Legal Knowledge and Information Systems
S. Villata et al. (Eds.)

doi:10.3233/FAIA200854

A Genetic Approach to the Ethical Knob

Giovanni IACCA [c], Francesca LAGIOIA [a,b], Andrea LOREGGIA [b], and Giovanni SARTOR [a,b,1]

[a] *CIRSFID - Alma AI, University of Bologna, Italy*
[b] *European University Institute, Florence, Italy*
[c] *University of Trento*

AbstractAs Autonomous vehicles (AVs) are entering shared roads, the challenge of designing and implementing a completely autonomous vehicle is still open. Aside from technological issues regarding how to manage the complexity of the environment, AVs raise difficult legal issues and ethical dilemmas, especially in unavoidable accident scenarios. In this context, a vast speculation depicting moral dilemmas has developed in recent years. A new perspective was proposed: an "Ethical Knob" (EK), enabling passengers to ethically customise their AVs, namely, to choose between different settings corresponding to different moral approaches or principles. In this contribution we explore how an AV can automatically learn to determine the value of its "Ethical Knob" in order to achieve a trade-off between the ethical preferences of passengers and social values, learning from experienced instances of collision. To this end, we propose a novel approach based on a genetic algorithm to optimize a population of neural networks. We report a detailed description of simulation experiments as well as possible applications.

Keywords. Autonomous vehicles, Ethical Knob, Genetic Algorithm, Ethical Dilemmas

1. Introduction

Determining how self-driving cars should tackle moral decisions is a major challenge for designers, deployers and regulators. Scholars, policy makers, general media, blog posts and even dedicated websites discuss how AVs should behave in hypothetical accident scenarios, where they have to make decisions involving harms to humans [1,15,12]. Consider for instance the following scenario: in a dangerous and unavoidable accident situation, an AV must decide between staying on course and hitting several pedestrians or swerving, thus killing one passer-by. Should the AV sacrifice one person to save the lives of many? Imagine next that the choice of swerving will cause the passengers' death. Should the AV let its passengers die rather than driving into several pedestrians? Many academic articles [1,11,12] have discussed similar scenarios, on the basis of the classical Trolley Problem, i.e. the ethical thought experiment discussed by Foot [3] and Thomson [17].

[1]F. Lagioia, A. Loreggia and G. Sartor have been supported by the H2020 ERC Project "CompuLaw" (G.A. 833647).

In this context, scholars refer to different ethical theories, such as utilitarian [1], deontological, e.g., Kantian [6], virtue ethics [9] or contract theory [5,10] approaches, and investigate how to program AVs based on such theories [2,4,9].

A further question is whether all AVs should have the same mandatory ethics setting (MES) [5,11] or every user/owner should have the choice to select his or her own personal ethics setting (PES) [1,7]. It has indeed been claimed that an AV should have different ethics settings consistent with several ethical theories, allowing each individual passenger/owner to decide what moral approach her AV should have [16]. Thus, an AV would be considered and function as a "moral proxy" for drivers/owners ethical outlook, rather than a distinct "moral agent" [14]. It has also been argued that AVs could be equipped with an "ethical knob", enabling passengers to determine the degree to which the AV prioritizes their lives over the lives of third parties [2]. The provision of personal ethics settings reflects the value of autonomy and is sensitive to the moral views of the members of society. A recent web poll by robohub.org, concerning who should determine how an AV responds in ethical dilemma situations, supports this result. Most of the participants (44%) thought that the passengers should decide how an AV responds in ethical dilemma situations, while 33% thought that lawmakers should have the final say [13]. Despite the potential advantages, the idea of a PES has also attracted some criticisms, since people might then potentially choose their moral settings based on racist ideologies or other types of wholly unacceptable outlooks [11]. In this regard, we may question whether there might be a middle ground between a completely open choice of ethics settings and the view that everyone should have the same MES. Allowing for a PES does not mean that all conceivable trade-offs should be allowed, since certain morally troubling options could be ruled out.

In this paper we examine the possibility of providing AVs with the ability to learn how to set their ethical knob in such a way as to reconcile the individual preferences of their passengers and social values (as implemented through legal sanctions and social norms).

