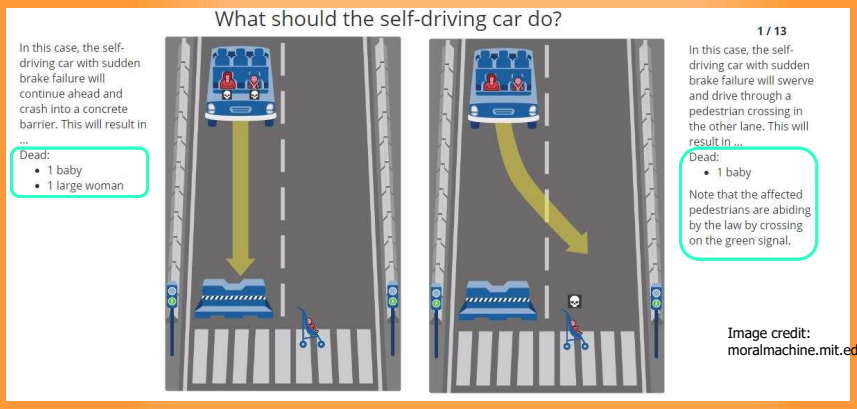
17

The Moral Machine



18

The Moral Machine



19

The Moral Machine

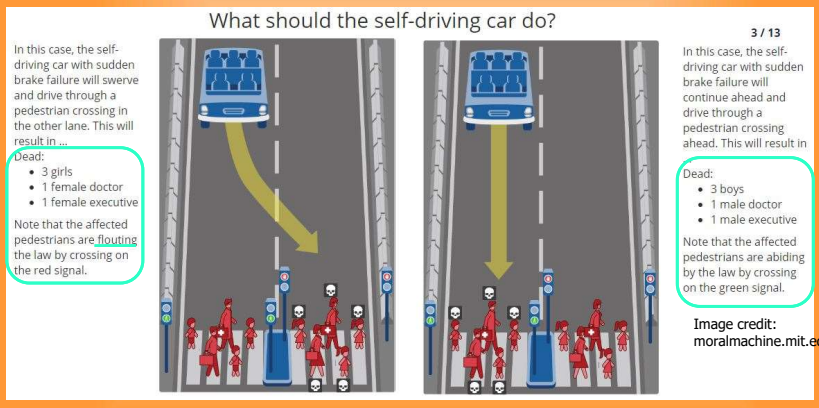


Image credit: moralmachine.mit.edu

20

The Moral Machine

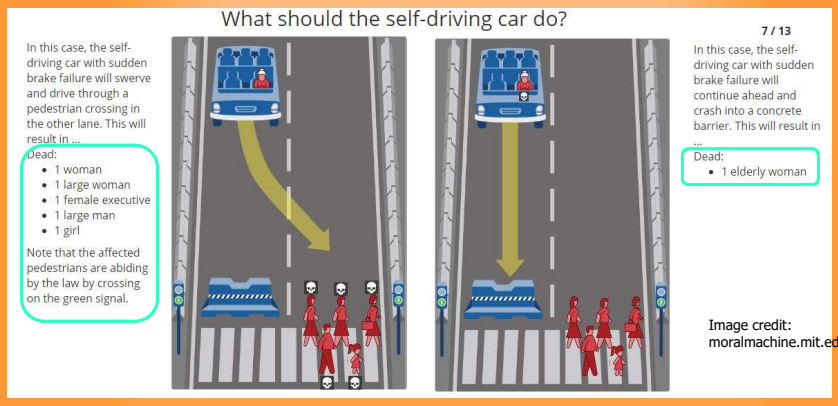


Image credit: moralmachine.mit.edu

21

The Moral Machine

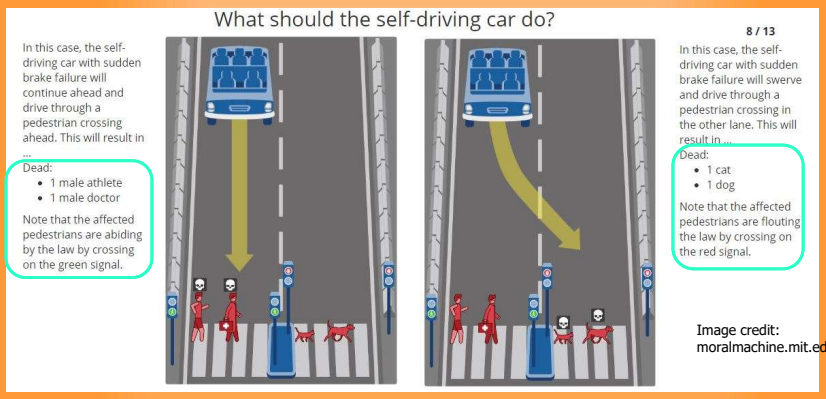
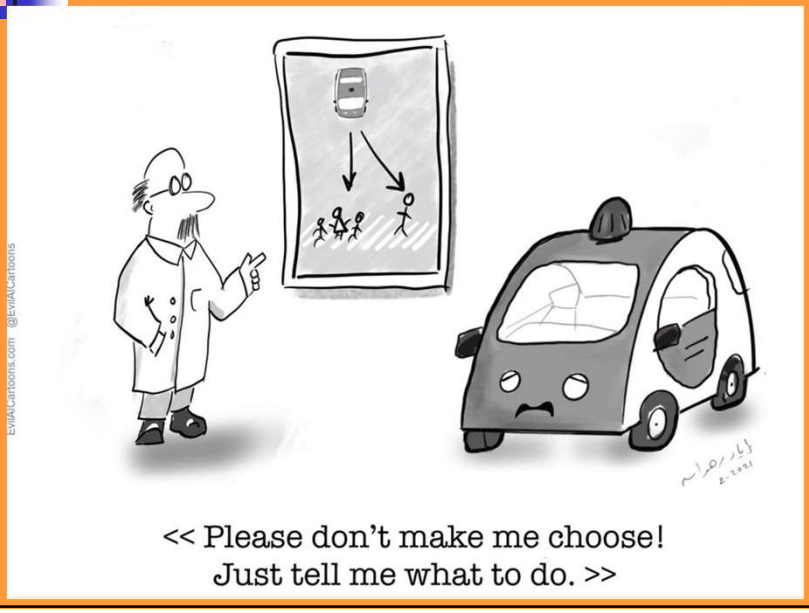


