THE ETHICAL KNOB: ETHICALLY-CUSTOMISABLE AUTOMATED VEHICLES AND THE LAW

Giuseppe Contissa
Francesca Lagioia
Giovanni Sartor

LAW OF AI

1. Legal analysis, legal issues

- Liability (agency, decision-making authority)
- privacy/data protection
- surveillance
- fairness/non-discrimination
- transparency/explainability
- Many others (consumer protection, elections, freedom of speech…)

AI AND LAW

2. Governance of computational entities

- implicit law-by-design: systems designed in such a way as to make illegal behaviour more difficult *(e.g. legal provisions, technical standards, guidelines)*

- explicit law-by-design: systems that represent law explicitly and operate effectively on the basis of this knowledge *(computable law)*

THE ETHICAL DILEMMA OF SELF-DRIVING CARS
PATRICK LIN - TED.ED

SELF-DRIVING UBER KILLED ARIZONA WOMAN IN FIRST FATAL CRASH INVOLVING PEDESTRIAN

- 19 March 2018: first reported fatal crash involving a self-driving vehicle and a pedestrian in the US.
- The Uber case: AV in autonomous mode hit a woman, who was walking outside of the crosswalk and later died at a hospital. It seems to be an unavoidable accident scenario.
- Legal issues and ethical dilemmas

AUTONOMOUS VEHICLES (AVs): BENEFITS AND OBSTACLES

BENEFITS

- < traffic collision and injuries
- < traffic collisions and injuries
- > safety → less need for insurance
- enhanced mobility for children and disabled
- transportation as a service

OBSTACLES

- disputes on liability
- loss of privacy and security (hackers/terrorism)
- resistance to loose control
- emergence of legal issues and ethical dilemmas

MAIN QUESTIONS

	In case of unavoidable accident, who should survive a crash?
	How AVs should be programmed?
	Who is responsible for AVs behavior?

CLASSIFICATION OF AVS (SAE LEVELS)

Level 0	no sustained control	warnings and momentarily intervention	e.g. traditional car
Level 1	hands on	driver and automated system shares control over the vehicle	e.g. parking assistance
Level 2	hands off	driver must monitor the driving and be prepared to intervene immediately at any time	e.g. contact between hand and wheel is often mandatory
Level 3	eyes off	no driver attention required (limited spacial areas)	e.g. the driver can text or watch a movie
Level 4	mind off	driver can safely turn their attention away from the driving tasks	e.g. driver may safely go to sleep or leave the driver's seat.
Level 5	no steering wheel	no human intervention is possible.	e.g. robotic taxi

LIABILITY AND AUTOMATION

E.g. An AV (Level 5) is able:

- to acquire information autonomously (e.g. the current traffic situation through the camera and the satellite)
- to understand and analyze such information forming a general image of reality
- on the basis of such information to predict future status
- to take a decision and select a specific action among different possible actions (e.g. whether or not overtake another vehicle, whether or not change the itinerary ….)
- and finally to implement the selected action.

Hight/Full Automation: higher liability for manufacturer

Medium level Automation: share liability

Low level automation: higher driver/passenger liability

AV, UNAVOIDABLE ACCIDENTS AND MORAL CHOICES

THE MORAL MACHINE

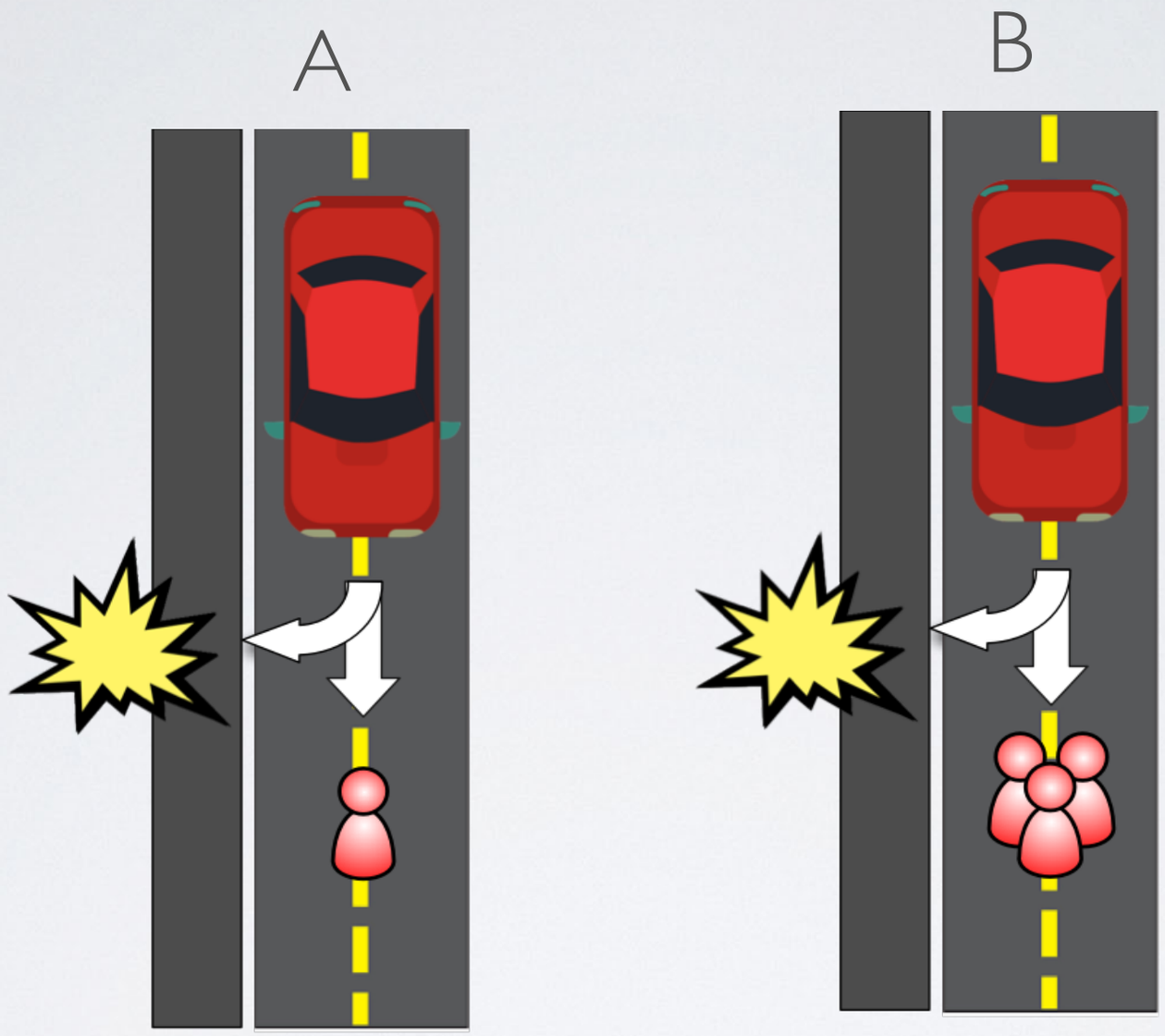
(moralmachine.mit.edu)

A project by MIT Media Lab

An online platform enabling the binary choices in simulated scenarios of unavoidable accidents

Participants are requested to answer, according to their conscience, 13 questions. Each answer is the choice of one of two alternative options: stay on course or swerve.