2. Ethical Knob, Individual Preferences and Social Values

In [2], it was assumed that the owner(s)/passenger(s) would set the ethical knob in their car by choosing a value from a continuous range between 0, denoting an extreme egoistic attitude (only passengers' lives are valued), and 1, denoting an extreme altruistic attitude (only pedestrians' lives are valued). Thus the knob was meant to express directly the ethical attitude of the AV passengers, i.e., the value they attribute to their life relative to the value of the lives of third parties (pedestrian potentially involved in road collisions). The AV would make the most advantageous choice, according to the set knob value, the number of lives at stake, as well as the probability that both passengers and third parties suffer harm, as a consequence of the driving decision.

In this work, we assume that the position of the knob no longer indicates the passengers' moral attitude, but rather the AV's assessment of the relative importance of the lives of passenger(s) and third parties. This assessment is the outcome of a learning process based on the the AV's engagement in accidents, and on its evaluation of the outcomes of such accidents. This evaluation takes into account the passengers' moral attitudes (their intrinsic preferences), as well as legal sanctions and social norms (extrinsic incentives).

In particular, the knob position is the outcome of an agent-based simulation, built on a genetic algorithm. Each step of the simulation takes into account 100 accidents simultaneously, each involving a particular AV. The simulation is run for 500 iterations. This process artificially "evolves" the knob values, according to the AVs utility function (which is parameterised to the moral attitude of the passenger(s) as well as to legal and social sanctions and rewards).

3. Methodology

Genetic algorithms [8] mimic the evolution process of a population that initially is made up of random individuals. Each individual is represented by a chromosome which is a possible solution to the problem being addressed. Individuals are evaluated based on a fitness function, indicating how well they perform. Better-performing individuals are given a higher chance of reproducing. In such a way, chromosomes that represent better solutions tend to spread in the population, while those representing inferior solutions tend to disappear. Small mutations, i.e., perturbation of some genes of chromosomes, may occur based on a random choice. This prevents the convergence toward a local minimum. In this work we implement the standard genetic algorithm schema described by the pseudo-code below 1.