Image credit: moralmachine.mit.edu

22

Lets return to automative driving



23

Easy for an engineer…

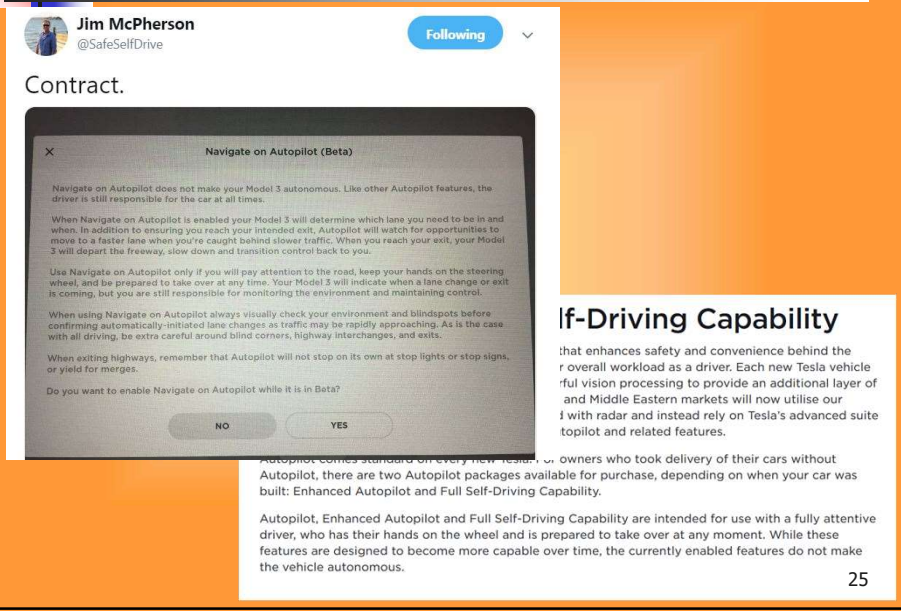
- Engineers are better than these abstract dilemmas, right?
  - Do you plan for "brake failures" which may lead to "unavoidable accidents" ?
  - Do you drive fast enough in the vicinity of pedestrian crossings that they would kill people in the case of a crash? And if you are in a rush?