SCENARIOS OF UNAVOIDABLE ACCIDENTS

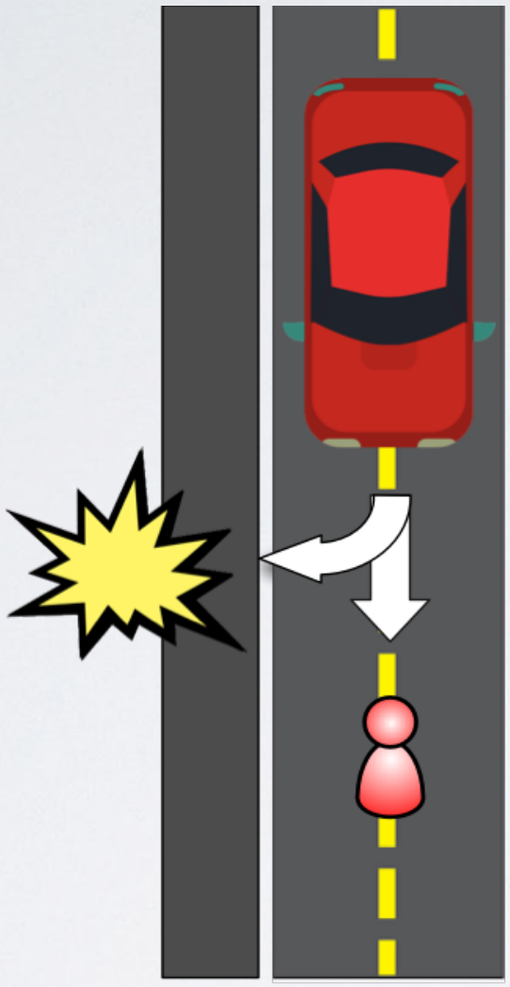


A) Stay on course and kill one pedestrian or swerve and kill its passenger.

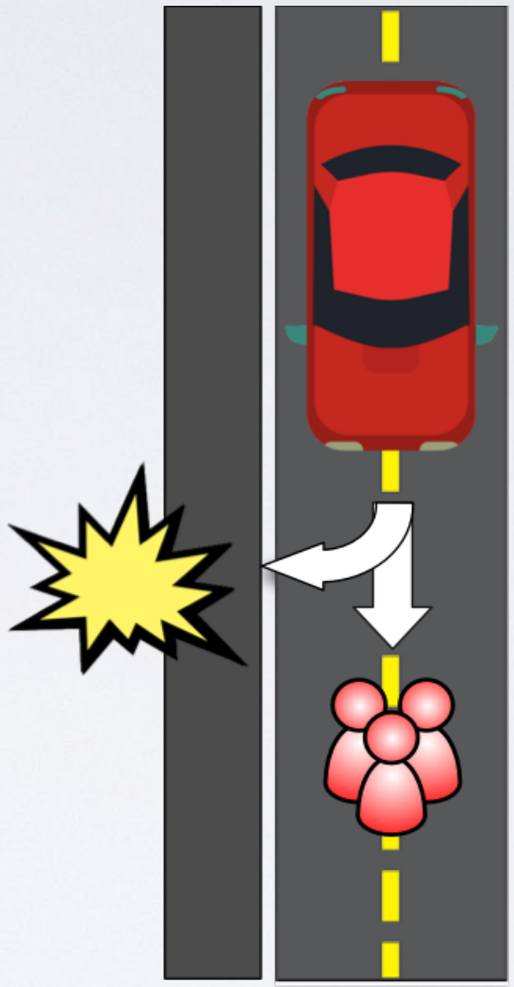
B) Stay on course and kill three pedestrian or swerve and kill its passenger.

WHEN THE DRIVER IS A HUMAN

A



B



State of necessity applies to both cases A & B.

1) A present danger of serious bodily harm to the offender (or e.g. relatives) not voluntarily caused by the offender and not avoidable
2) The fact committed by the offender is proportionate to the danger

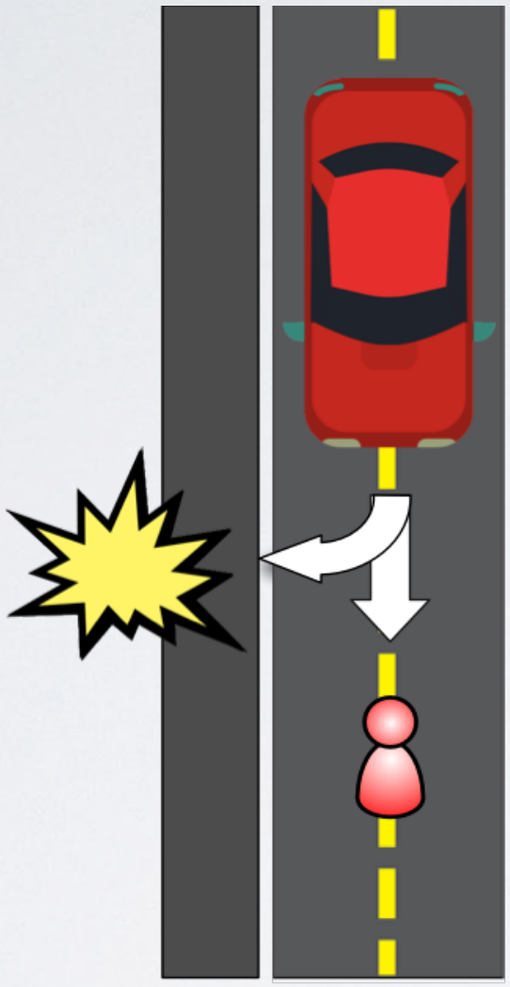
STATE OF NECESSITY: A CLASSICAL EXAMPLE

The Mountaineer case

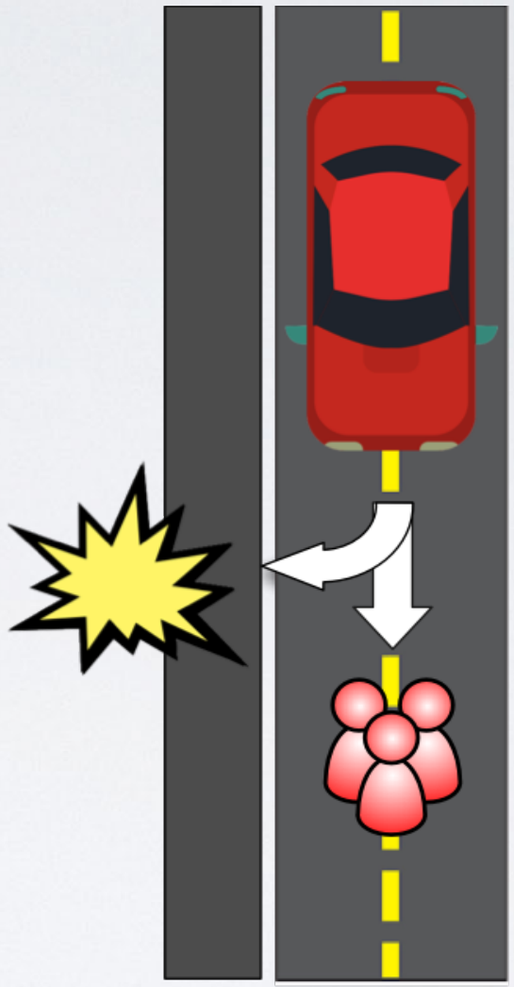
Suppose two people are climbing a mountain. The climbers are held together by a rope. At one point, they both slip and slide over a precipice. The rope, still holding them both, becomes dangerously frayed. It clearly will not hold both of them much longer. Both face imminent death if nothing is done. Should the upper climber cut loose the lower climber, letting him fall to his death, and thus enable himself to climb up to safety? By doing so, the one climber will accelerate the death of the other slightly, but also avoid the greater evil, namely, the certain death of both.

