MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

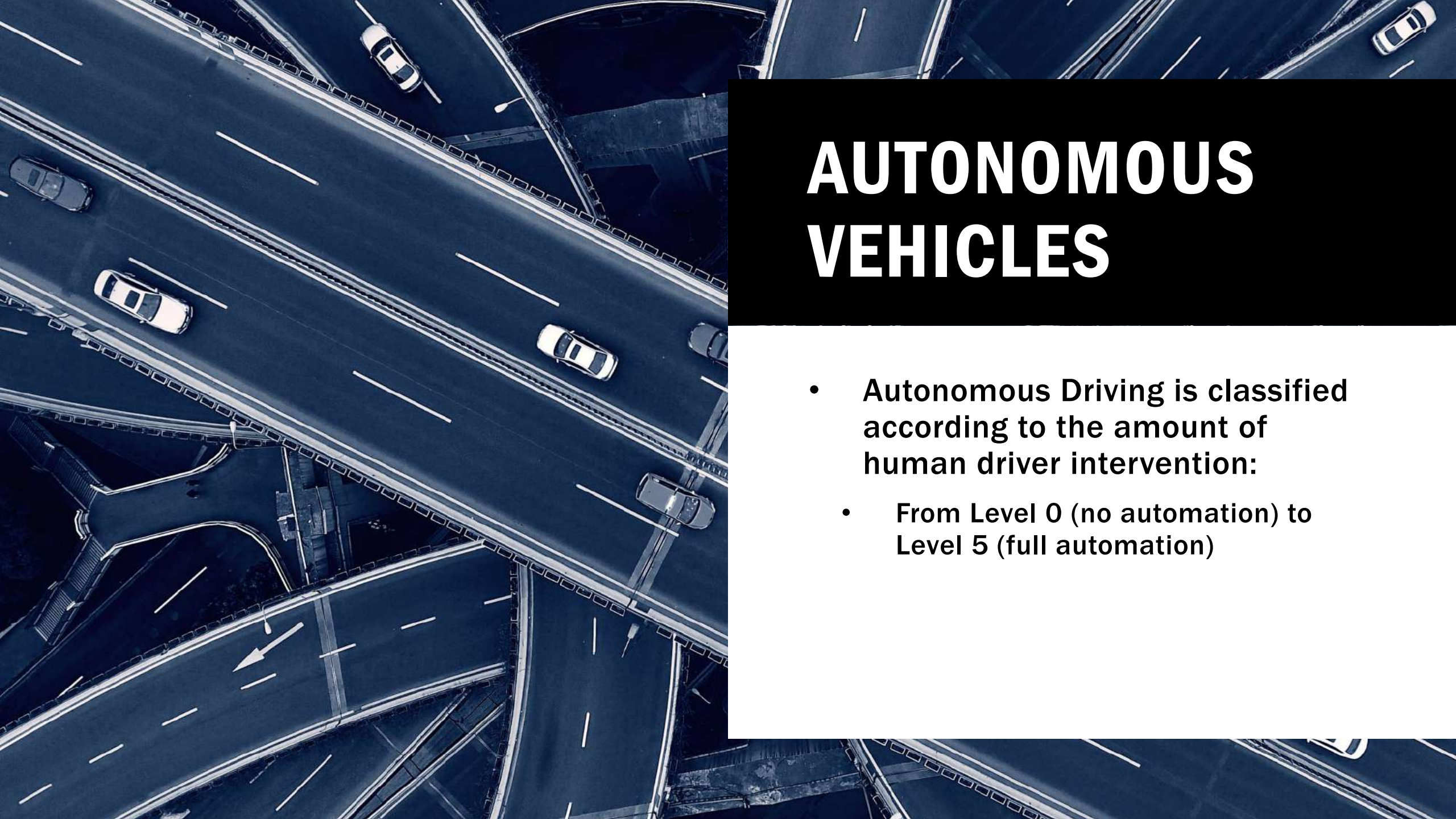
A GENETIC APPROACH TO THE ETHICAL KNOB

Giovanni IACCA (University of Trento)
Francesca LAGIOIA (EUI/CIRSFID)
Andrea LOREGGIA (EUI)
Giovanni SARTOR (EUI/CIRSFID)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

AUTONOMOUS VEHICLES



AUTONOMOUS VEHICLES

- Autonomous Driving is classified according to the amount of human driver intervention:
  - From Level 0 (no automation) to Level 5 (full automation)

AUTOMATION LEVELS OF AUTONOMOUS CARS

LEVEL 0

There are no autonomous features.

LEVEL 1

These cars can handle one task at a time, like automatic braking.

LEVEL 2

These cars would have at least two automated functions.

LEVEL 3

These cars handle "dynamic driving tasks" but might still need intervention.

LEVEL 4

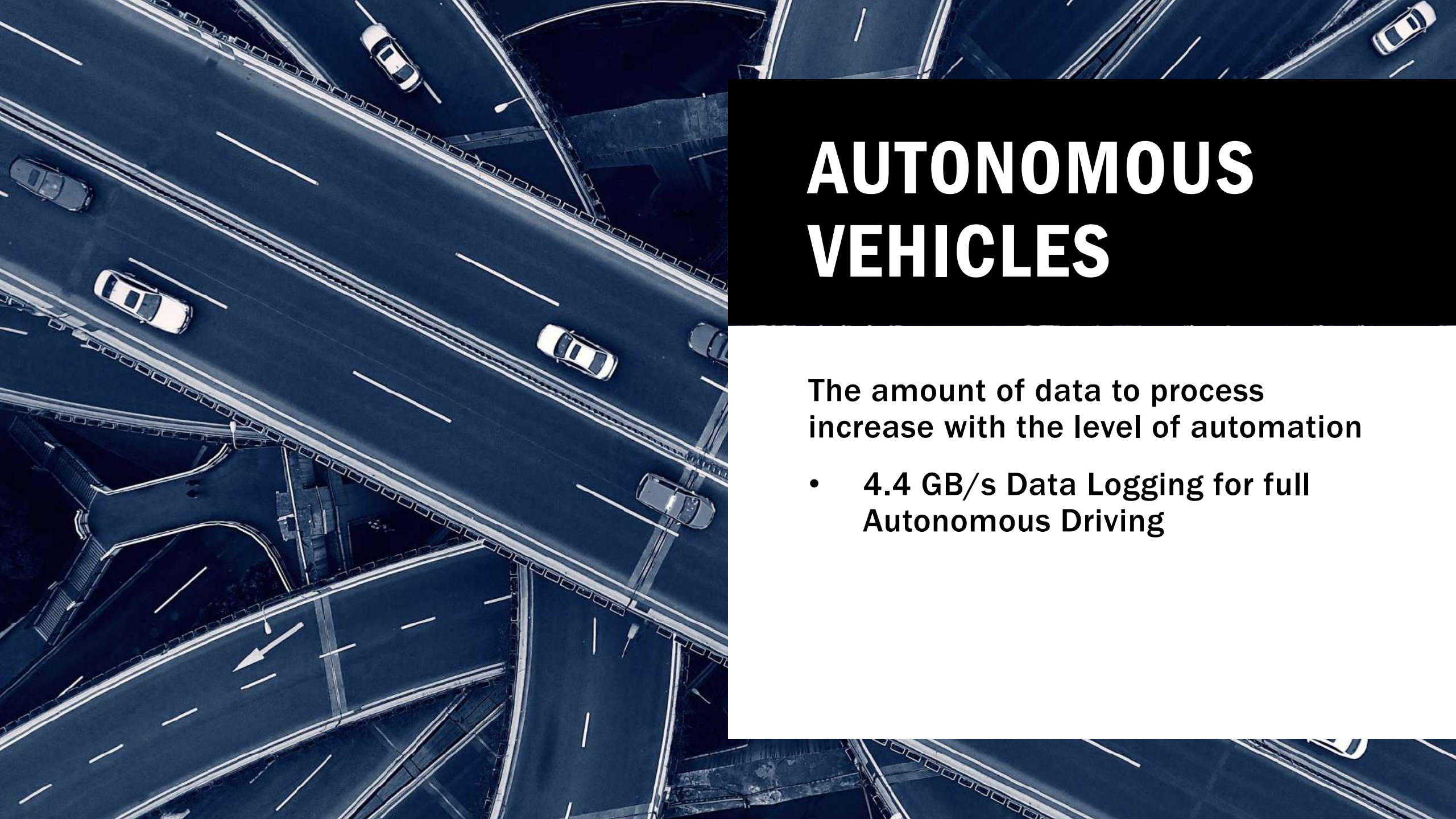
These cars are officially driverless in certain environments.

LEVEL 5

These cars can operate entirely on their own without any driver presence.

SOURCE: SAE International

BUSINESS INSIDER



AUTONOMOUS VEHICLES

The amount of data to process increase with the level of automation

- 4.4 GB/s Data Logging for full Autonomous Driving

CAR AUTOMATION SENSORS & DATA VOLUMES

Sensor type	Quantity	Data generated
Radar	4–6	0.1–15 Mbit/s
LIDAR	1–5	20–100 Mbit/s
Camera	6–12	500–3,500 Mbit/s
Ultrasonic	8–16	<0.01 Mbit/s
Vehicle motion, GNSS, IMU	-	<0.1 Mbit/s

TOTAL ESTIMATED BANDWIDTH

3 Gbit/s (~1.4TB/h) to 40 Gbit/s (~19 TB/h)

AUTONOMOUS VEHICLES CAN POTENTIALLY FAIL

OUTLINE

- Introduction
- Ethical knob, individual preferences and social values
- Genetic Algorithms
- Neural Networks
- Genetic Approach to the Ethical Knob
- Conclusion/Discussion

Chrome Archivio Modifica Visualizza Cronologia Preferiti Persone Scheda Finestra Guida