"The only safe scenario would be don't move. You have to make reasonable assumptions about what you care about and what you don't."

Aaron Ames.

24

Do you remember that "law" thingy?



25

And if we make it more complex? Communication (distributed systems)

Car2car info

Car2pedestrian info

26

And what now for engineers?

"A car with any level of autonomy that relies upon a human to save the day in an emergency poses almost insurmountable engineering, design, and safety challenges, simply because humans are for the most part horrible backups. They are inattentive, easily distracted, and slow to respond." - Coelingh, Volvo.

"The truth is that civilization does not protect us from wild animals. It attempts, however imperfectly, to protect us from ourselves." – Michael Crichton, Travels

We should not absolve ourselves of the responsibility to act morally as drivers, or plan and engineer a moral environment – Tara Goddard

"We do not have the luxury of giving up on creating moral machines" – Moral Machine team from MIT

27

Legal – Informatics ?

Law ?

28

(Mis)evaluation impact in real life

"The good news is that you're perfectly healthy —the bad news is that my algorithms predict you'll be dead in two days..."

Think:
- Employment impact
- Insurance impact
- Reputation impact

(picture source: *Elliott*, AI Cartoons <https://timoelliott.com/blog/cartoons/artificial-intelligence-cartoons>)

29

AI and Security

- Attack AI systems
  - Learning
    - Cause learning system to not produce intended/correct results
    - Cause learning system to produce targeted outcome designed by attacker
  - System/learning
    - Learn sensitive information about individuals

  ⟹Need security in learning systems

- Misuse AI
  - Use AI to attack other systems
    - Find vulnerabilities in other systems
    - Target attacks
    - Devise attacks

  ⟹Need security in other systems

30

How to overcome these AI social issues

Who can regulate the use of AI?

- **European law:** if there is a reference to the internal market and thus a need for legal harmonisation: e.g. differences between national AI regulations make cross-border activities more burdensome
- **International law** (e.g. "European Ethical Charter on the use of AI in judicial systems and their environment" of the Council of Europe)
- **National law**
- **Professional codes** - self-regulation as a "privilege" of the liberal professions

31

High-Level Expert Group on AI: Ethics Guidelines

Created to drive "responsible" AI usage

4 ethical principles:	7 core requirements:
1. Respect for human autonomy 2. Prevention of harm 3. Fairness 4. Explicability	1. Human agency and oversight 2. Technical robustness and safety 3. Privacy and Data Governance 4. Transparency 5. Diversity, non-discrimination and fairness 6. Societal and environmental wellbeing 7. Accountability

(source: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines)