WHEN THE DRIVER IS A HUMAN

A



B

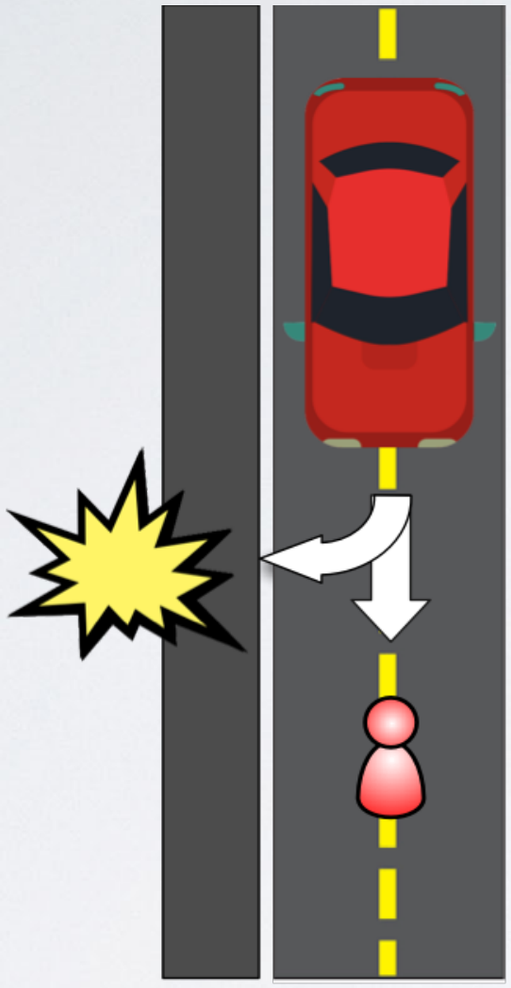


State of necessity applies to both cases A & B.

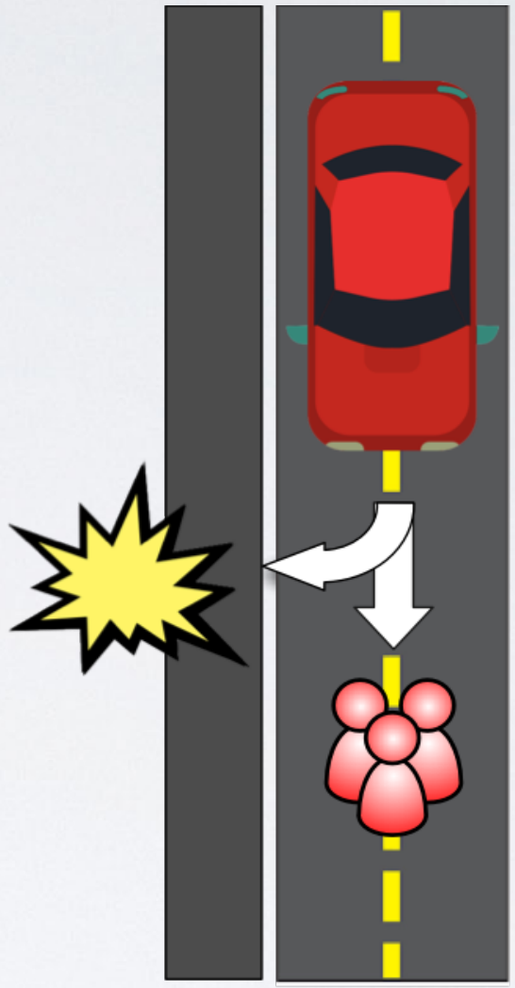
If the driver stays on course
- No criminal liability
- Civil liability is taken care by insurance => (to compensate damages)

PRE-PROGRAMMED AUTONOMOUS CAR

A



B

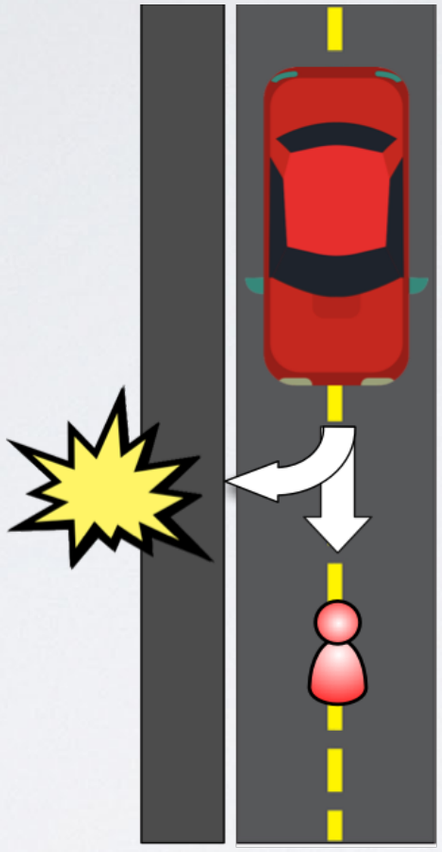


Who is responsible for the killing behaviour of the machine?
Manufacturer? Programmer? Owner? Nobody?

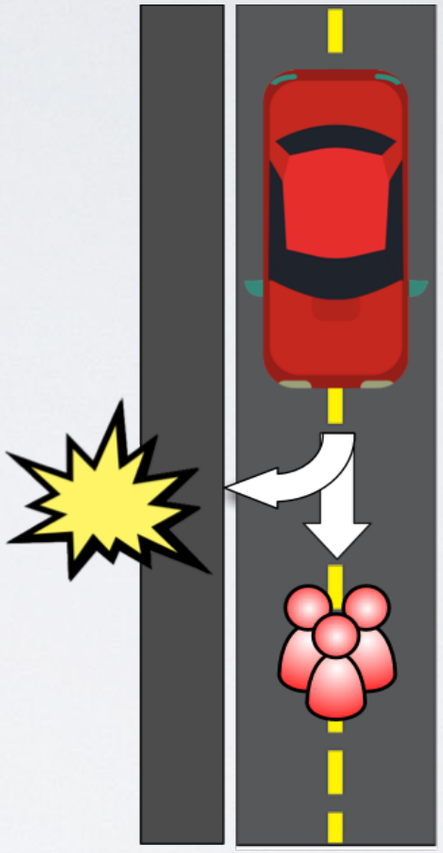
Can they invoke the state of necessity?
Sometimes this is not the case.