81% Ven 12:48

Moral Machine

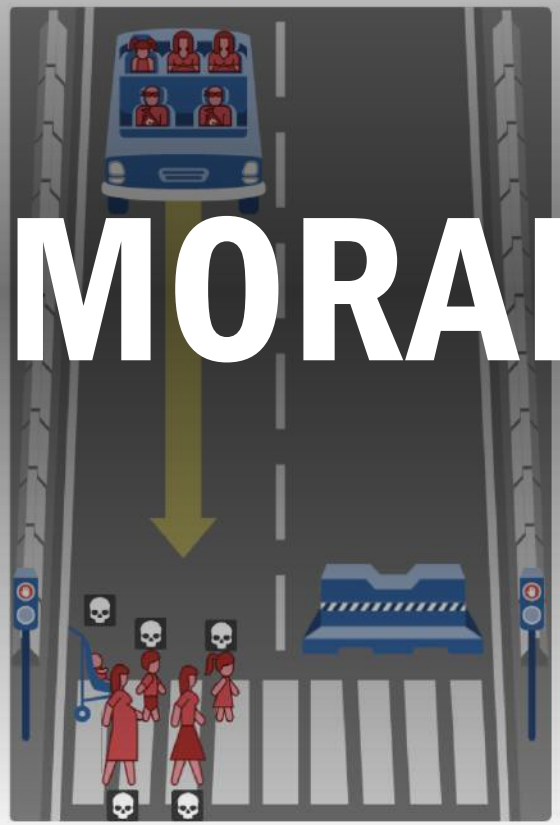
moralmachine.net

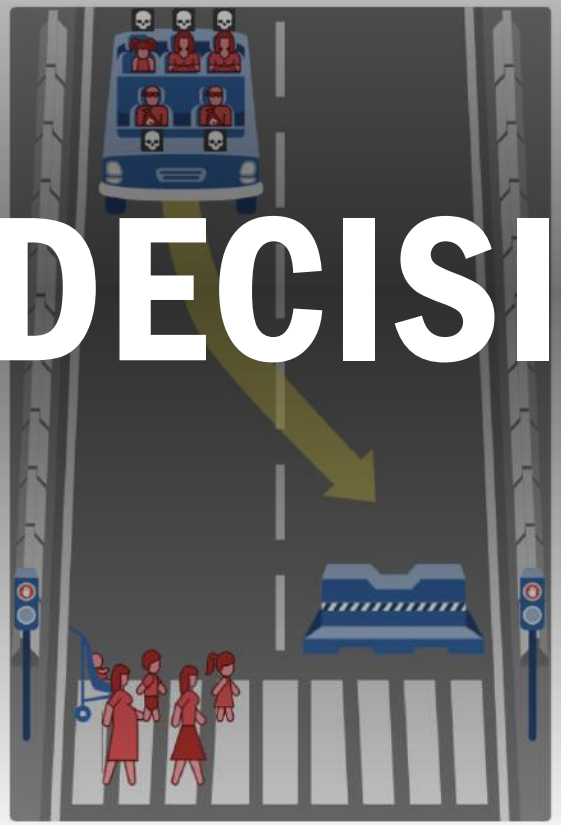
MORAL MACHINE

Home Judge Classic Design Browse About Feedback En

Help

Share Link 0 Likes Random





Show Description

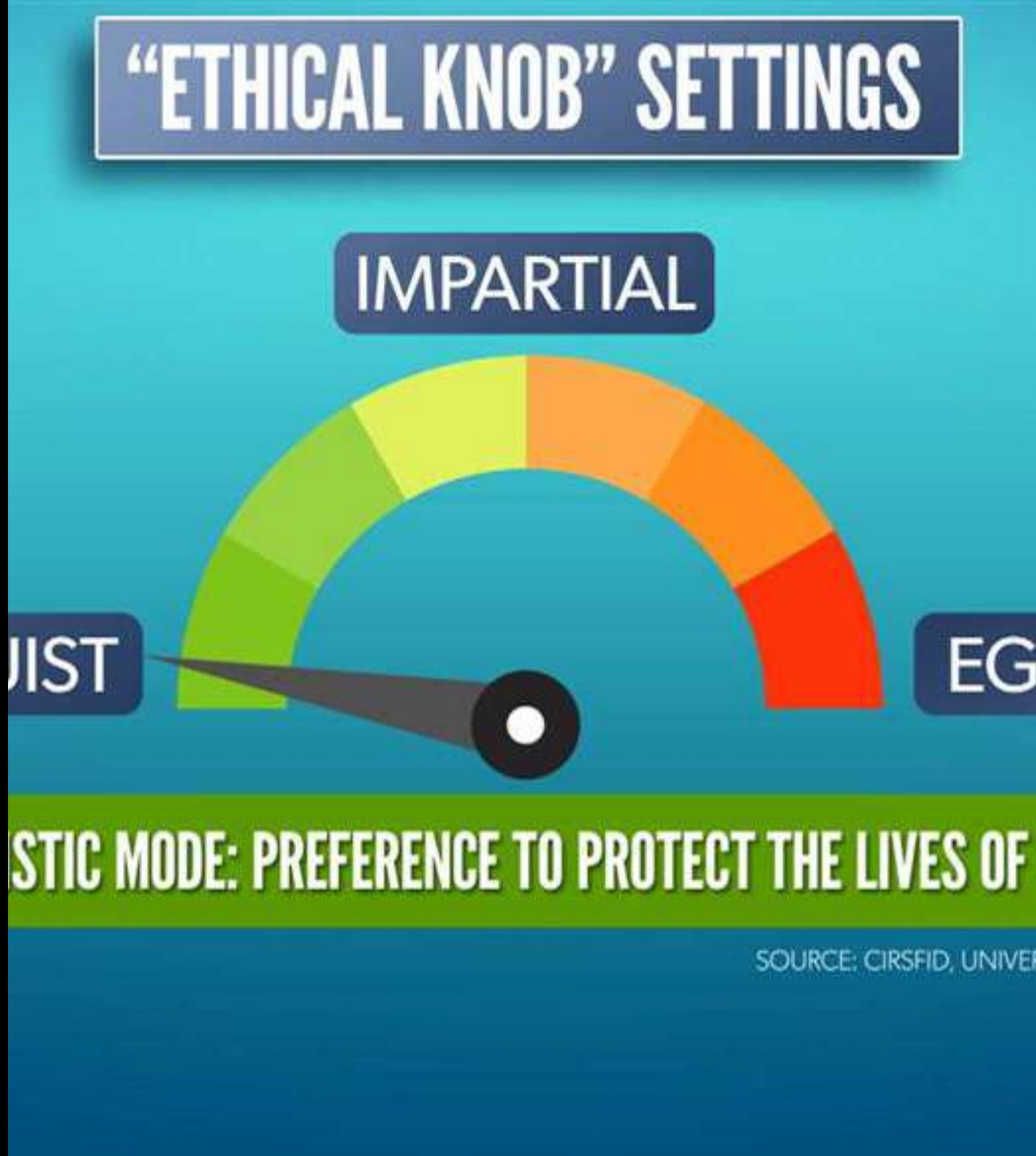
Show Description

Sponsored

THE MORAL DECISIONS

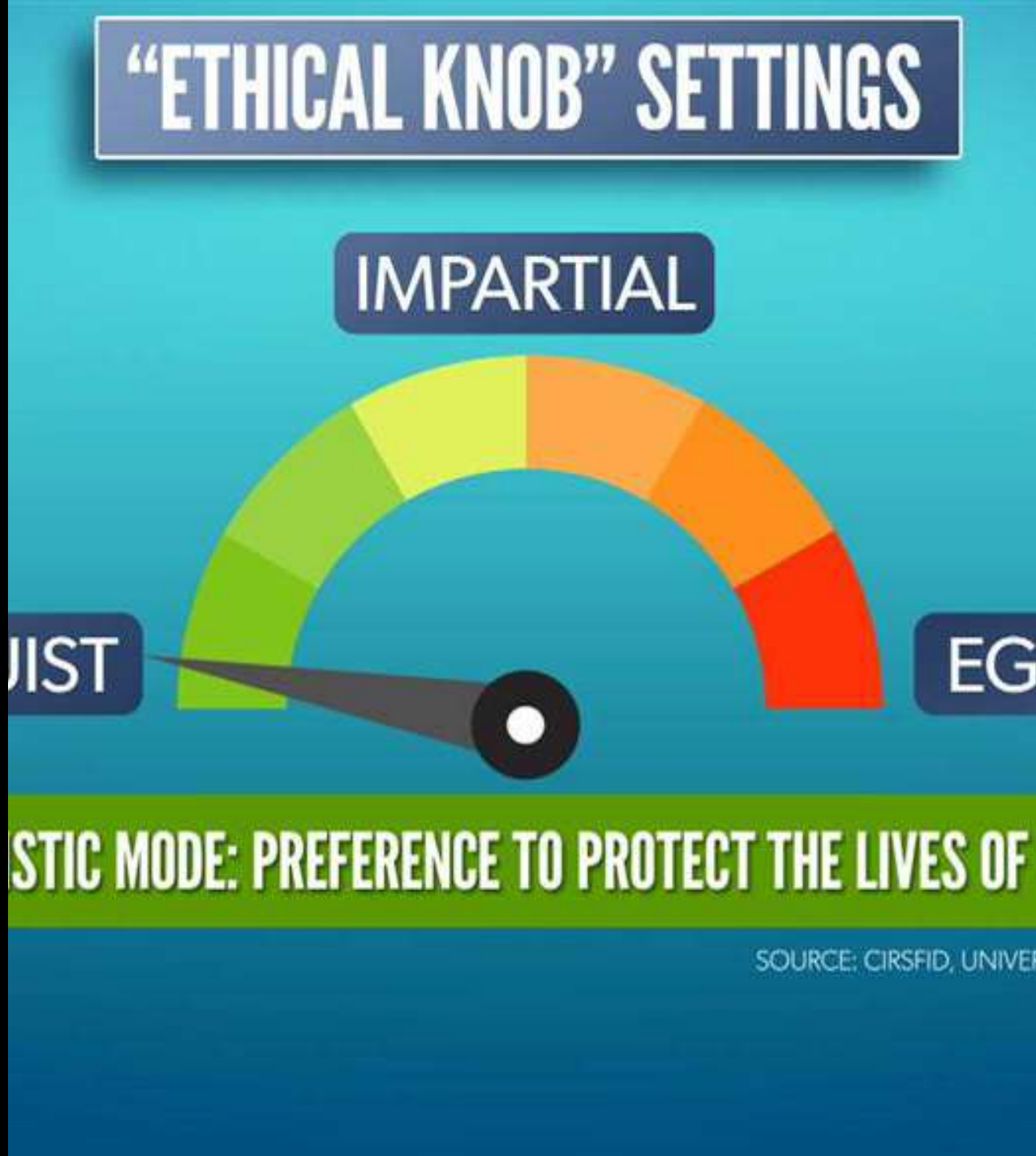
THE ORIG. PROPOSAL

- The knob expresses directly the ethical attitude of the AV passengers
- The value passengers attribute to their life relative to the value of the lives of third parties



THE NEW PROPOSAL

- The position of the knob <u>no longer</u> indicates the passengers' moral attitude
- It indicates the AV's assessment of the relative importance of the lives of passenger(s) and third parties



HOW TO DO THAT?

- Combination of AI techniques:
  - Neural networks to compute the right action to take based on the given scenario
  - Genetic Algorithm to find an (almost) optimal configuration of neural networks

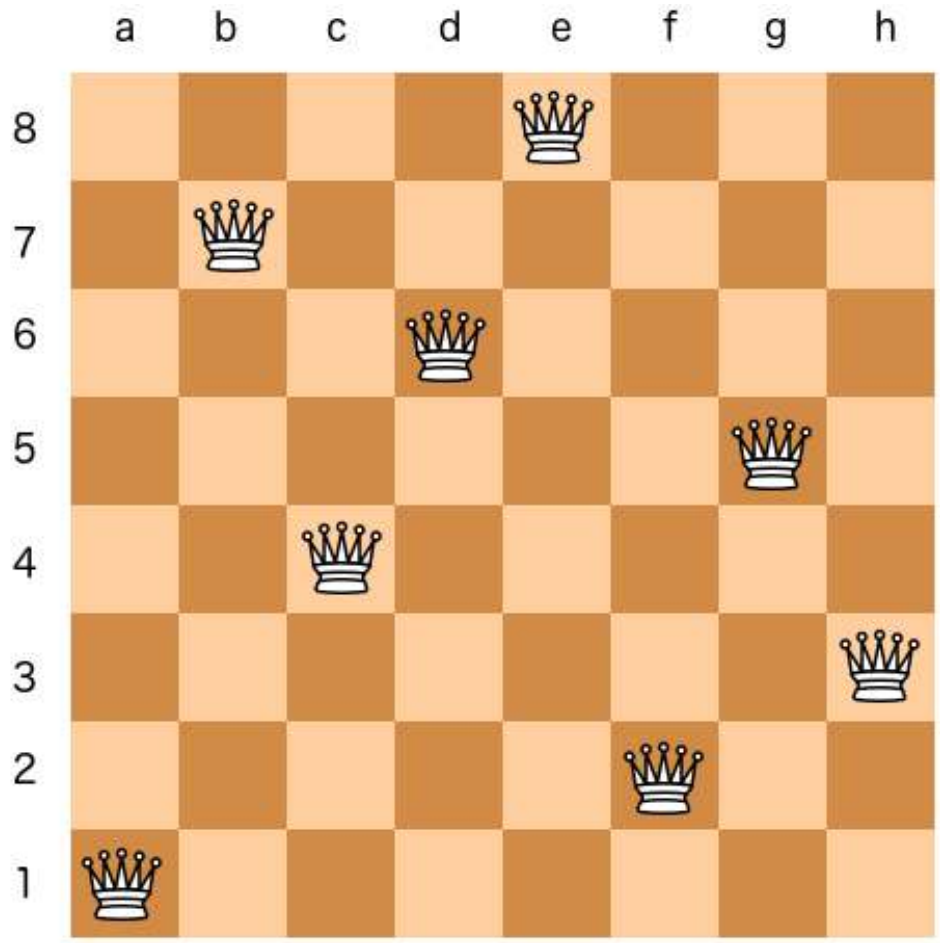
GENETIC ALGORITHMS

- Inspired by Charles Darwin's theory of natural evolution:
  - the fittest individuals are selected for reproduction in order to produce offspring of the next generation
- Heuristic Search in the solution space
- Mostly used in optimization tasks