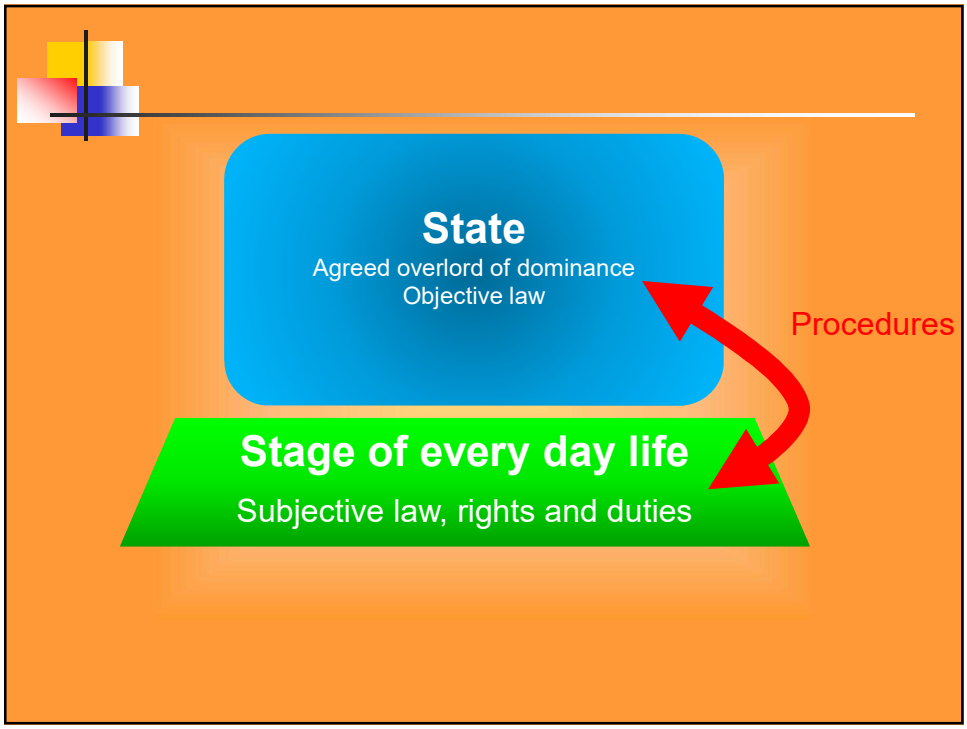
32

AI and human rights:
Charter of Fundamental Rights, ECHR, constitutions

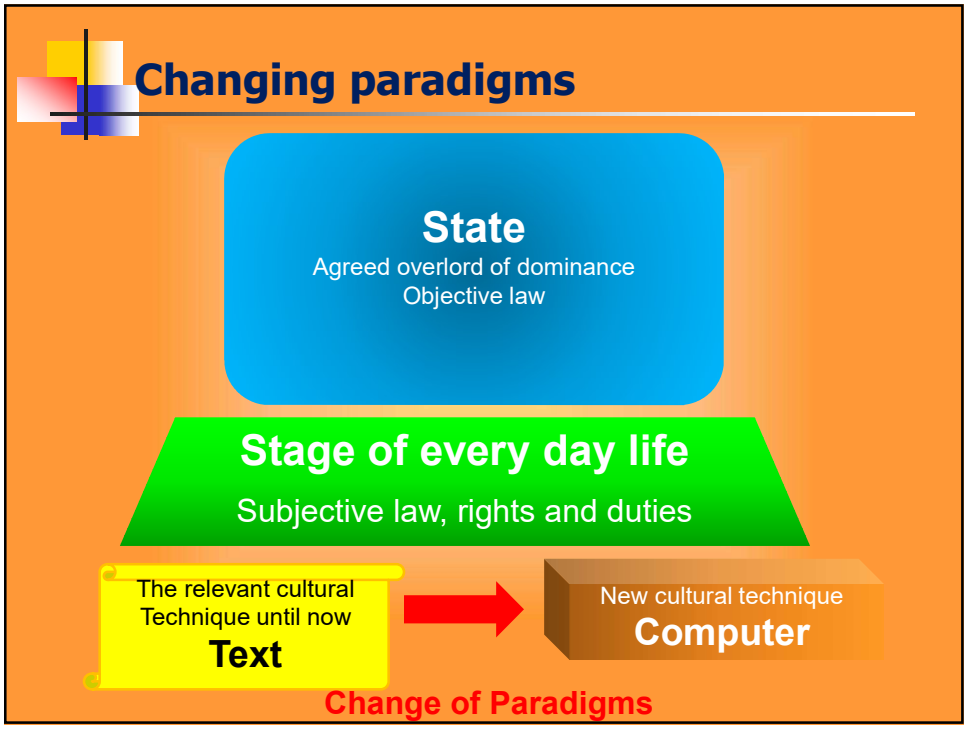
- **Responsibility for the consequences of innovation**:
  - The state guarantees protection from negative effects of technological innovation
  - Principle of non-discrimination – **Attention: correlation instead of causality**
- **Freedom of innovation**: Securing the freedom for technical development - Freedom to conduct business, right to (intellectual) property

(**Example: Necessary standard** of medical treatments: **Obligation to use AI?**
(e.g. ECHR 30.8.2016, 40448/06 *Aydoğdu/Turkey:* functioning hospital system)

33


State
Agreed overlord of dominance
Objective law
Procedures
Stage of every day life
Subjective law, rights and duties

34


Changing paradigms
State
Agreed overlord of dominance
Objective law
Stage of every day life
Subjective law, rights and duties
The relevant cultural
Technique until now
Text
New cultural technique
Computer
Change of Paradigms

35

Law

- Laws are brought about by tension, agitation and conflict by dramatic situations.
- Laws are societal rules or regulations that are obligatory to observe.
- Laws protect the welfare and safety of society, resolve conflicts, and are constantly evolving.