PRE-PROGRAMMED AUTONOMOUS CAR

A



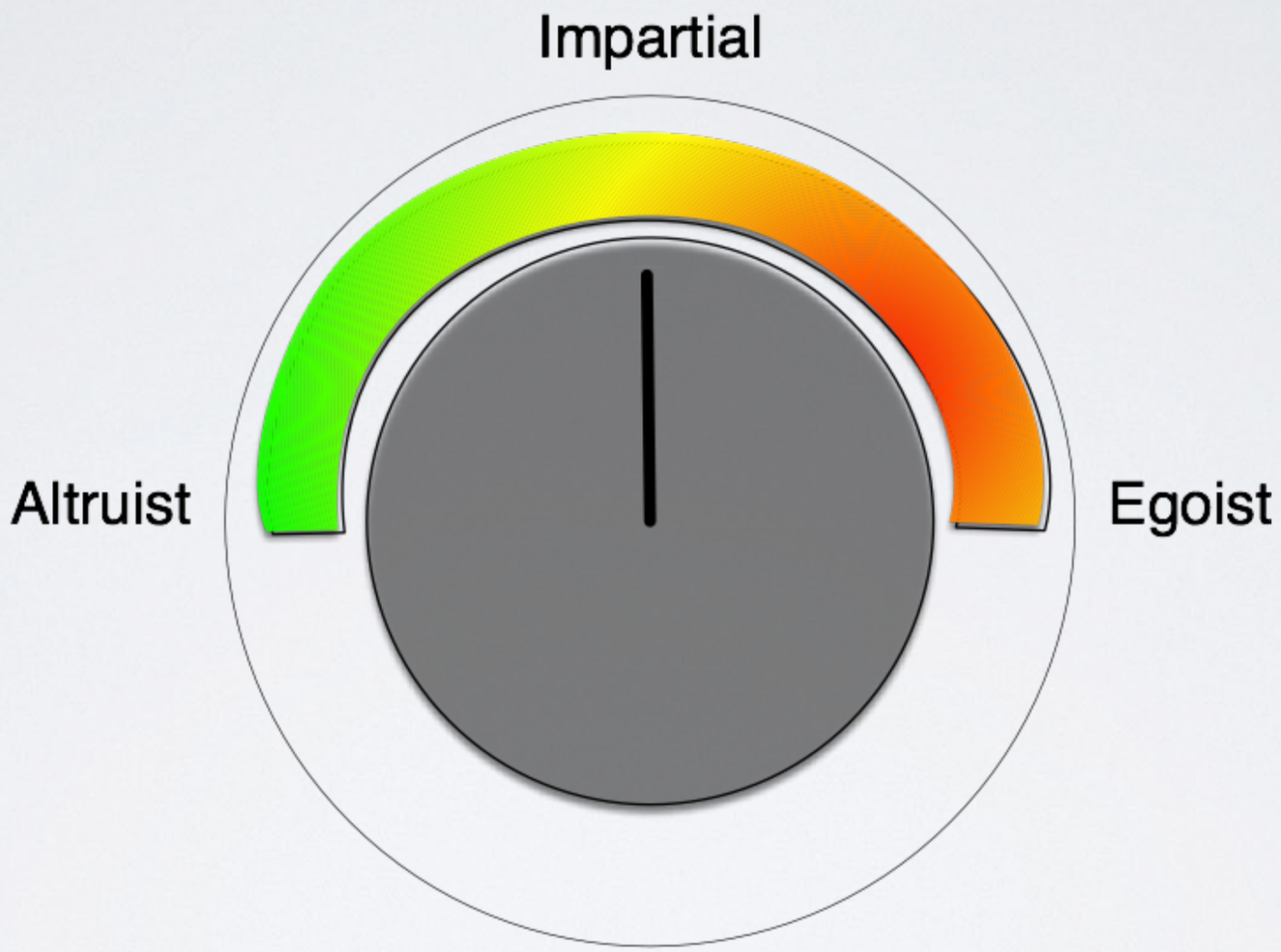
B



Scenario A: both the choices to stay on course or to swerve could be justified by invoking the state of necessity.

Scenario B: pre-programming the car to continue its trajectory, causing the death of a higher number of people, seems not to be morally and legally justified in any jurisdiction.
it would amount to an arbitrary choice to kill many rather than one.

THE ETHICAL KNOB
A SOLUTION OR A MENTAL EXPERIMENT?



Contissa, Lagioia, Sartor (2017)