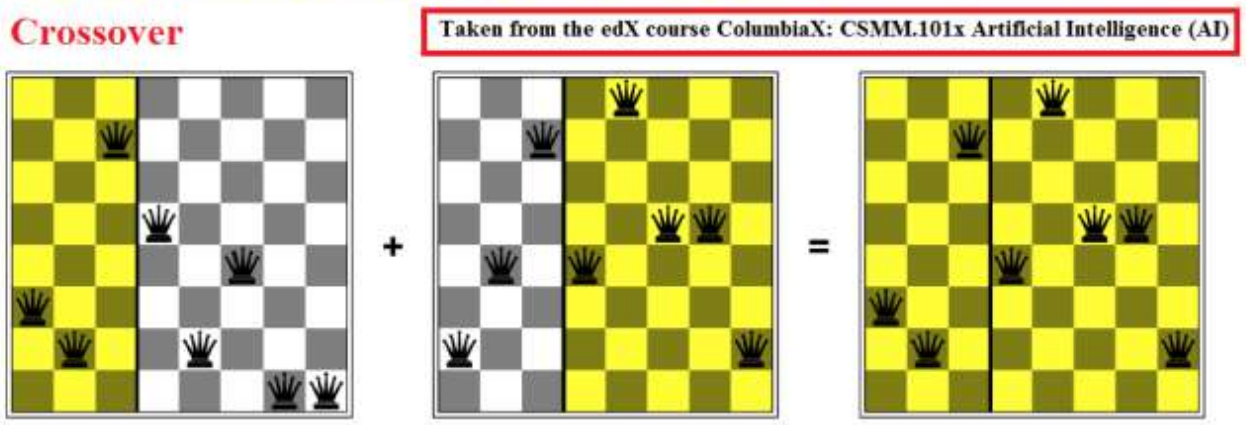
SIMPLE EXAMPLE

N-queens puzzle: place n chess queens on an $n \times n$ chessboard so that no two queens threaten each other

SIMPLE EXAMPLE

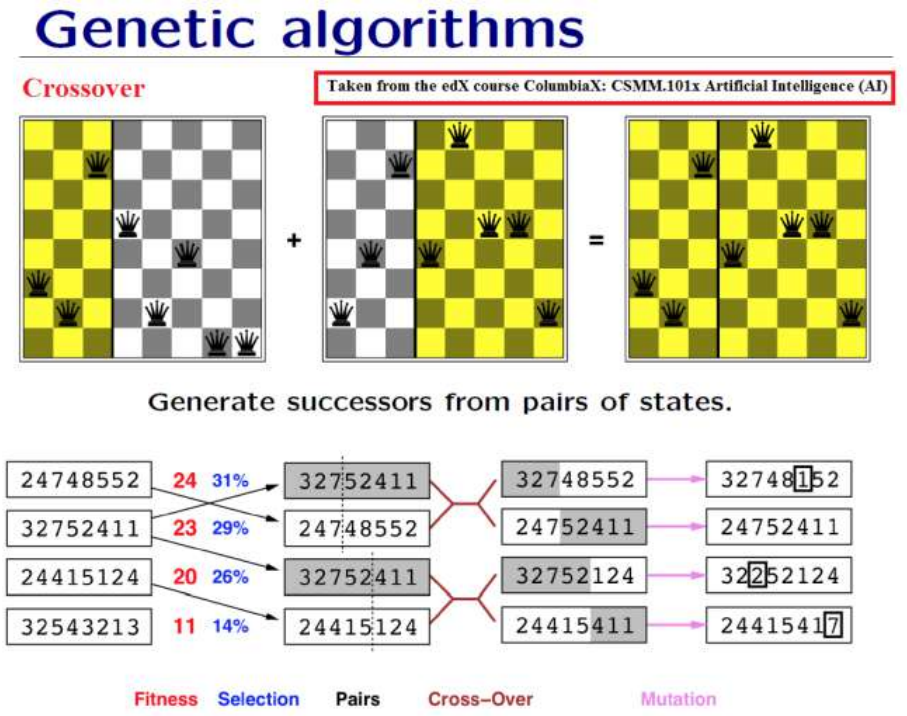


HOW IT WORKS

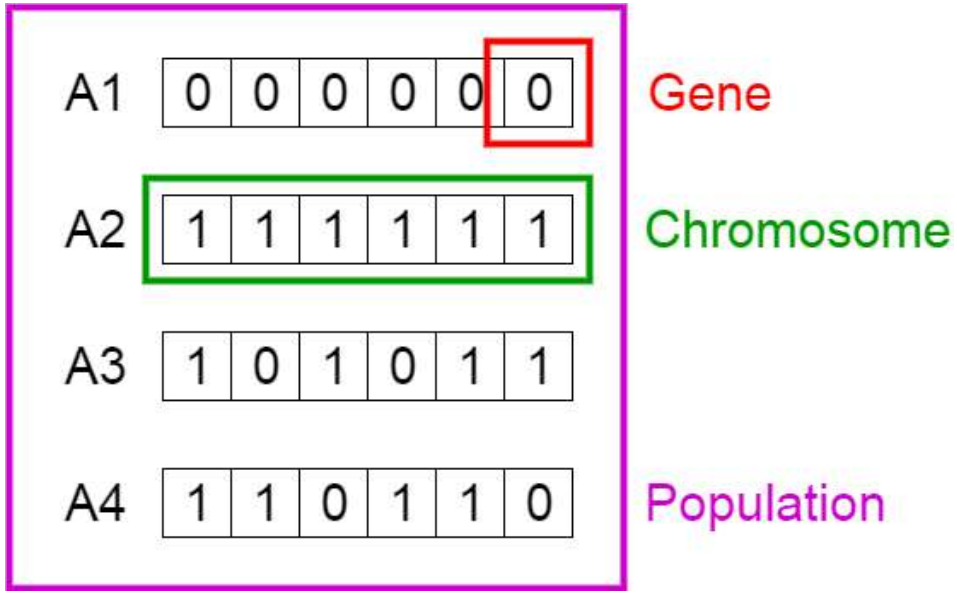


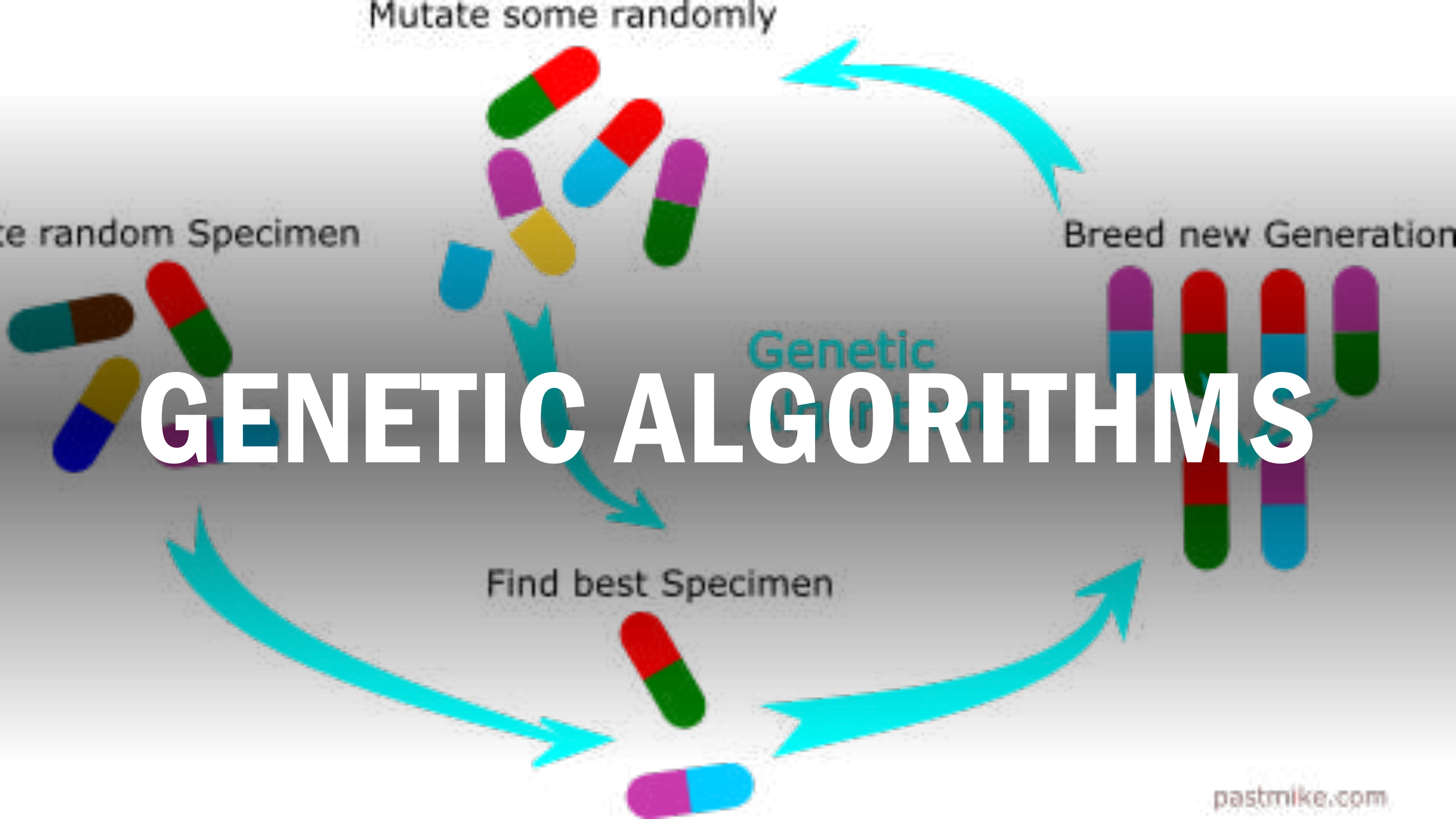
- Individuals corresponds to solutions of the problem
- Initially, solutions are generated at random
- Each individual is evaluated
- The best are selected and used to combined to produce the new generation

HOW IT WORKS


Genetic algorithms
Crossover
Taken from the edX course ColumbiaX: CSMM.101x Artificial Intelligence (AI)
+
=
Generate successors from pairs of states.
24748552
32752411
24415124
32543213
24 31%
23 29%
20 26%
11 14%
32752411
24748552
32752411
24415124
32748552
24752411
32752124
24415411
32748152
24752411
32252124
24415417
Fitness
Selection
Pairs
Cross–Over
Mutation