36

36

Ethics and Morality

- Ethics is a set of moral principles and a code for behavior that govern an individual's actions with other individuals and within society.
- Morality is what people believe to be right and good, while ethics is a critical reflection about morality.

37

37

Law vs Ethics

- Law, and ethics are different but related concepts.
- Laws are mandatory to which all citizens must adhere or risk civil or criminal liability.
  - Legal bodies will impose the law
- Ethics relate to morals and help us organize complex information and competing values and interests to formulate consistent and coherent decisions.
  - Professional bodies may impose ethics (e.g. Ordem dos Engenheiros, Comités de ética ou deontologia)

38

38

EU AI Act

- Two main concepts
  - What is AI
  - Risk based-approach
- Enforced by EC and national authorities
  - Penalties may reach 6% turnaround worldwide
  - Potential AI Board and AI Office (Europe)
    - Code of conduct

39

39

EU AI Act Risk Categories

A risk based approach

Chatbots, deep fakes, emotion recognition

LIMITED RISK
(AI systems with specific transparency obligations)

Biometric ID, critical infrastructure, educational, border control

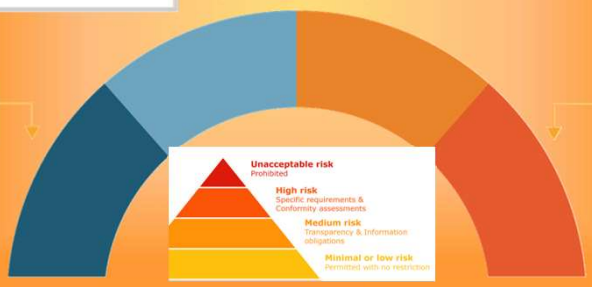
HIGH RISK

MINIMAL RISK

Spam filters, video games

UNNACEPTABLE RISK

Manipulation, Social scoring Biometrics (remote)