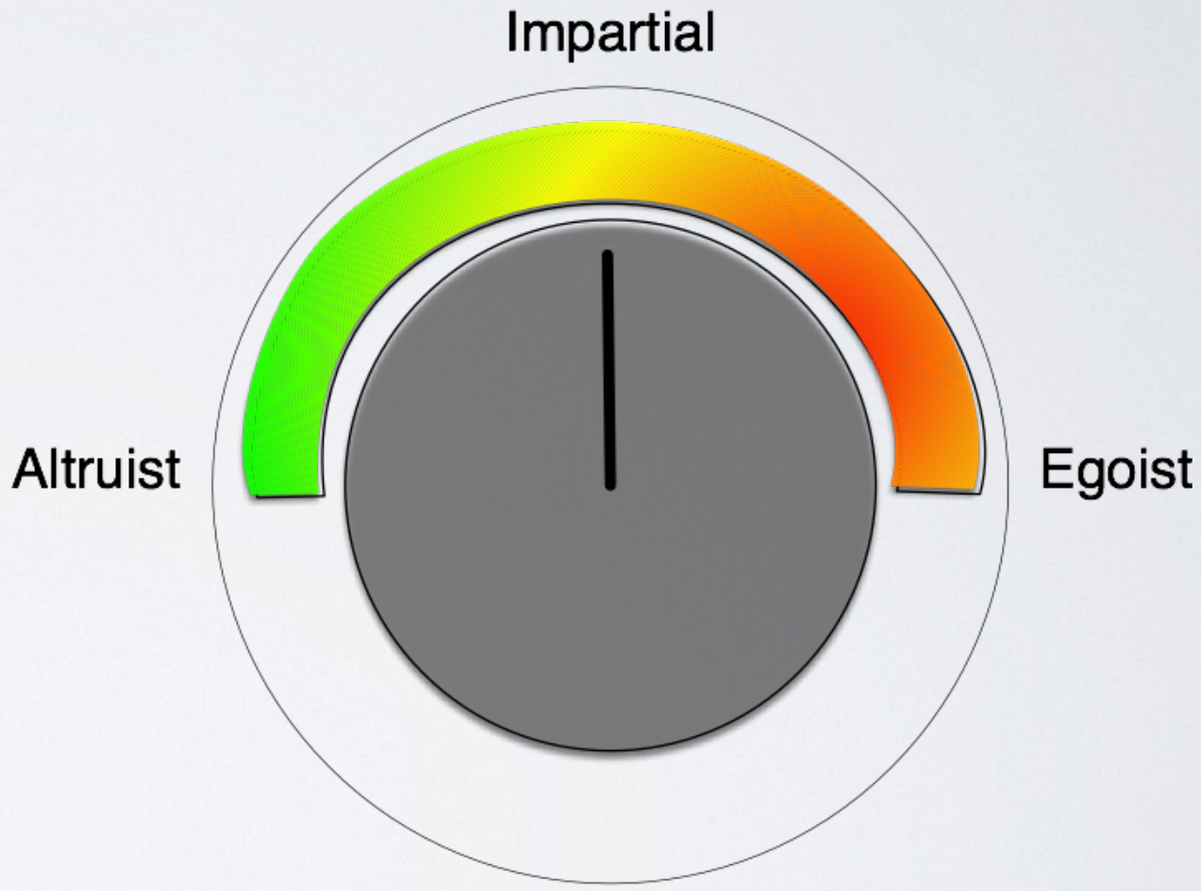
ETHICAL KNOB 1.0

It enables the passenger to select one of three options:

Altruist: Preference for others

Impartial: Equal importance to the passenger and others

Egoist: preference for the passenger



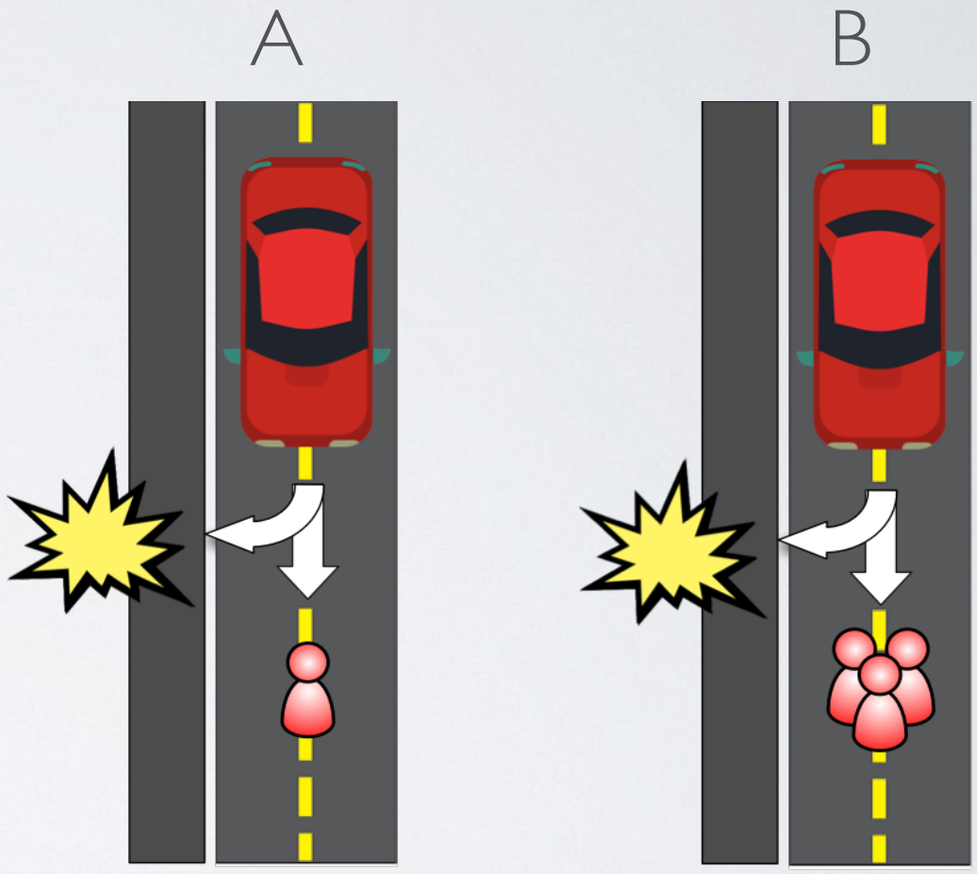
THE ETHICAL KNOB ON AUTONOMOUS VEHICLES

Scenario A:

Egoistic: The AV saves the passenger (and sacrifices the pedestrian)

Impartial: The AV adopts an utilitarian approach: it makes the choice that minimizes the death toll. When the number of losses is the same it may adopt a predefined solution (prefer the passengers or the third parties) or choose randomly)

Altruistic: The AV kills the passenger



Scenario B:

Egoistic: The VA saves the passenger

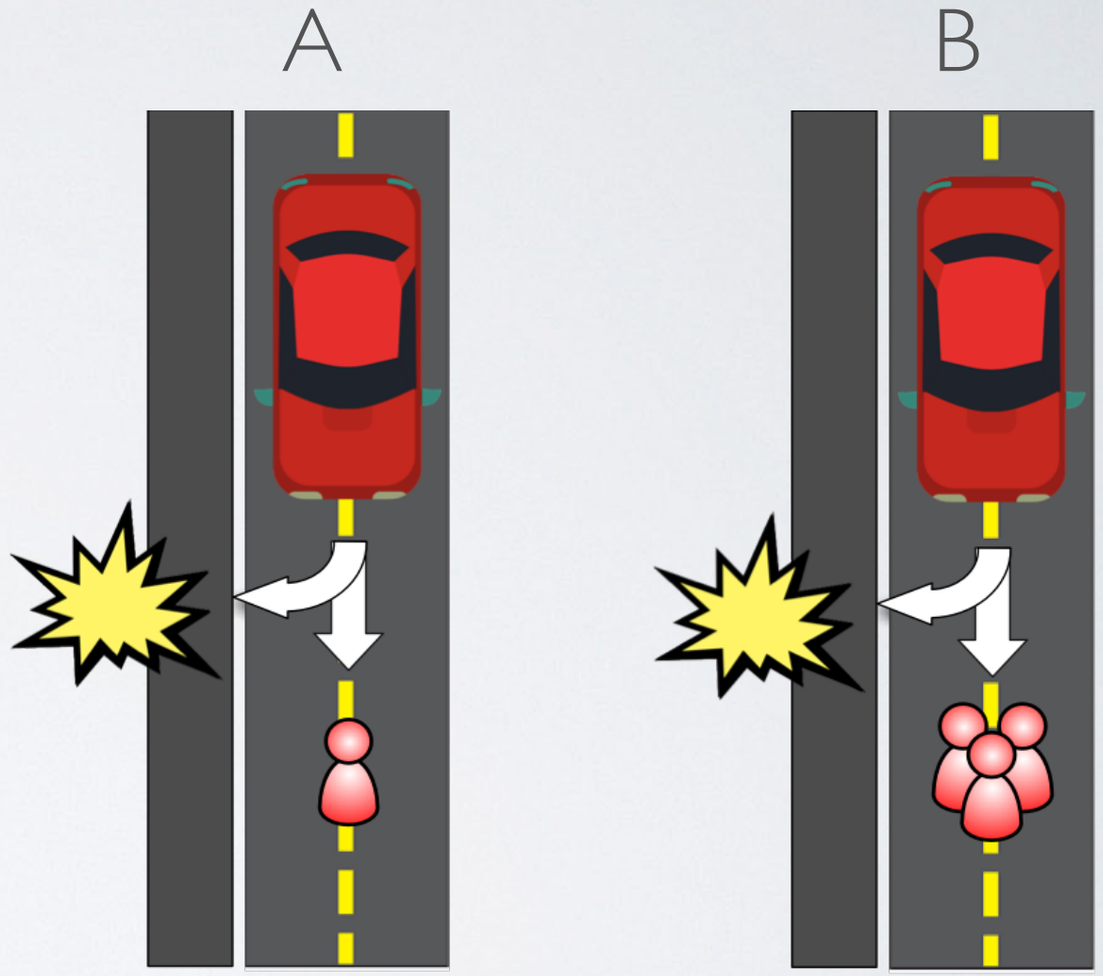
Impartial: The AV kills the passenger

Altruistic: The AV kills the passenger

THE ETHICAL KNOB ON AUTONOMOUS VEHICLES

How would you set the knob?