GENETIC ALGORITHMS

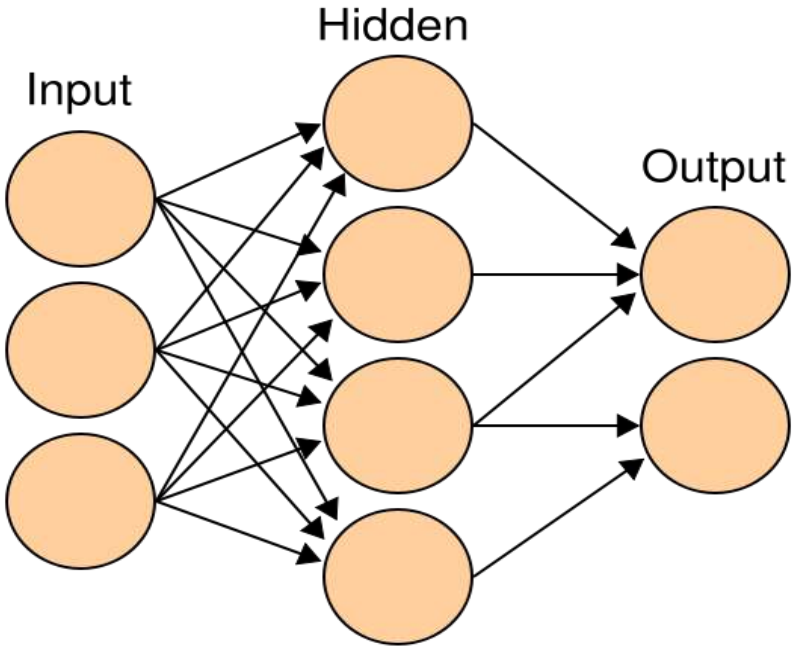




GENETIC ALGORITHMS

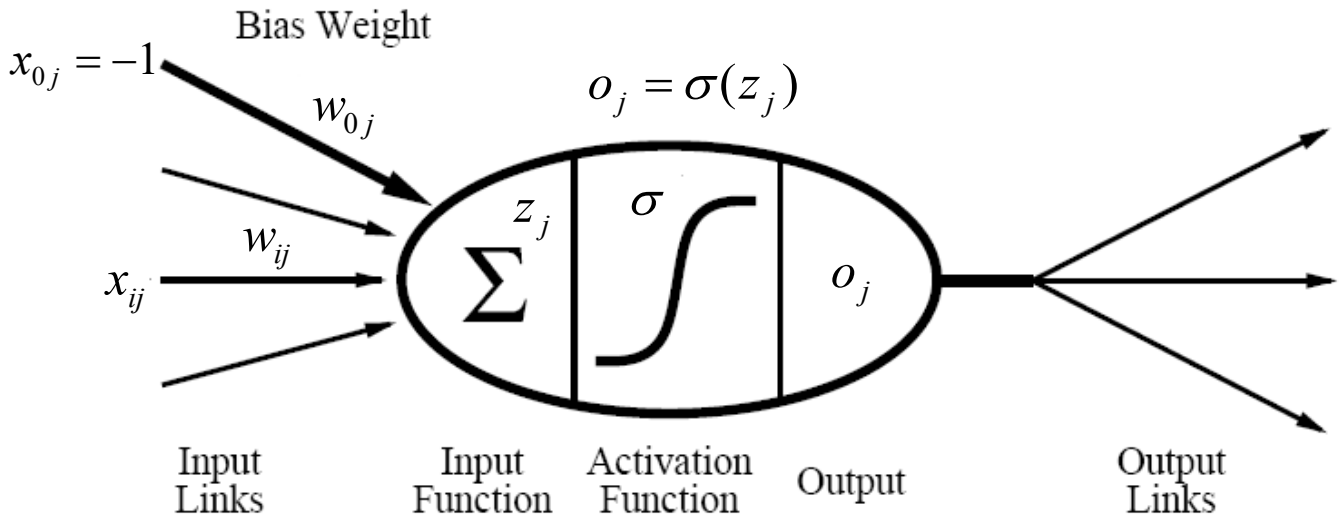
ARTIFICIAL NEURAL NETWORK

- Inspired by natural neural network
  - Classification / Regression
- Adaptive Model, the internal state is adjusted during the training phase



ARTIFICIAL NEURAL NETWORK: ORIGINS

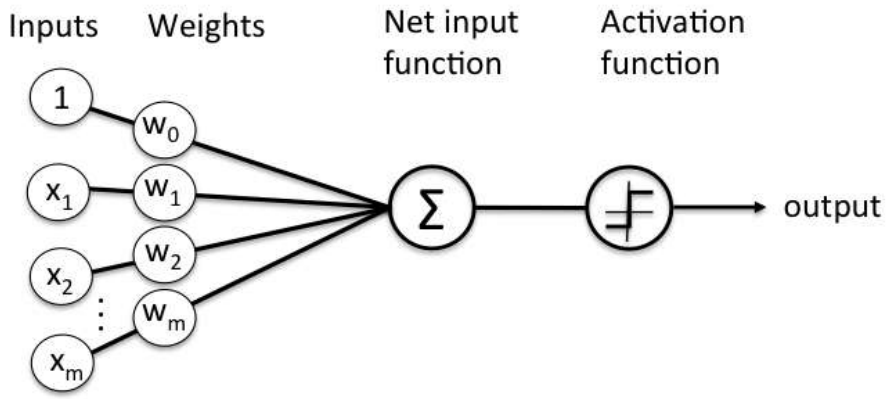
Formal model of a neuron:



McCulloch, W. S. and Pitts, W. (**1943**). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–137.

ARTIFICIAL NEURAL NETWORK: ORIGINS

- Input values are weighted based on "importance"
- Weighted input are summed up
- The sum is transformed using an activation function

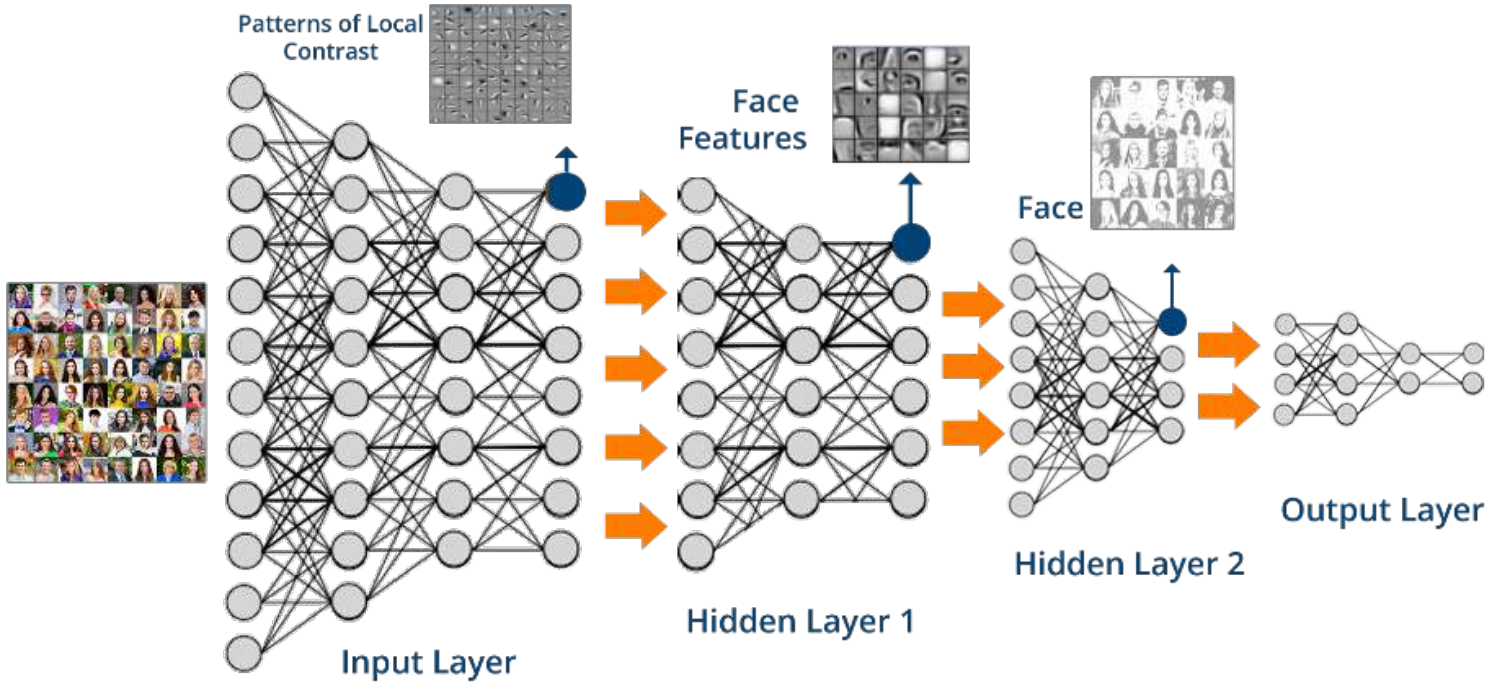


Schematic of Rosenblatt's perceptron.

Rosenblatt, Frank. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms.* No. VG-1196-G-8. Cornell Aeronautical Lab Inc Buffalo NY, **1961**.

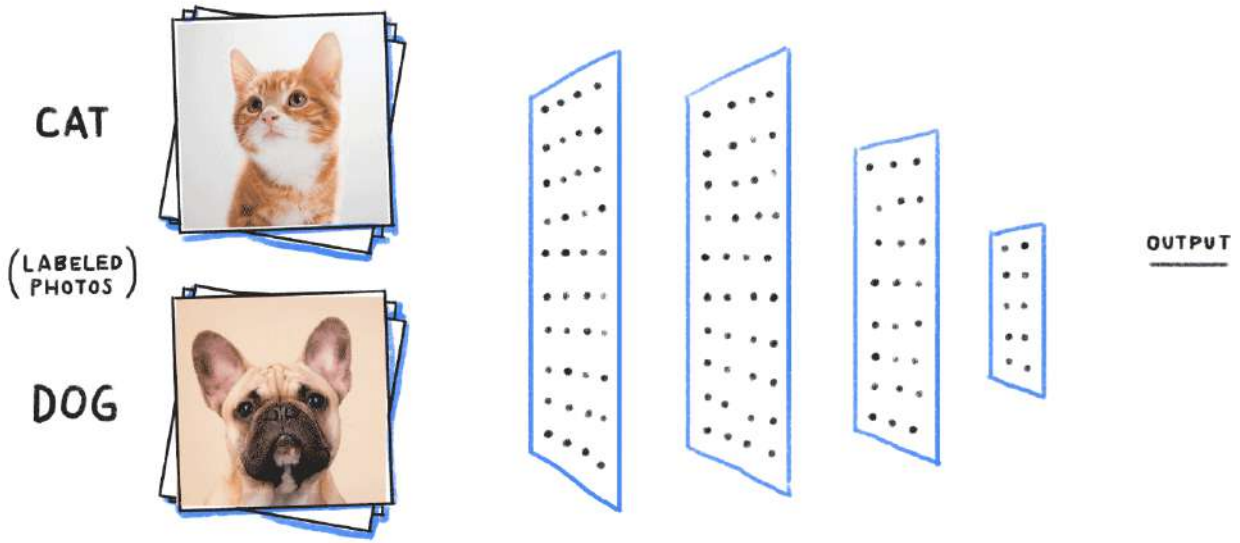
ARTIFICIAL NEURAL NETWORK

Neural network are made of several layers of **perceptron,** the main idea is to mimic the cerebral cortex



ARTIFICIAL NEURAL NETWORK

Neural network are made of several layers of **perceptron,** the main idea is to mimic the cerebral cortex



ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

- The network is trained based on sample pairs (*x*,y) (training set).
- The training set is used several times (each time is called an epoch), weights are adjusted in order to decrease the error.
- Gradient descent is efficient, but it can stuck in a local minimum.
- Training is in general NP-Complete.