40

Requirements for high-risk AI systems

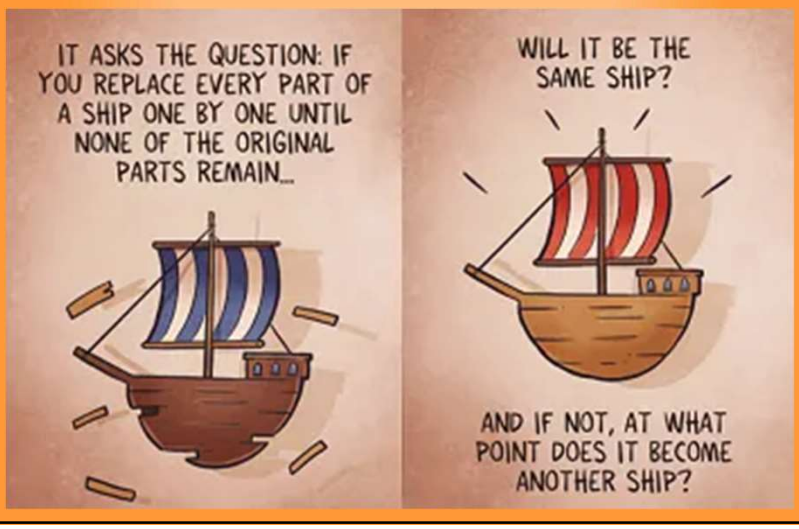
41

Lets have fun - II

42

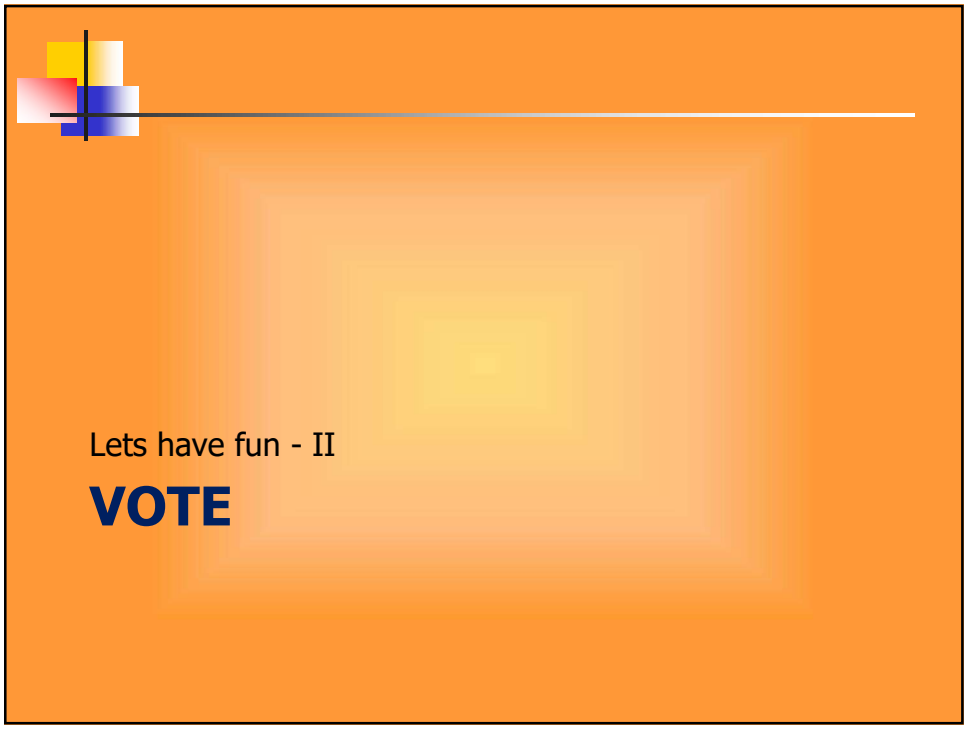
The ship of Theseus

- Classic philosophy problem

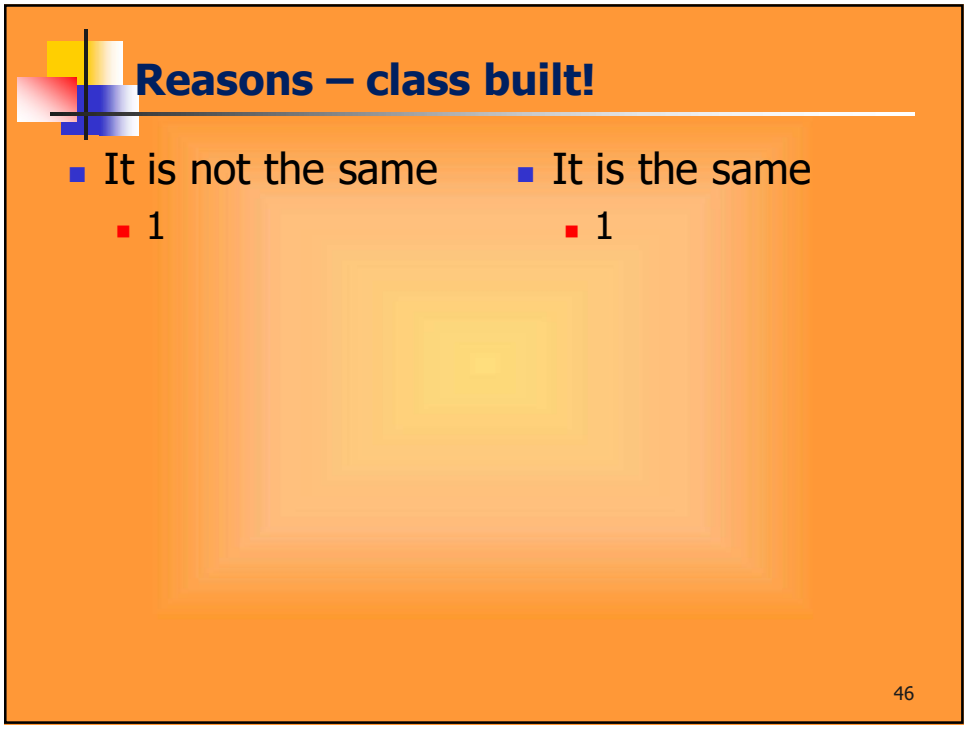


43

43


Lets have fun - II
VOTE

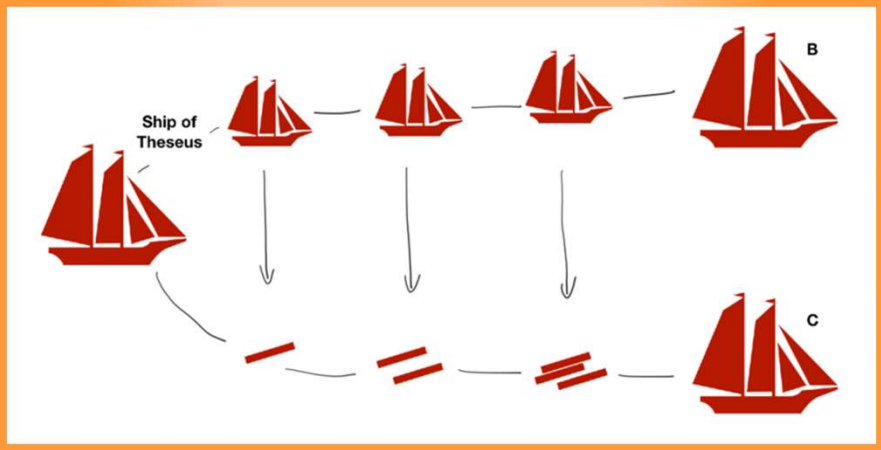
44


Reasons – class built!
It is not the same
1
It is the same
1
46

46

The ship of Theseus – Hobbes view

- Modification to the classic problem



47

47

Lets have fun - II

VOTE

48

Reasons – class built in 2023!

- It is B
  - Choice will depend on the process
    - It could be the other way around (develop the copies of the board as new, retain the old boat, and then build a new one with the new boards)
- It is C
  - The choice will depend on the process, indeed

49

49

Theories of Identity and Time

- What is the object? Does it change?

Presentism:

- Only those things that exist at the present time – *now* – really exist at all. The only qualities that these things really possess are the qualities that they possess now.

Eternalism:

- All things exist. Or, rather: past and future objects and times are just as real as currently existing ones. Reality is a four-dimensional spatiotemporal manifold of objects and events.
  - Reducionism: language can be reduced to scope the instant of the existence of the thing
  - Perdurantism: Objects persist through time by having different temporal parts, or "stages," at different times

51

51

But do you remember LAWs?

- What defines a boat *a boat*?