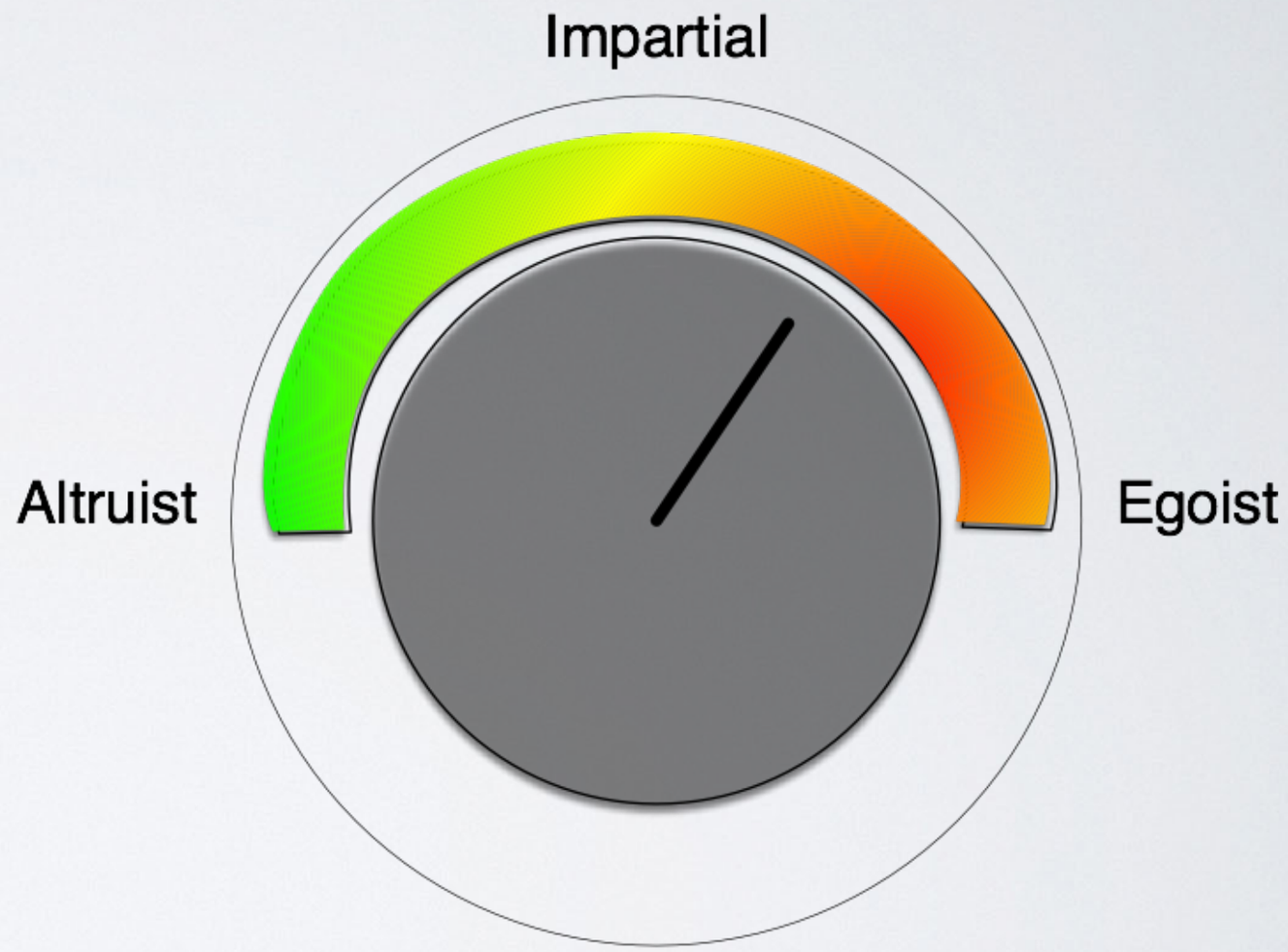
Would it make a different if other people were on board (children, friends, family)?



ETHICAL KNOB 2.0

It enables the passenger to indicate the proportional importance of his life relatively to the importance of the life of others