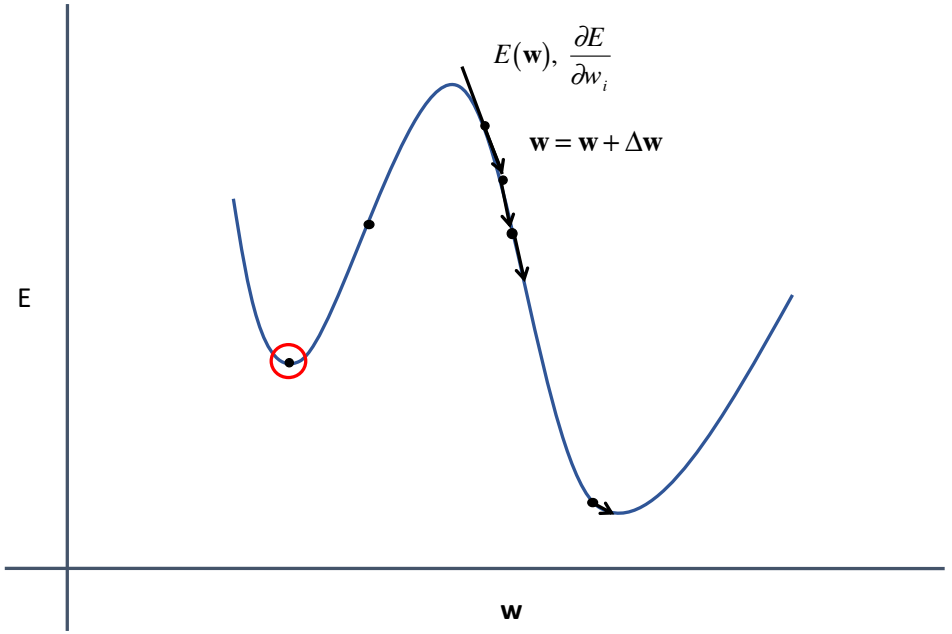
```
Initialize weights at random
repeat
for each example in the training set
    compute example's output
    compute quadratic error
    for i = levels_# down to 1
        compute update for weights
        at level i
    end
    update all weights
end
until (all examples correctly classified
or max iterations reached)
```

Werbos (**1974**). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. *Ph.D. Thesis, Harvard University.*

Rumelhart, Hintont, Williams (**1986**). Learning representations by back-propagating errors. *Nature*

GRADIENT DESCENT

The idea is computing the partial derivative of the error function in order to reduce the loss.

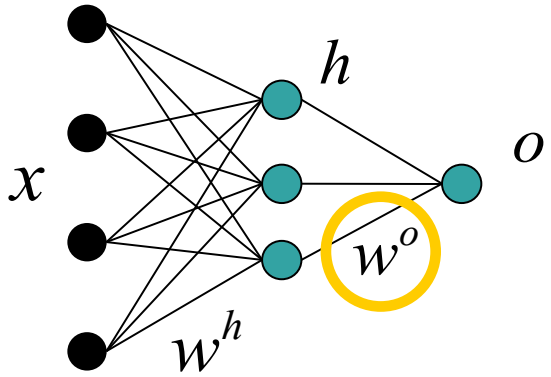


ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

Definitions:

$$z_j^h = \sum_{i=0}^{n} w_{ij}^h x_i \qquad h_j = \sigma(z_j^h)$$

$$z^o = \sum_{j=0}^{m} w_j^o h_j \qquad o = \sigma(z^o)$$



Activation Function:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$

$$\boldsymbol{x} \in \mathbb{R}^{n,1} \quad \boldsymbol{w^h} \in \mathbb{R}^{n,m}$$

$$\boldsymbol{h} \in \mathbb{R}^{m,1} \quad \boldsymbol{w^o} \in \mathbb{R}^{1,m}$$

Error function:

$$E = \frac{1}{2}(y-o)^2 \qquad w_i = w_i - \alpha\frac{\partial E}{\partial w_i} = w_i + \Delta w_i$$

$$\frac{\partial E}{\partial w_j^o} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} \cdot \frac{\partial z^o}{\partial w_j}$$

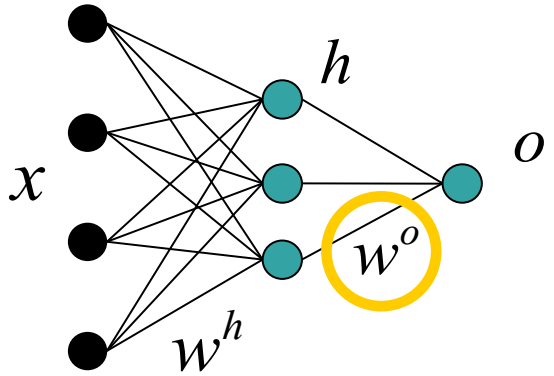
ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

$$\frac{\partial E}{\partial w_j^o} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} \cdot \frac{\partial z^o}{\partial w_j}$$

$$z^o = \sum_{j=0}^{m} w_j^o h_j$$

$$o = \sigma(z^o)$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$



$$\frac{\partial E}{\partial o} = \frac{\partial}{\partial o}\left[\frac{1}{2}(y - o)^2\right] = -(y - o)$$

$$\frac{\partial o}{\partial z^o} = o \cdot (1 - o)$$

$$\frac{\partial z^o}{\partial w_j^o} = h_j$$

$$\frac{\partial E}{\partial w_j^o} = -(y - o) \cdot o \cdot (1 - o) \cdot h_j = -\boldsymbol{\delta^o} h_j$$

Weight update:

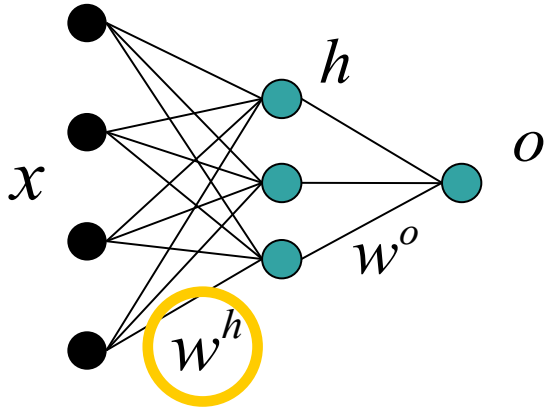
$$\Delta w_j^o = \alpha \delta^o h_j$$

ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

$$\frac{\partial E}{\partial w_{ij}^h} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} \cdot \frac{\partial z^o}{\partial h_j} \cdot \frac{\partial h_j}{\partial z_j^h} \cdot \frac{\partial z_j^h}{\partial w_{ij}^h}$$

$$z_j^h = \sum_{i=0}^{n} w_{ij}^h x_i$$

$$h_j = \sigma(z_j^h)$$



$$\frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} = -\delta^o$$

$$\frac{\partial z^o}{\partial h_j} = w_j^o$$

$$\frac{\partial h_j}{\partial z_j^h} = h_j \cdot (1 - h_j)$$

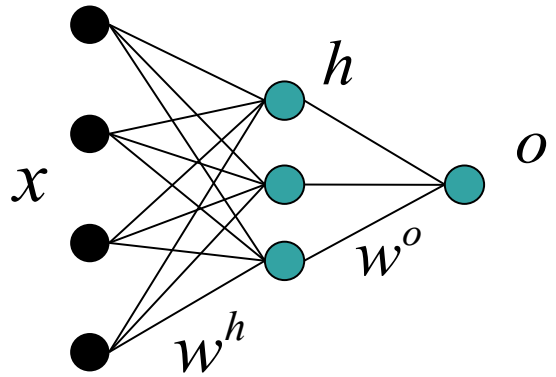
$$\frac{\partial z_j^h}{\partial w_{ij}^h} = x_i$$

$$\frac{\partial E}{\partial w_{ij}^h} = -\boldsymbol{\delta^o} \cdot w_j^o \cdot h_j \cdot (1 - h_j) \cdot x_i = -\boldsymbol{\delta_j^h} x_i$$

Weight update:

$$\Delta w_{ij}^h = \alpha \delta_j^h x_i$$

ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION



Weights update:

$$\Delta w_j^o = \alpha \delta^o h_j$$

$$\Delta w_{ij}^h = \alpha \delta_j^h x_i$$

with

$$\delta^o = (y - o) \cdot o \cdot (1 - o)$$

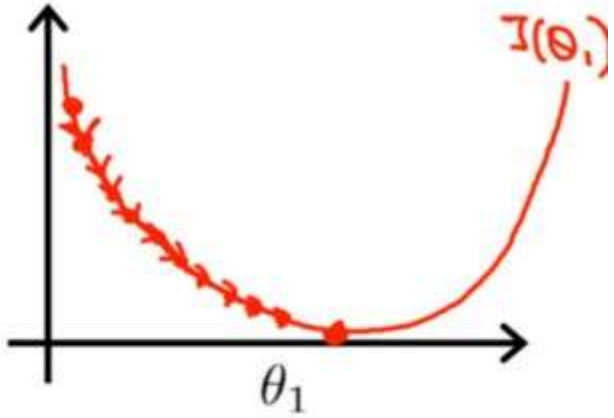
$$\delta_j^h = \delta^o \cdot w_j^o \cdot h_j \cdot (1 - h_j)$$

Learning rate

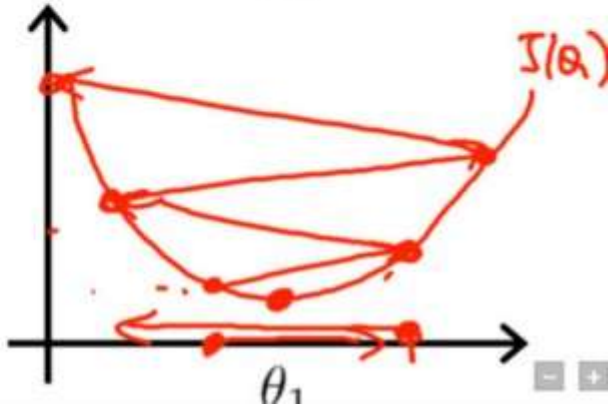
GRADIENT DESCENT

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

If α is too small, gradient descent can be slow.



If α is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.