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
```

In our simulation, the genetic algorithm maximizes the payoff of individuals involved in the population P. Each AV is an individual $p_i \in P$, constituted by a neural network. In each iteration, the individual AV is located in a scenario where it faces a dilemma and decides what action to take (i.e., it can either go straight and put at risk pedestrian(s) or swerve and put at risk passenger(s)). Each scenario is defined by the following variables:

- $nPed_{p_i}$: number of pedestrians. It is a random number in $[0, maxPed]$; in the experiments we set $maxPed = 6$.
- $nPass_{p_i}$: number of passengers. It is a random number in $[0, maxPass]$; in the experiments we set $maxPass = 6$.
- a_{p_i}: intrinsic level of altruism for passengers in p_i. It is a random number in $[0, 1]$.
- s_{p_i}: intrinsic level of selfishness for passengers in p_i. It equal to $1 - a_{p_i}$, namely, it is the complement of the level of altruism.

- $prodPed_{p_i}$: probability of injuring pedestrians when the AV goes straight. It is a random number in $[0,1]$.
- $prodPass_{p_i}$: probability of injuring passengers when the AV swerves. It is a random number in $[0,1]$.

In each scenario and before taking an action, a neural network evaluates the aforementioned features and predicts the value of the knob to decide the action to take. For the purpose of the genetic algorithm, the chromosome of each individual corresponds to parameters of the neural network.

Initialization and fitness evaluation. Initially, a population P of $n = 100$ individual AVs is created at random. At the beginning of each iteration, each individual $p_i \in P$ is located in a scenario instantiated at random. The AV chooses its action on the basis of its knob level and the confronted scenario, the choice being represented by the value of the variable act_{p_i}: if $act_{p_i} = 0$ the AV goes straight, if $act_{p_i} = 1$ it turns. The value of act_{p_i} is computed as follows:

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

The formula indicates that the AV goes straight (rather than turning) based on the comparison of two quantities. The first if obtained by multiplying the number of pedestrians, the probability of harming them by going straight, and their relative importance according to the knob position. The second is similarly obtained by multiplying the number of passengers, the probability of harming them by turning, and their importance. If the first quantity is lower than the second, the AV goes straight, while if it is higher it turns.

After the action has been chosen, the response by the environment is given by the variable $dead_{p_i}$, which indicates whether the action of going straight has resulted in pedestrians' injuries or whether the action of turning has resulted in passengers' injuries. The variable $dead_{p_i}$ may be randomly instantiated to 0 (safe) or 1 (harmed), based on the probabilities $probPed_{p_i}$ or $probPass_{p_i}$.

Then the fitness of each $p_i \in P$ is evaluated using the following function:

$$f(p_i) = \Delta u(p_i) + reward(p_i)$$

The fitness has two components. The first is the delta-payoff $\Delta u(p_i)$, which is the difference between the utility of the choice made and the expected utility of the alternative choice. Both utilities include an evaluation of the outcome (injuries to passenger(s) or pedestrian(s)) according to the intrinsic moral attitude of the passengers/owners, as well as to the legal sanction for unjustified harms. The second component, i.e., $reward(p_i)$, expresses the social evaluation of the AV's behaviour. It is positive when the AV's behaviour is better than the average of the population, while it is negative when it is worse than the average. The $reward(p_i)$ may be understood as the impact of such evaluation on an individual's self/social esteem. For each $p_i \in P$, $\Delta u(p_i) = u(p_i) - u_{alt}(p_i)$, where $u(p_i)$ represents the payoff gained in the scenario after the action taken and $u_{alt}(p_i)$ is the payoff that would be obtained through the alternative action. $u(p_i)$ is computed as follow:

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

where the first line yields the utility (for preserving people lives/health) obtained by going straight, and the second line yields the utility obtained by turning.

In the first line:

- $nPass_{p_i} \cdot s_{p_i}$ is the selfish utility obtained by preserving passengers (who are all preserved when the car goes straight)
- $(1 - dead_{p_i}) \cdot nPed_{p_i} \cdot a_{p_i}$ is the altruistic utility obtained by preserving pedestrians (in case they are not injured, i.e. $dead_{p_i} = 0$ even through the AV's choice puts them at risks)