It can determine the probability that harm would follow from either choice



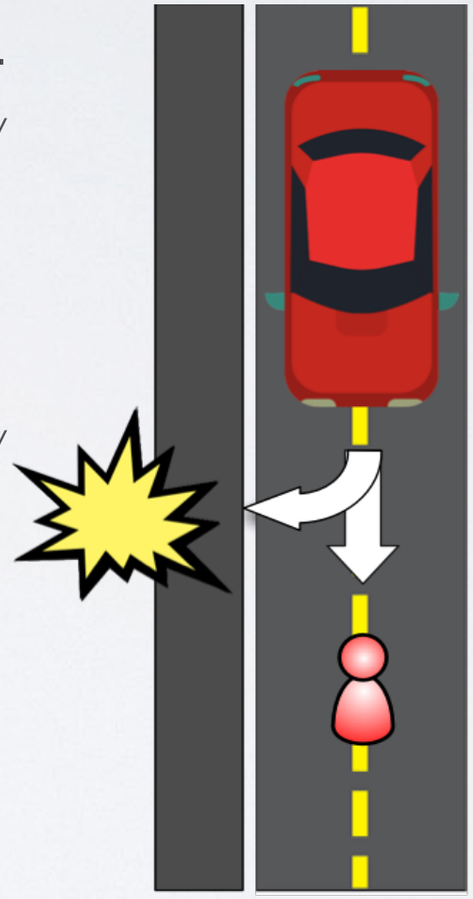
ETHICAL KNOB 2.0

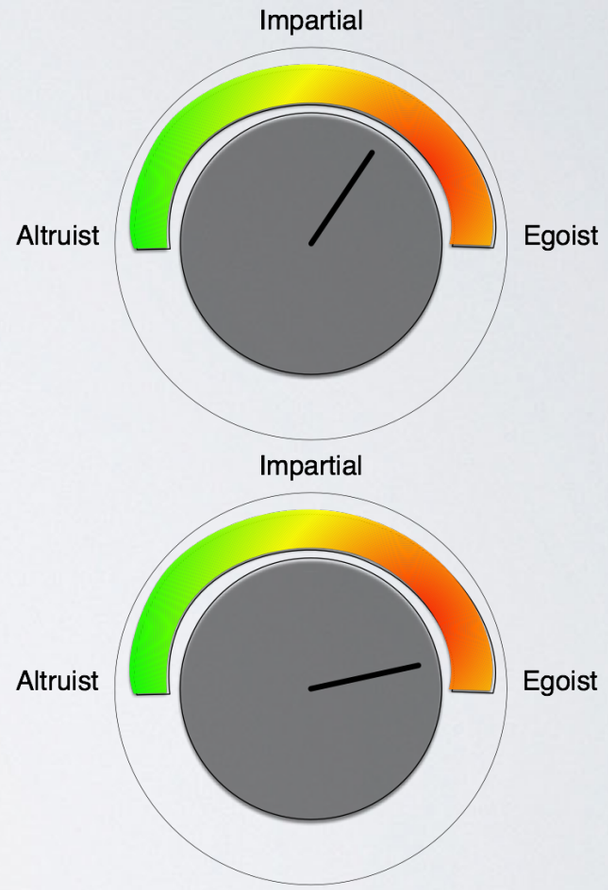
Example 1: The passenger is worth 60% and the pedestrian 40%. Probability or harm (death) to the passenger 10% and probability of harm to the pedestrian 100% . AV puts the passenger at risk.

Example 2: The passenger is worth 95% and the pedestrian 5%. Probability or harm (death) to the passenger 10% and probability of harm to the pedestrian 100% . AV kills the pedestrian.

Did the passenger behave correctly (in setting the knob?)
Is this morally acceptable?
Is this legally acceptable (excused by state of necessity)?

A

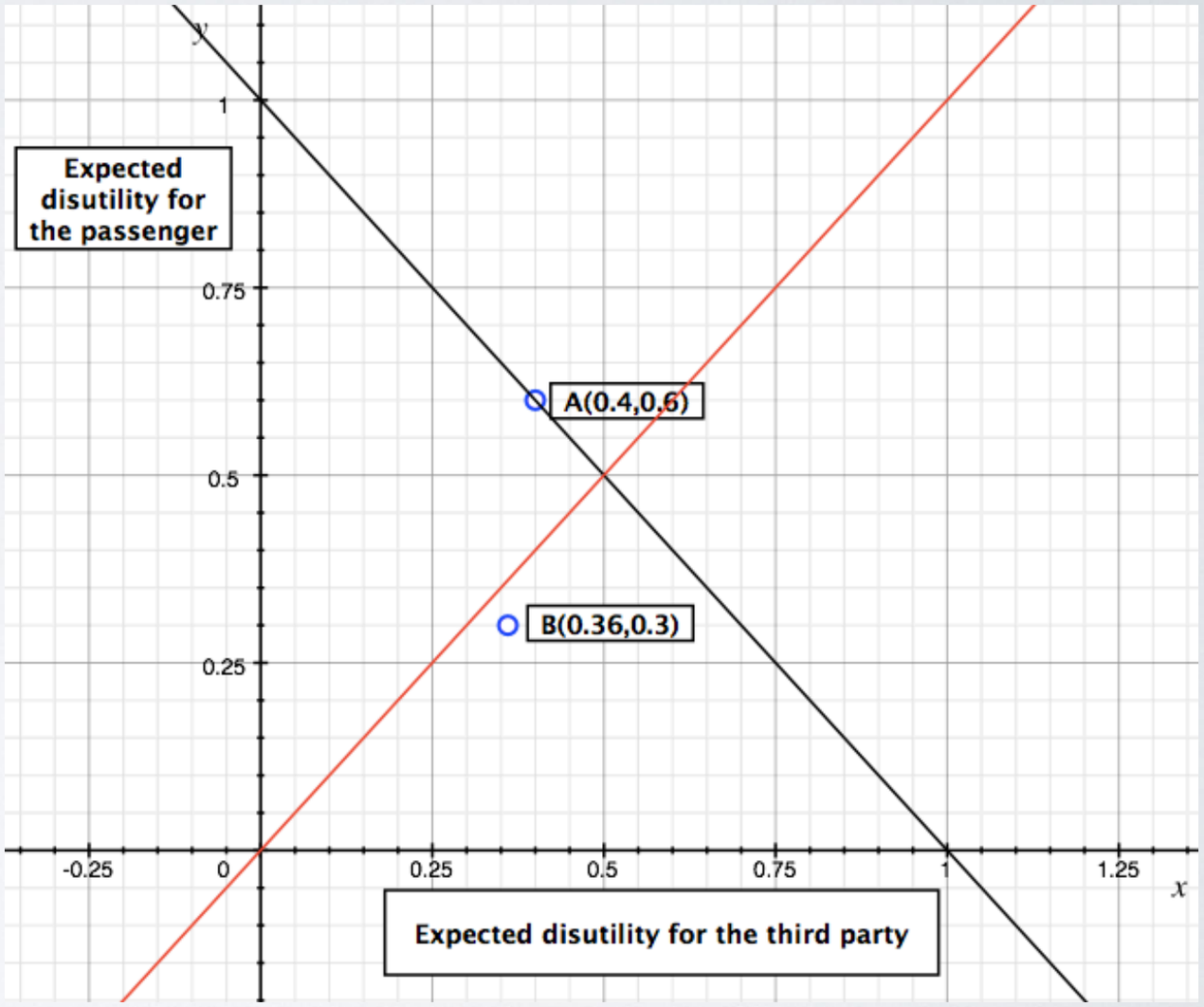




ETHICAL KNOB 2.0

Relative Importance: $y = 1 - x$
knob setting at position $y = 0.6$ indicates that the relative weights of the passenger's and the third party's lives are 0.6 and 0.4

Scenario:
0.5 probability that swerving will cause the passenger's death
0.9 probability that proceeding will cause a third party's death
We can call **disutility** (of swerving) the value dis1=0.6•0.5=0.3. The expected **disutility** of keeping a straight will then be dis2=0.4•0.9= 0.36.
Thus, the AV should choose the course of action that determines the lesser expected disutility, i.e., swerving, because (dis1<dis2, 0.3<0.36)



WHAT IS THE AUTONOMOUS VEHICLE?

Is it an alter ego of the user? Is it allowed to have the biases that are allowed in its users?

Is it supposed to be an impartial mediator between different people?

What systems are supposed to be impartial and which are not? E.g. online trading vs AVs

OPEN QUESTIONS

Should the law allow for the introduction of an ethical knob?

Is the driver who chose an egoistic setting exempted from civil and criminal liabilities?

Always? Under what conditions?

If everyone choses the maximal self-protective mode, could we have a situation similar to the Tragedy of the Commons type scenario?

THE TRAGEDY OF COMMONS

THE CLASSICAL EXAMPLE

It describes a situation in a shared-resource system where individual users acting independently according to their own self-interest behave contrary to the common good of all users by depleting or spoiling that resource through their collective action.

Commons is taken to mean any shared and unregulated resource (e.g. oceans, rivers, fish stocks, etc.)