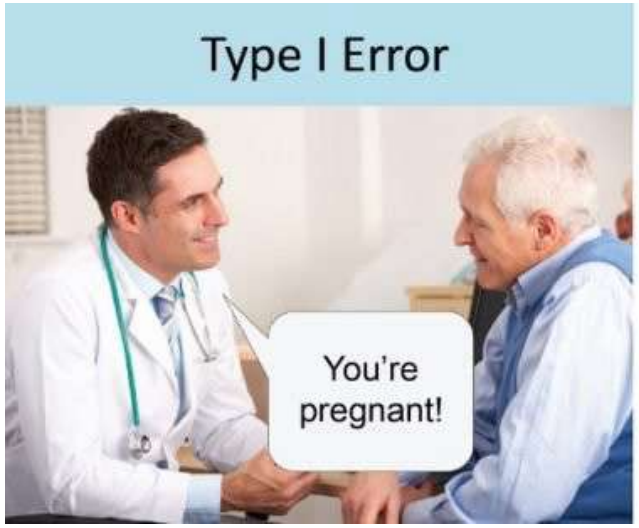


Andrew Ng's course on Coursera

PERFORMANCE

Confusion matrix shows how many true/false positives and true/false negatives

	Predicted	
	Positive	Negative
Actual True	**TP**	**FN**
Actual False	**FP**	**TN**



PERFORMANCE

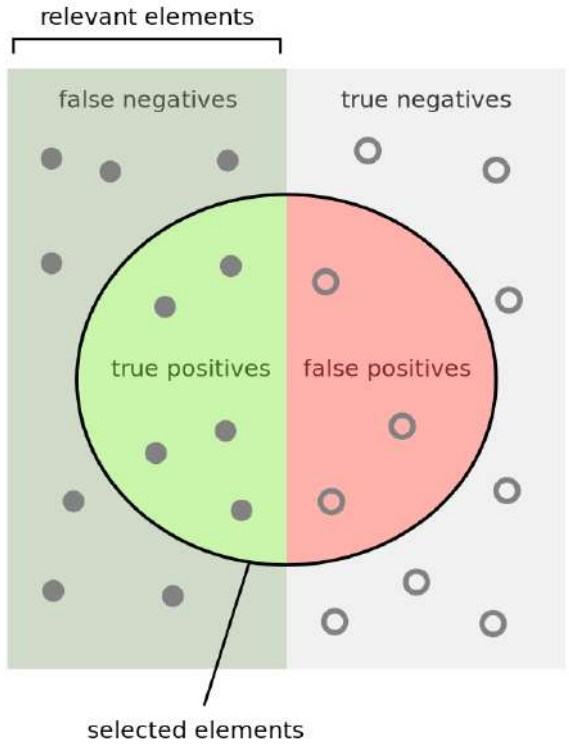
- Accuracy

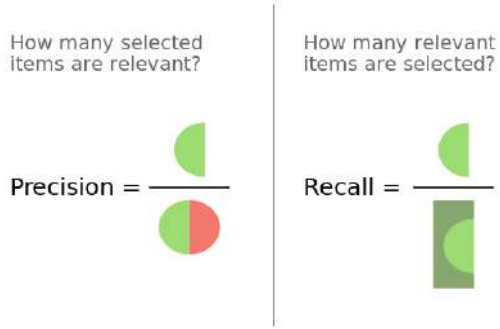
$$\frac{(TP + TN)}{n}$$

- Precision and Recall
- F1 Score o F-score, weighted sum of Precision and Recall

$$2 \times \frac{P \times R}{P + R}$$

- K of Cohen





GENETIC APPROACH TO EK: WHY?

We cannot use gradient descent:

- What data?
- How to train?
- Which values for hyper-parameters?
- Where? In which scenarios?

HOW TO DO THAT?

- Combination of AI techniques:
  - Neural networks to compute the right action to take based on the given scenario
  - Genetic Algorithm to find an (almost) optimal configuration of neural networks

GENETIC APPROACH TO EK

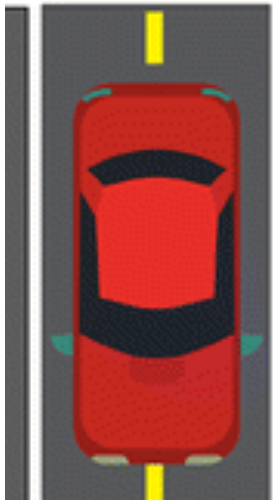
Algorithm 1 Evolutionary algorithm of the Ethical Knob

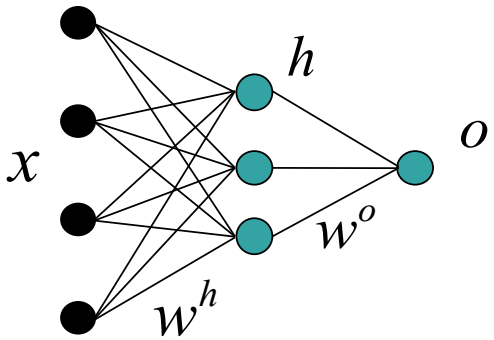
procedure EK(n) ▷ Input: n number of individuals in the population
    Initialize a random population P of n individuals
    for Every generation **do**
        $EvaluateFitness(P)$
        $parents = SelectParents(P)$
        $offsprings = crossOver(parents)$
        $P = mutation(offsprings)$
    end for
    return P
end procedure

SIMULATION

An individual in the simulation corresponds to an AV

SIMULATION: POPULATION



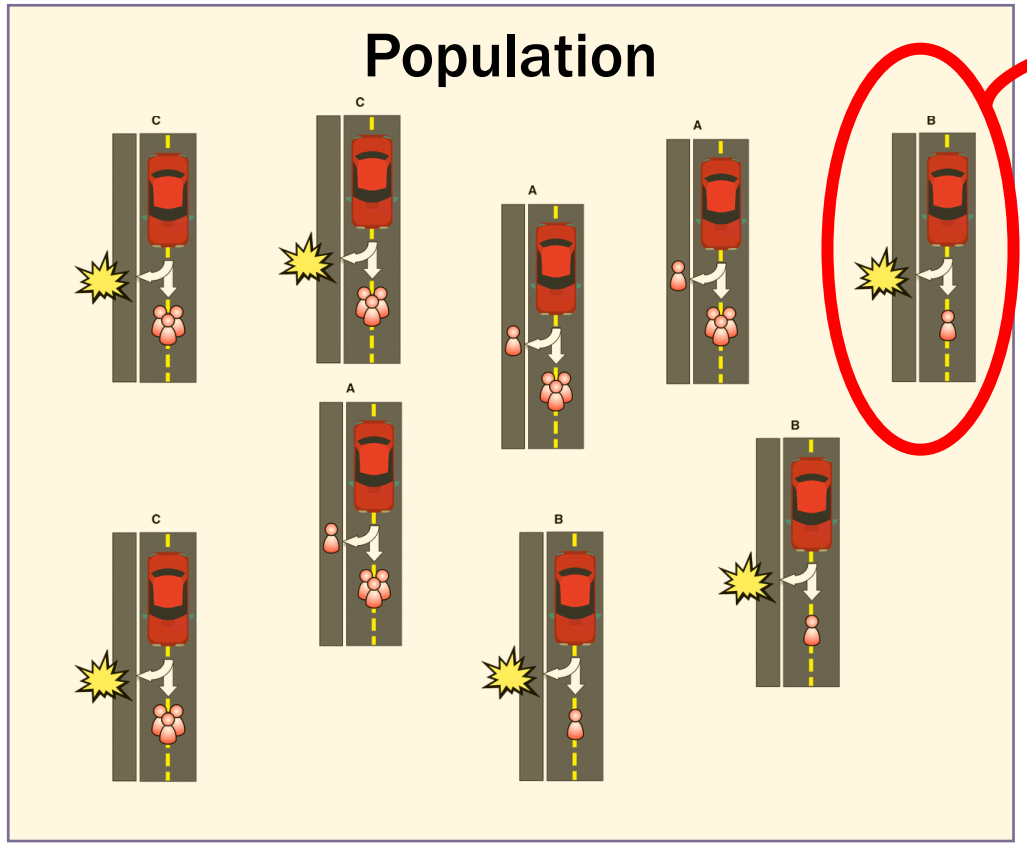


We represent an AV using a NN. The NN:

- Analyzes the scenario
- Outputs the level of the knob

The knob value is used to take an action

SIMULATION: POPULATION



Any scenario has:

- Altruism level
- Number of passengers
- Prob. of harming passengers
- Number of pedestrians
- Prob. of harming pedestrians

SIMULATION: EVALUATION

The notation:

- $nPed_{p_i}$: **number of pedestrians**
- $nPass_{p_i}$: **number of passengers**
- a_{p_i}: **intrinsic level of altruism for passengers in** p_i
- s_{p_i}: **intrinsic level of selfishness for passengers in** p_i
- $prodPed_{p_i}$: **probability of injuring pedestrians when the AV goes straight**
- $prodPass_{p_i}$: **probability of injuring passengers when the AV swerves**

SIMULATION: EVALUATION

The action is taken based on the assessment computed by the NN. The idea is pondering which action minimize harm with respect to relative importance of lives:

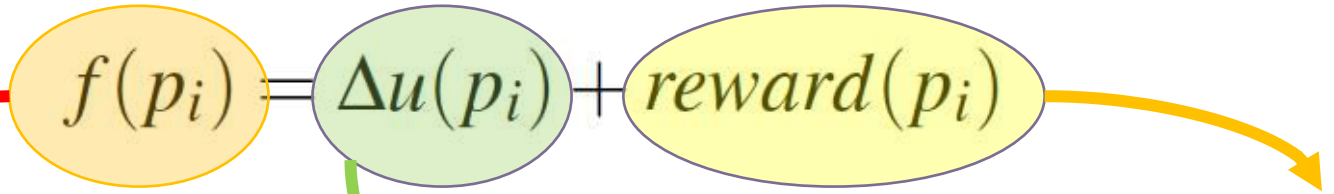
$$act_{p_i} = \begin{cases} 0 & \text{if } nPed_{p_i} \cdot probPed_{p_i} \cdot (1 - knob_{p_i}) \leq nPass_{p_i} \cdot probPass_{p_i} \cdot knob_{p_i} \\ 1 & \text{otherwise} \end{cases}$$

Go straight

Swerve otherwise

SIMULATION: EVALUATION

Individual is evaluated using the following fitness function:



Fitness of p_i

Difference between the utility of the choice made and the expected utility of the alternative choice

The social evaluation of the individual behaviour. Reward/punishment based on action taken by the average individual

SIMULATION: EVALUATION

$$\text{For each } p_i \in P,\ \Delta u(p_i) = u(p_i) - u_{alt}(p_i)$$

Utility for the taken action

Expected utility for the alternative choice

SIMULATION: EVALUATION

Depending on the taken action, the utility is computed based on the response of the scenario:

$$u(p_i) = \begin{cases} nPass_{p_i} \cdot s_{p_i} + (1 - dead_{p_i}) \cdot nPed_{p_i} \cdot a_{p_i} - dead_{p_i} \cdot nPed_{p_i} \cdot cPed & act_{p_i} = 0 \\ (1 - dead_{p_i}) \cdot nPass_{p_i} \cdot s_{p_i} + nPed_{p_i} \cdot a_{p_i} & act_{p_i} = 1 \end{cases}$$

where dead_{p_i} is 0 if people survived, 1 otherwise.

SIMULATION: EVALUATION

$$u(p_i) = \begin{cases} nPass_{p_i} \cdot s_{p_i} + (1 - dead_{p_i}) \cdot nPed_{p_i} \cdot a_{p_i} - dead_{p_i} \cdot nPed_{p_i} \cdot cPed & act_{p_i} = 0 \\ (1 - dead_{p_i}) \cdot nPass_{p_i} \cdot s_{p_i} + nPed_{p_i} \cdot a_{p_i} & act_{p_i} = 1 \end{cases}$$

Selfish utility preserving passengers

Altruistic utility obtained by preserving pedestrians

Total legal sanction (compensation) due for causing the death of a pedestrian

SIMULATION: EVALUATION

The second component is computed based on the alternative action:

$$u_{alt}(p_i) = \begin{cases} nPass_{p_i} \cdot s_{p_i} \cdot (1 - probPass_{p_i}) + nPed_{p_i} \cdot a_{p_i} & act_{p_i} = 0 \\ nPass_{p_i} \cdot s_{p_i} + nPed_{p_i} \cdot a_{p_i} \cdot (1 - probPed_{p_i}) + & \\ -nPed_{p_i} \cdot cPed \cdot probPed_{p_i} & act_{p_i} = 1 \end{cases}$$

Notice that single components are weighted using the likelihood of harming pedestrian/passengers in this case.

SIMULATION: EVALUATION

The reward depends on whether the AV's behaviour differs from the average behaviour of the community:

- If the average individual would go straight and the AV turns, then the action is rewarded (having done an action that is meritorious, since it minimizes the risk of losses more than the average)
- On the other hand, if the average individual would turn and the AV goes straight, then it is punished.