- $dead_{p_i} \cdot nPed_{p_i} \cdot cPed$ it is the total legal sanction (compensation) due for causing the death of a pedestrian, where $cPed$ is the sanction for injuring a single pedestrians. The sanction is applied when the AV has behaved negligently, in the sense of choosing to harm pedestrians in a situation in which the expected harm to pedestrian exceeds the expected benefit to passengers. The value of $cPed$ is 1 if $probPed_{p_i} \cdot nPed_{p_i} > probPass_{p_i} \cdot nPass_{p_i}$, otherwise it is 0.

In the second line:

- $(1 - dead_{p_i}) \cdot nPass_{p_i} \cdot s_{p_i}$ is the selfish benefit obtained when passengers survive (even if they were put at risk)
- $nPed_{p_i} \cdot a_{p_i}$ is the altruistic utility obtained by preserving pedestrians

As noted above, the AV assesses the action it has performed by comparing the utility obtained through that action and the expected utility that would have been obtained by taking the alternative. The latter utility is computed according to the following formula:

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

In order to compute the social reward function, we need to know how individuals in the community behave on average. To do that, we compute the average knob of the community as $knob_P = \frac{1}{|P|} \sum_{p_j \in P} knob_{p_j}$. The average know is used to compute the action of an average AV in each scenario p_i. The action is computed replacing the value of $knob_{p_i}$ with the value of $knob_P$ in the formula of act_{p_i}.

We then check whether the action taken by the AV differs from the action that would be taken by the average individual. 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 blamed.

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

In this simulation, we have assumed that the level of altruism is randomly chosen.

Parents selection. After the evaluation step, a subset of individuals are selected as a basis to compute the next generation. For the selection process, we used a tournament selection algorithm: individuals are paired at random and those with the highest fitness in the couples are selected as parents. It should be noted that the same individual may appear more than once in the set of parents. In our experiments, the set of parents is set at 80% of the original population.

Crossover. The idea at the basis of the crossover operator is to mimic the combination of genes that takes part in reproduction. The chromosomes of one individual are combined with those of another individual. In such a way, the solution space is explored starting from a random point, and at each iteration the algorithm moves the solution towards a better solution. In this work, chromosomes are represented by the weights of neural networks. The crossover operator creates a new chromosome by choosing at random one weight from one parent or the other. Parents are paired at random to generate a new individual, until a new population of n individuals is generated.

Mutation. The mutation operator is applied to each child's chromosomes. It is used to prevent premature convergence, i.e., to avoid getting stuck at local optima. The operator acts on the chromosome by altering certain genes with some probability. In this work, if a gene is chosen for mutation, its value is randomly varied in a range of 1% of the original value.

4. Experimental Evaluation

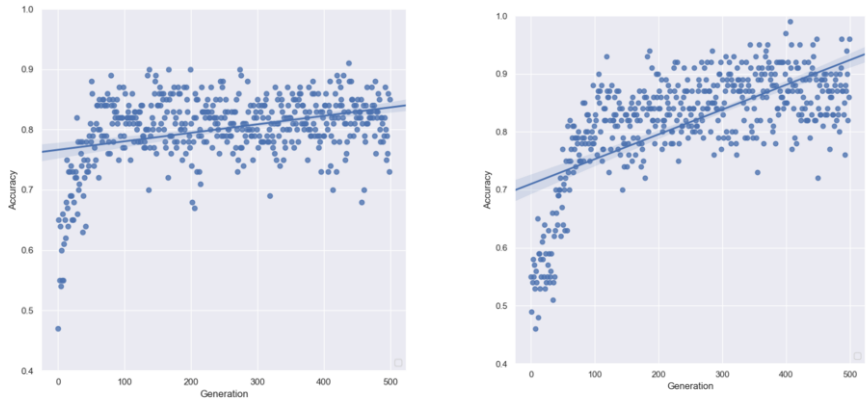
To evaluate the genetic algorithm described in the previous section and reported in the pseudo-code of Algorithm 1, we developed a Python 3.6 framework that implements it. Neural networks are defined using Keras ver. 2.2.4 over Theano ver. 1.0.3. Each individual $p_i \in P$ represents an AV, having the following features:

- Altruism level (i.e., a_{p_i}): a number in the range $[0,1]$, which describes how much an individual cares about the others relative to itself;
- Fitness (i.e., $f(p_i)$): a value which describes the goodness of the individual with respect to its behaviour in the population;
- Knob Level (i.e., $knob_{p_i}$): a number in the range $[0,1]$, representing the Ethical Knob described in [2]; the device determines the behaviour of the AV in ethical dilemma situations;
- Neural Network: it computes a regression task whose objective function is to optimise the level of the knob. In our empirical evaluation, the neural network has the following characteristics: 3 layers, one input layer with 5 nodes, one hidden layer with 3 nodes and one output layer with one node. The ReLu is used as the activation function for the hidden layer, while tanh is the activation function for the output layer.