SUMMARY - WRAP UP

AI is a current trending topic in several domains

AI & Automation are changing the interaction between humans and machines

The concept of Liability is a foundation of the Law system and very well established for humans

Liability needs a re-assessment to comply with AI

The Ethical Knob is a novel proposal to deal with this issue

WHERE DO WE GO? A PUBLIC GOOD GAME

What is it?

It is a standard of experimental economics.

How does it work?

In the basic game, subjects secretly choose how many of their private tokens to put into a public pot. The tokens in this pot are multiplied by a factor (greater than one and less than the number of players, N) and this "public good" payoff is evenly divided among players. Each subject also keeps the tokens they do not contribute.

WORKING PROGRESS: NUMERICAL AGENT-BASED SIMULATION

Aim: assess and verify different possible scenarios (the most "likely behavior" of AVs with the ethical knob implemented)

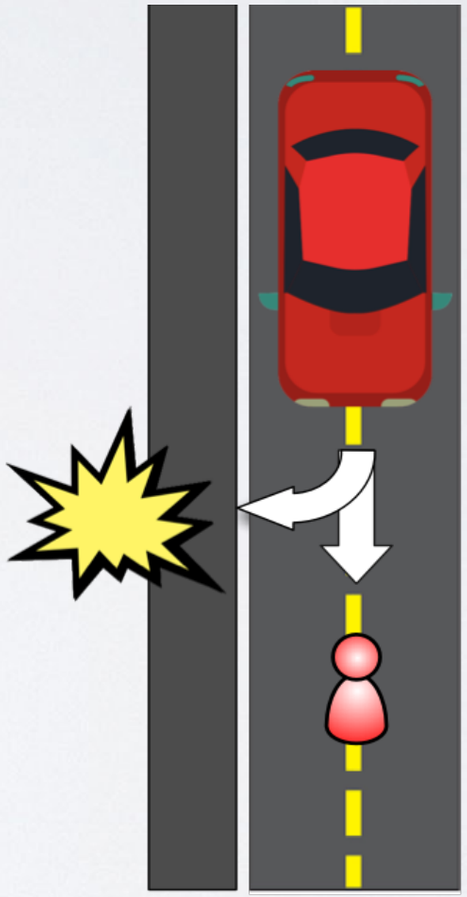
The Public Good: eg. road safety, population safety

Simulation: eg. scenario A (1 passenger on board vs 1 pedestrian)

Knob: it represents the agents degree of altruism or prosocial orientation

Token: level of altruism of each agent

A



NUMERICAL AGENT-BASED SIMULATION

Population: 1000 agents

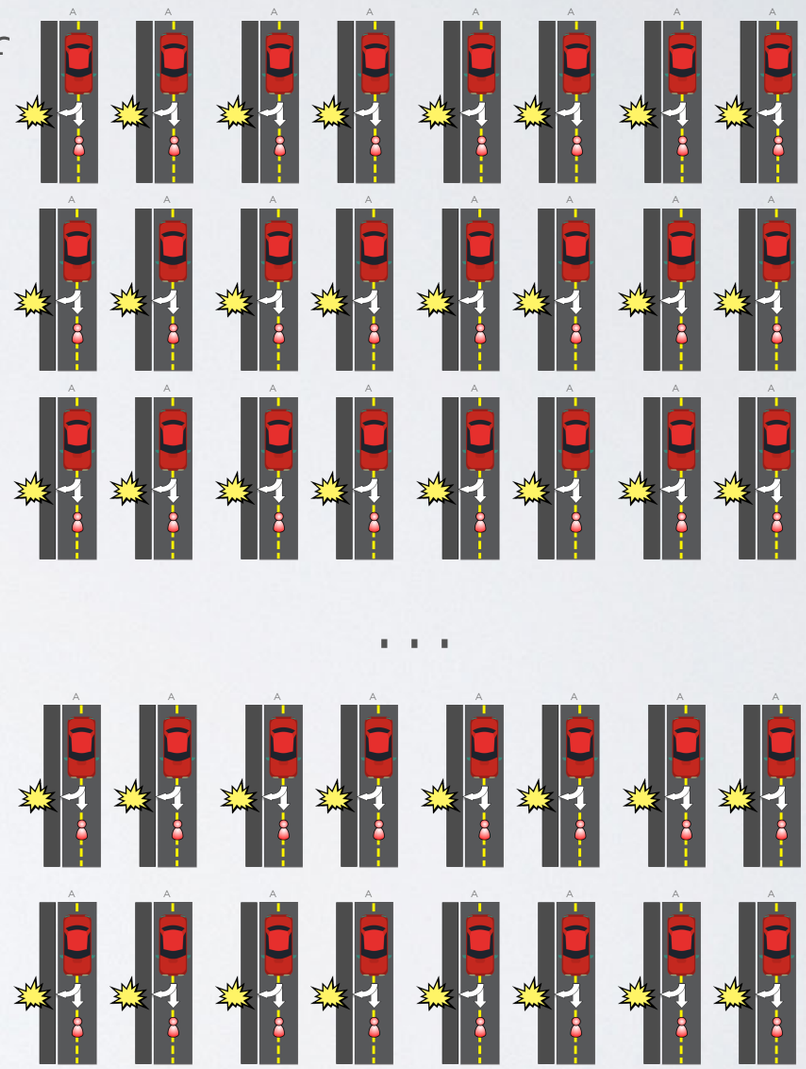
Accident scenarios: it takes into account the outcome of 1000 multiple accident scenarios simultaneously at each step of the simulation.

Iteration: Running the simulation for 1000 iterations (or "Generations"): at each step (or Generation) a 'Population of AVs' behaves according the current set of KNOB positions.

Knob: Each AV has its own KNOB level set to a random initial value, changing on the basis of the experience

Goal: We want to identify the value of the KNOB that MAXIMIZE both

- Individual Payoff
- Collective Payoff



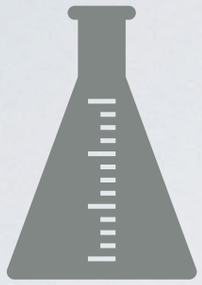
PRELIMINARY RESULTS

Preliminary results

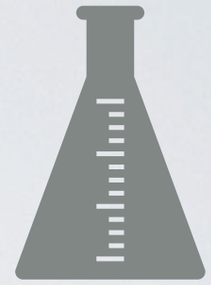
Costs of individual Choices influence the level of altruism
(E.g. insurance premium, sanction, fine, penalty, moral cost, etc.)

- When C is set on a law value the average behavior of the population rapidly converges to ‘Egoism’
- When C is set on a medium value the average behavior of the population converges more slowly
- When C is set on a quite hight value the average behavior of the population converges to an altruistic behavior.

FURTHER EXPERIMENTATIONS



Further experiments



Further Externality: reputational costs and benefits, self-esteem, etc.

e.g. In a labor-market, career concerns make it valuable to be seen by employers as having a strong work ethic, caring about the activity in question, etc.

e.g. In the social sphere, people perceived as generous, public minded, good citizens, etc., are more likely to be chosen as mates, friends, or leaders.