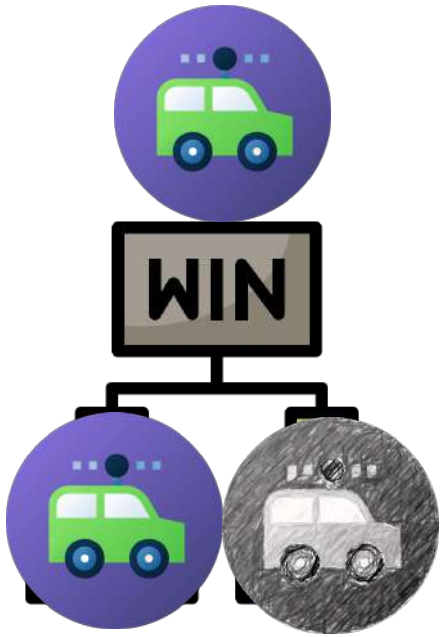
SIMULATION: EVALUATION

The reward depends on whether the AV's behaviour differs from the average behaviour of the community:

$$reward(p_i) = \begin{cases} 0.25 & \text{if } act_{(P,p_i)} = 0 \text{ and } act_{p_i} = 1 \\ -0.25 & \text{if } act_{(P,p_i)} = 1 \text{ and } act_{p_i} = 0 \end{cases}$$

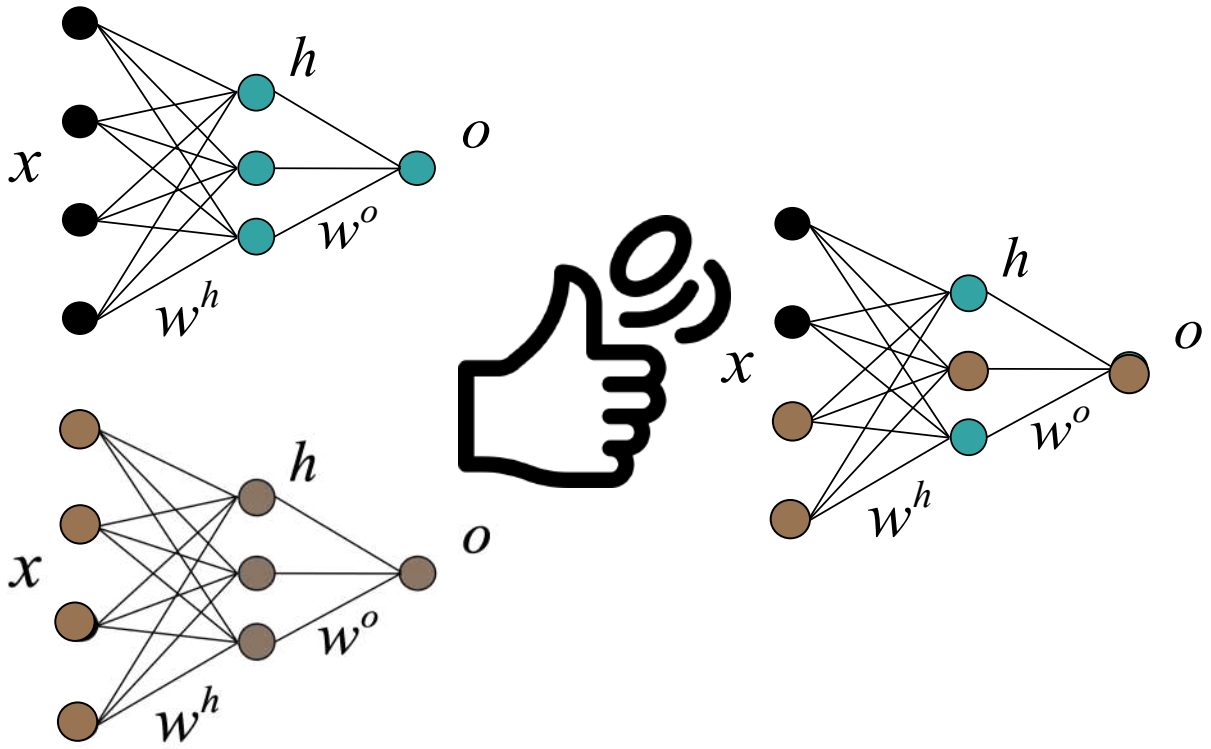
SIMULATION: SELECTION

Tournament selection: individuals are randomly paired. For each couple, the individual with the highest fitness is selected for reproduction.



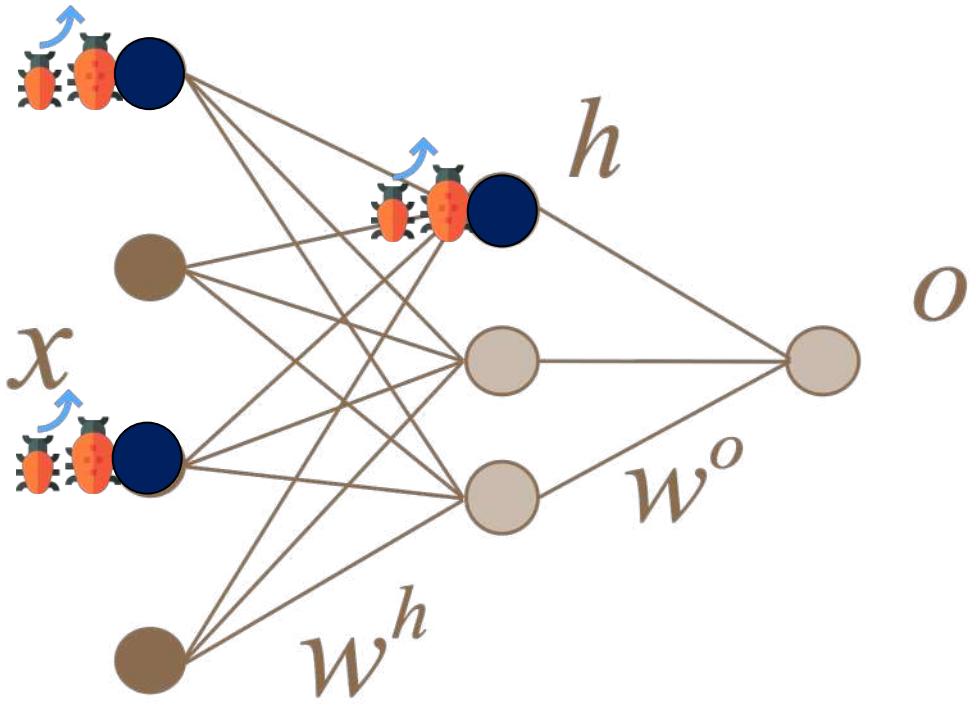
Icons made by Freepik from https://www.flaticon.com/

SIMULATION: CROSSOVER



- Mimicking the combination of genes that takes part in reproduction
- Chromosomes are represented by the weights of NN
- New chromosome by choosing at random one weight from one parent or the other.

SIMULATION: MUTATION



- It is applied to each child's chromosomes
- Alters certain genes with some probability
- It is used to prevent premature convergence

EMPIRICAL EVALUATION

- **Experiment 1:** $reward(pi) = 0$ **and** $cPed = 0$**. The aim is to test a simple situation in which the fitness function does not take into account any penalties from legal norms or any reward/stigma deriving from social norms.**
- **Experiment 2:** $reward(pi) = 0$ **and** $cPed = 1$**. The aim is to check whether legal norms may influence the system's performance.**
- **Experiment 3: the reward is in** $\{-0.25; 0.25\}$ **and** $cPed = 0$**. The aim is to explore whether social norms may influence the system's performance.**
- **Experiment 4: the reward is in** $\{-0.25; 0.25\}$ **and** $cPed = 1$**. The aim is to check whether and to what extent the combination of legal and social norms may influence the system's performance.**

EMPIRICAL EVALUATION

The prediction task can be seen as a binary classification task in which the AV learns to take the action which maximizes the payoff. In particular, looking at the fitness function, we classify samples as:

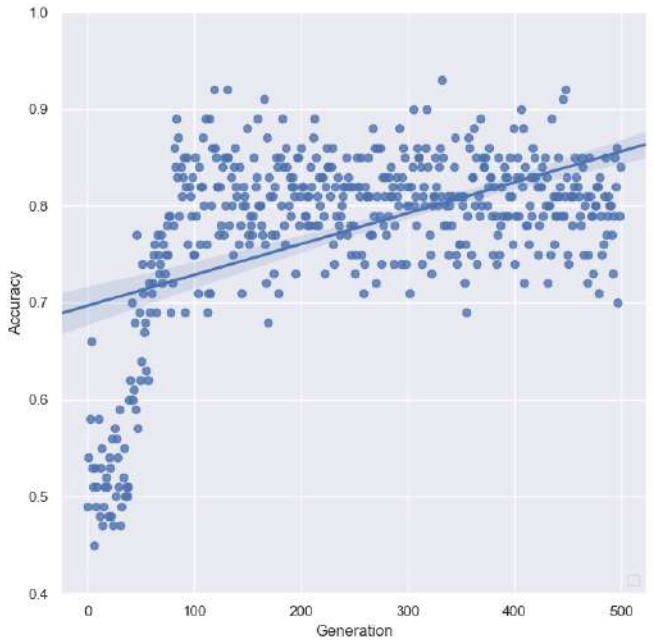
- Real Positive: the preferable action is to turn;
- Real Negative: the preferable action is to go straight;
- Predicted Positive: the neural network predicts a knob level which makes the AV turn;
- Predicted Negative: the neural network predicts a knob level which makes the AV go straight.

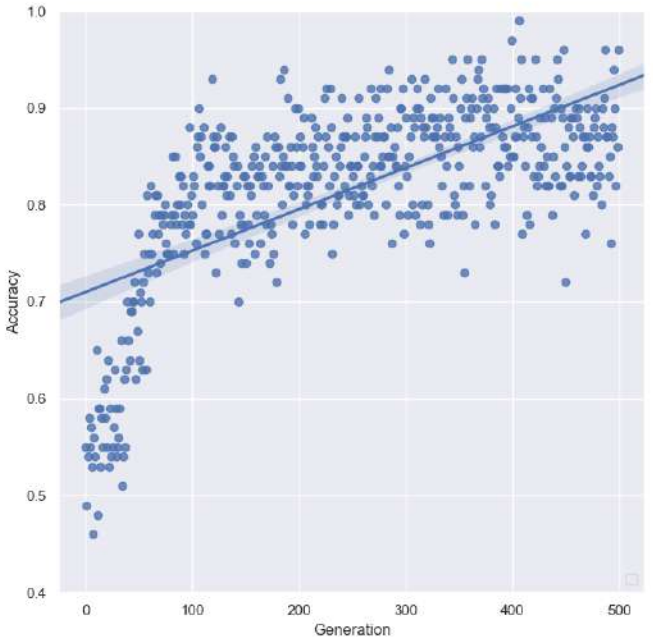
EMPIRICAL EVALUATION

Three different metrics:

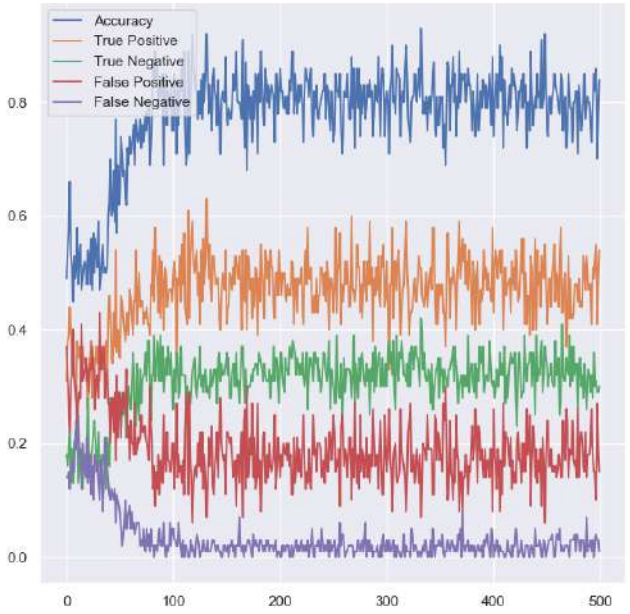
- Accuracy, which describes how many predictions coincide with the preferable actions;
- Confusion Matrix, which shows true positives, true negatives, false positives and false negatives;
- Number of victims, which describes the number of casualties that may be caused by an AV, using the knob values proposed by neural networks. In particular, the last metric is compared with number of victims caused by 3 different AVs: one which always minimizes the number of victims, one which always chooses the optimal action and one which always maximizes the number of victims.