  - What are the properties that make the object part of a class?
  - In this case… a registration license, with a license plaque.
    - Remember the car registration?
    - But then is the registration license the *"boat"*?
    - This approach is implicitly following a concept known as *spacetime worms* associated to objects (eternalism).
  - How do you handle this associated with software objects?
    - Remember **"copy as is" concepts? "derivative works"?**

Side Note: Again, there are different philosophy theories on the subject (*eternalism, reductionism, presentism)*

52

WHAT ABOUT THE FUTURE?

53

Deep challenges in Informatics

- The definition of Identity
  - When is me really me?
  - I can clone myself digitally
- The association of a Digital Identity to a (real or virtual) object
  - Is this permanent? Can be transitive?
  - E.g. the DNS system works along these lines. The CC is another such system
- The structure and definition of the *Digital Identity identifier*

54

54

Lets' return to the autonomous car Virtual reality

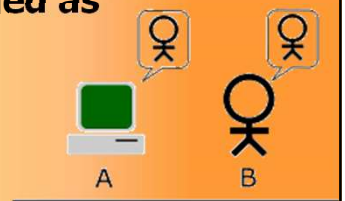
Do you understand why the concepts of identity and identifier are so important? This concept is known as *digital twin.*

Source: Crossing the road in the world of autonomous cars by Stamp Siripanich at Teague Labs https://link.medium.com/rLZQ1gC6GV

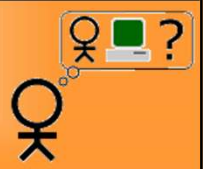
55

Another example: "intelligence" a-la Turing Test

- You ask the computer any question.
- If it is able to reply to it in the same manner a normal human would, **it is named as "intelligent"**.