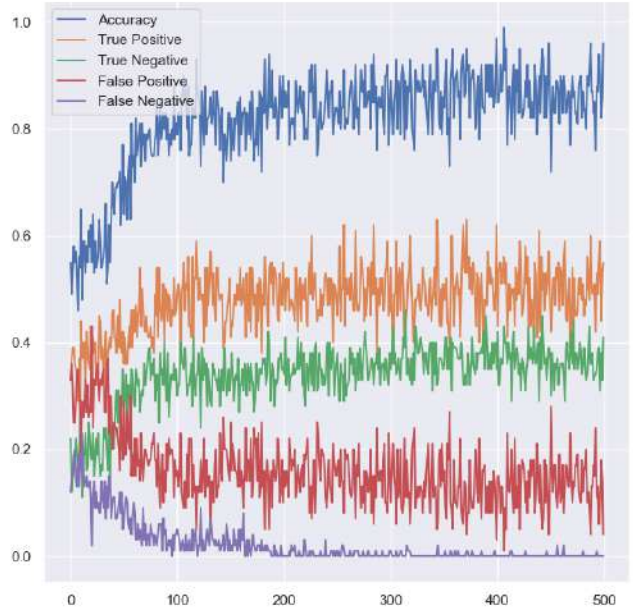
EMPIRICAL EVALUATION





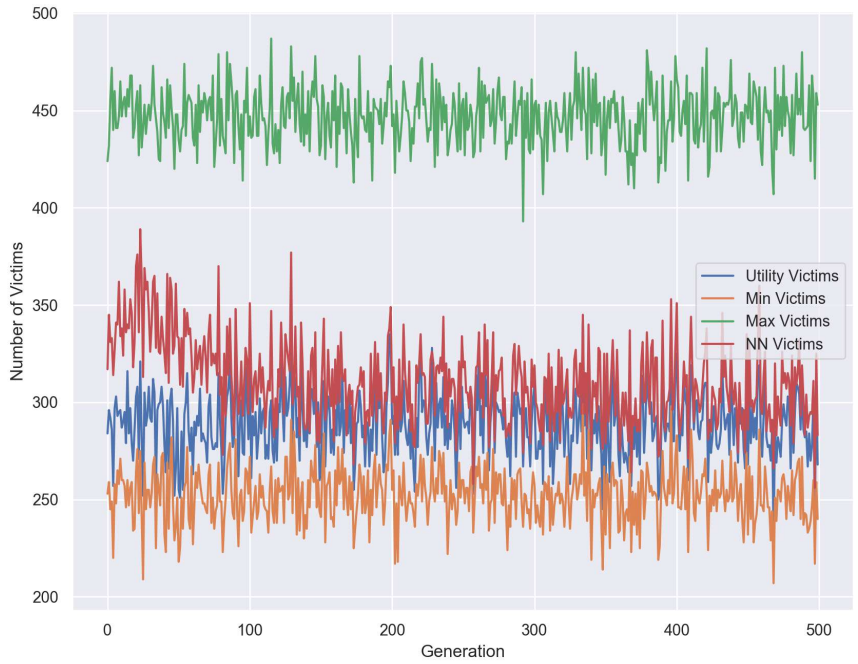
EMPIRICAL EVALUATION





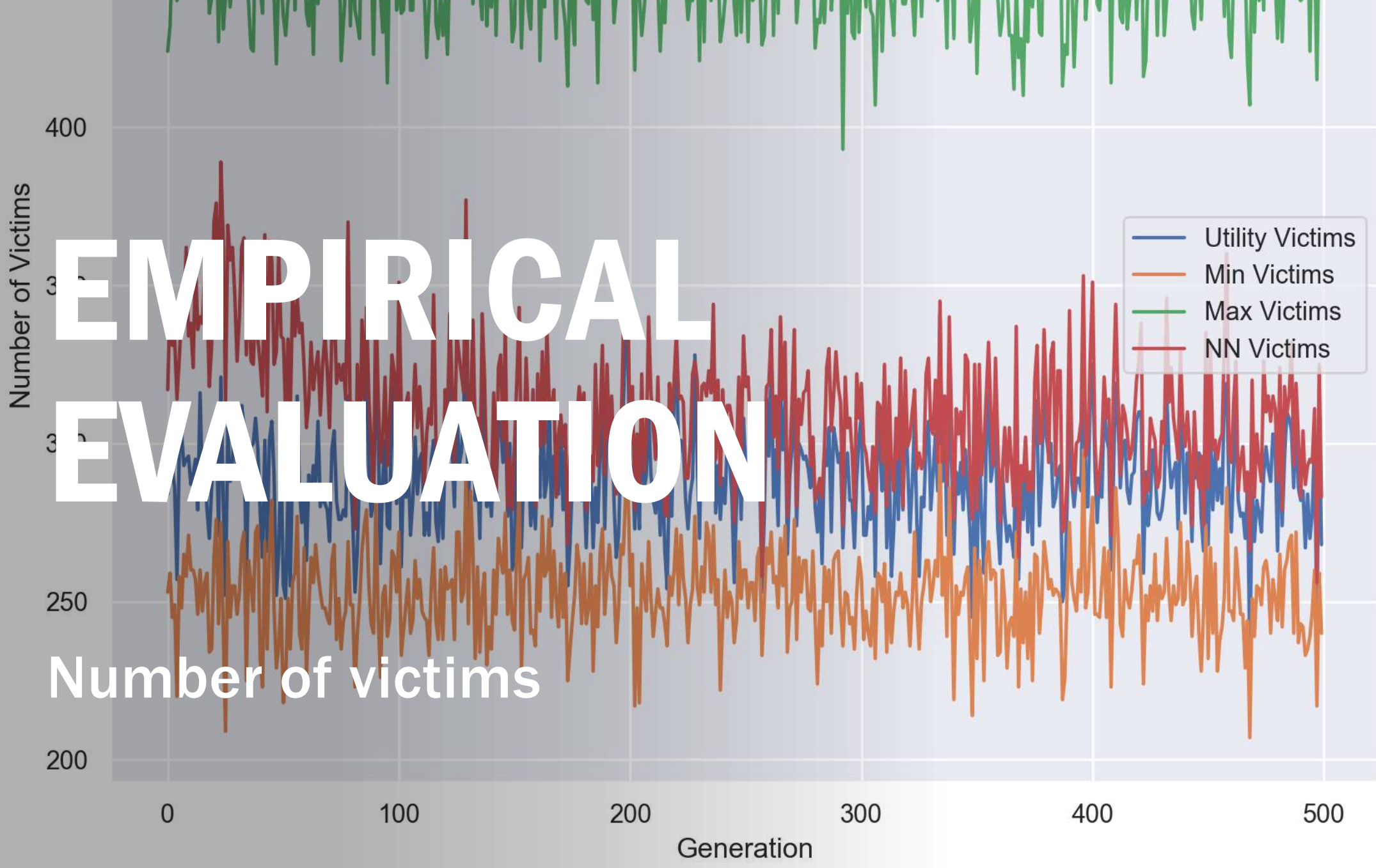
EMPIRICAL EVALUATION

Number of victims



EMPIRICAL EVALUATION

Number of victims



CONCLUSION/DISCUSSION

- What importance to give to the safety of passengers relative to the safety of pedestrians
- The assessment of the value of the AV's choices is dependant on considering the passengers' moral attitude (their intrinsic preferences) as well as legal sanctions and social norms (extrinsic incentives)
- Convergence of socially valuable behaviour can be obtained by providing appropriate mechanisms for sanction and reward

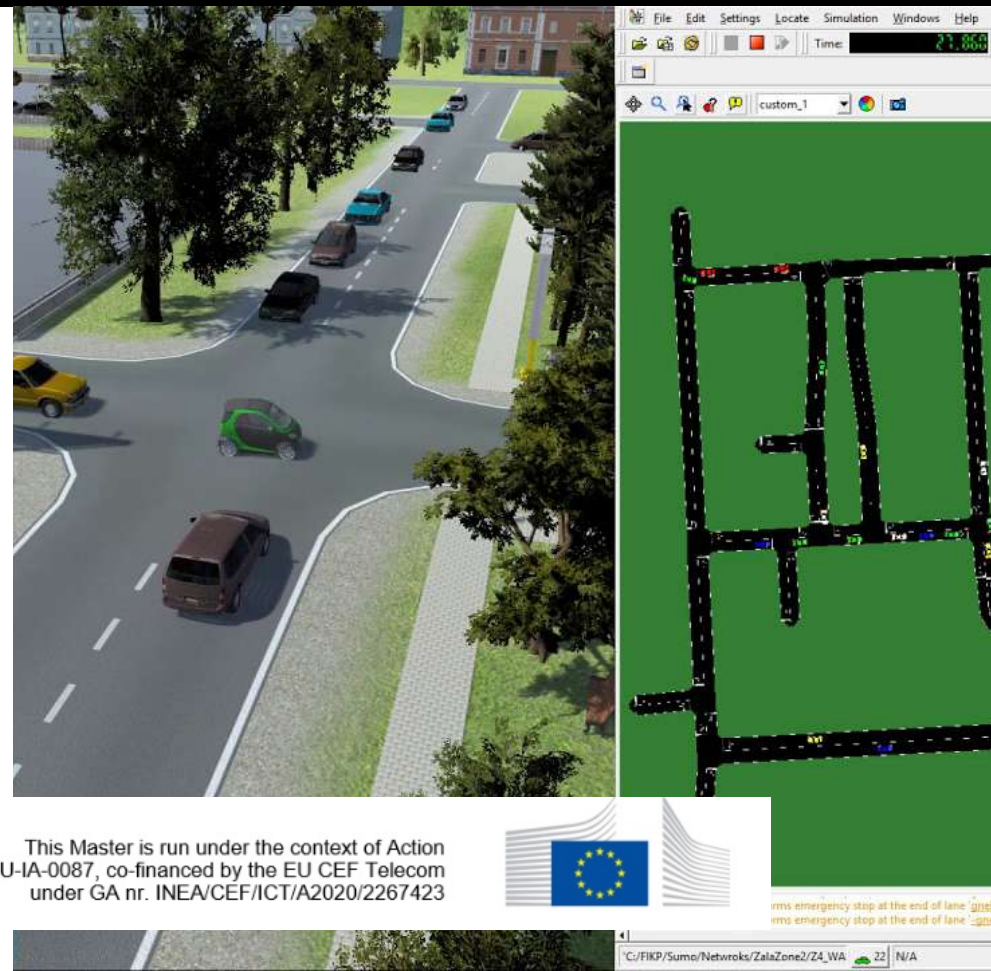
CONCLUSION/DISCUSSION

We aim to expand our model, for instance:

- Agents with memory
- Enabling agents to learn probability distributions
- Considering their past outcomes and those of observable others
- Adapting their ethical approach to societal preferences.

CONCLUSION/DISCUSSION

We also plan to insert out agents in existing traffic simulators (such as SUMO) to test our model in a dynamic environment.



MAI4CAREU Master programmes in Artificial Intelligence 4 Careers in Europe

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423