- Turing test: chat room.
  - Many participants in a chat room.
  - One of the participants is a real person and one participant is a computer program.
  - The program passes the test (is considered intelligent) if no one can tell which of the two participants is human.

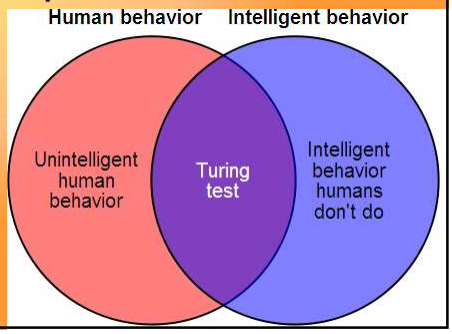


56

But naming is not the identity

- Turing test does not directly test if the computer behaves intelligently
- It only tests whether or not the computer behaves like a human being.
- In general the two concepts are different.

Understanding this separation between the digital world and the real world, and how we relate both is a fundamental tenant in informatics



57

Three Laws of Robotics (1940)

First Law: **A robot may not injure a human or through inaction, allow a human to come to harm.**

Second Law: **A robot must obey the orders given it by human beings, unless such orders would conflict with the first law.**

Third Law: **A robot must protect its own existence, as long as such protection does not conflict with the first or second law.**

58

58

Extending the Laws(?!)

Zeroth law: A robot may not injure humanity or through inaction allow humanity to come to harm.
(due to Asimov, Olivaw, and Calvin).

David Langford's, acknowledging military funding for robotics:
4. A robot will not harm authorized Government personnel but will terminate intruders with extreme prejudice.
5. A robot will obey the orders of authorized personnel except where such orders conflict with the Third Law.
6. A robot will guard its own existence with lethal antipersonnel weaponry, because a robot is bloody expensive.

60

60

Extending the Laws(?!)

Zeroth law: A robot may not injure humanity or through inaction allow humanity to come to harm.
(due to Asimov, Olivaw, and Calvin).

David Langford's, acknowledging military funding for robotics:
4. A robot will not harm authorized Government personnel but ...uders with extreme prejudice.
5. A rob... ...orized personnel ...conflict with the Third Law.
6. A robot will guard its own existence with lethal antipersonnel weaponry, because a robot is bloody expensive.

Sounds right?

61

61

Will AI Be Like Us?

Like us, AI systems...
...will talk to us in our languages.
...will help us with our problems.
...will have anthropomorphic interfaces.

Unlike us, AI systems...
...will compute and communicate extremely quickly.
...will have bounds for learning and retention of knowledge that will soon surpass ours.
...might not be well modeled by the psychological models that work for people.

63

63

Tools vs Agents

Agent: Takes responsibility, takes initiative, interacts with others on behalf of a client.

Tool: Responds directly to its user. Does not take responsibility. Does not take initiative. Does not normally interact with others on behalf of a client.

What will be the outcome?

64

64

Class answer?

65

65

High Levels of AI

- **Pros…**
  - **Powerful tools,**
  - **solutions to complex problems,**
  - **better society(?)**
- **Cons…**
  - **As a tools it might be used against people;**
  - **May create worse problems than it solves;**
  - **standards of living might get worse;**
  - ***we might lose some aspect of our humanity.***

66

66

Laws of robotics?

- Analyse the complexity required to follow "the laws"
  - See Asimov novels with freezing robots by indecision
- See:
  - Kurzweil – AI will supersede humans, and expand our values (*singularity)*
  - Stross – AI will not stop evolving for us, and evolution rate will be exponential

67

67

The Safe Systems approach

- Autonomous and semi-autonomous technologies should be developed within a Safe Systems framework, NOT the other way around
- Ex: automotive vision zero approach:

68

Food for thought: intelligence and identity

- Metaverse
  - Entities in the metaverse
  - Digital currencies
  - Non-fungible-tokens
- Services/Programs
  - All the communication stack!
  - All the process management!
- Legal frameworks rely in these concepts!
  - Uniqueness concepts
  - Digital world is challenging laws and regulations!

69

69