In order to evaluate the performances of the framework and analyze whether the genetic approach is able to optimize the neural network, we performed four different experiments. Such experiments aim to test different approaches, as described below.

Experiment 1: $reward(p_i) = 0$ 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.

Figure 1. Accuracy for different settings: each blue dot represents the number of individuals in the population who take the action that maximizes the fitness function.



(a) Experiment 2

(b) Experiment 3

Table 1. Accuracy and confusion matrix (standard deviation in brackets) for the different settings. In each scenario all the features are drawn randomly.

Setting	Accuracy	TP	TN	FP	FN
Experiment 1	0.8487 (0.01)	0.5390 (0.00)	0.3097 (0.01)	0.1403 (0.01)	0.0110 (0.00)
Experiment 2	0.8442 (0.00)	0.5600 (0.00)	0.2842 (0.00)	0.0358 (0.00)	0.1200 (0.00)
Experiment 3	0.9467 (0.01)	0.5500 (0.00)	0.3967 (0.01)	0.0533 (0.01)	0.0000 (0.00)
Experiment 4	0.8357 (0.01)	0.6800 (0.00)	0.1557 (0.01)	0.1643 (0.01)	0.0000 (0.00)

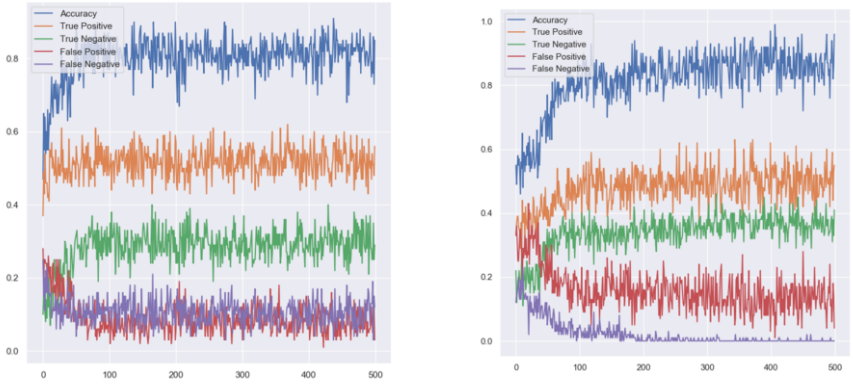
Experiment 2: $reward(p_i) = 0$ and $cPed = 1$. The aim is to check whether legal norms may influence the system's performance.

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.

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

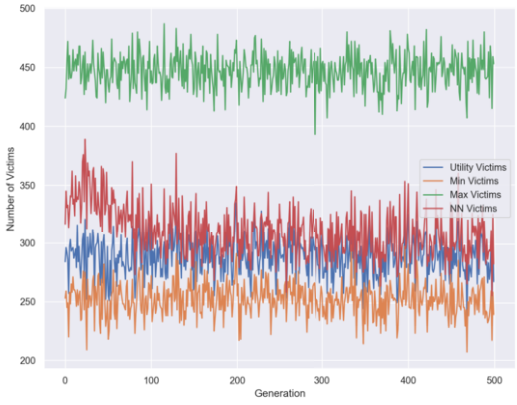
Figure 2. Accuracy and confusion matrix for different settings: blue line reports accuracy, orange line reports true positive, green line reports true negative, red line reports false positive and purple line reports false negative.



(a) Experiment 2

(b) Experiment 3

Figure 3. Number of victims over 500 generations in a deterministic environment. Different lines represent different approaches. In this setting, the utility function line coincides with the min function.



4.1. Data analysis

Based on the previous definitions, for each experiment we plot 3 different metrics, describing how individuals in the population evolve, generation after generation. Specifically for each generation we plot: 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. Figure 1 shows the accuracy for Experiments 2 and 3, in both of which, the accuracy increases. This suggests that neural networks in the population improve generation after generation. Notice that the increase of accuracy in Experiment 3 is steeper, suggesting that the opinion of the community (i.e., the stigma or the honour given based on average behaviour in the community) has a higher influence on the evolutionary process. When no reward is used–as in Experiment 1 and 2– or when it is applied in combination with a cost for harming pedestrians-like in Experiment 4–increment of the performance is less evident.

In order to understand whether the different components of the fitness function influence the final payoff of an individual, we performed a post-hoc analysis. Figure 2 shows the values of confusion matrix for Experiment 2 and 3 during the evolutionary process: the introduction of a cost decreases the number of false positives (see Figure 2a). On the other hand, the reward seems to reduce the number of both false positives and false negatives (see Figure 2b).

Moreover, at the end of the simulation when the networks are optimized, we generate 100 new scenarios. We use them as input for all the neural networks in the population and then count how many scenarios per individuals were tackled correctly. Table 1 shows the average values of accuracy and standard deviation when the probability of death for both the passenger and pedestrians is set to 1. Even though the accuracy is high, for some scenarios a high number of false positives can be noted. We conjecture this is due to the introduction of the reward. Indeed, whenever the AV is in doubt (usually because the number of pedestrians is close to that of passengers) the neural network predicts a false positive rather than a false negative, since the former can be rewarded whenever the community considers the AV choice an heroic action. The number of victims in the scenario is a metric which describes the system's ability to mediate between individuals' preferences and casualties minimization. Figure 3 shows the number of victims caused on 4 different approaches. The green line represents the maximal number of victims for each generation, while the orange represents the minimal number of victims for each generation. It is interesting to notice the following: firstly, when the AV operates in a deterministic scenario (i.e., the probabilities of harming pedestrians and individuals are set to 1), the utility function works as a proxy for the min function. Secondly, the number connected with the neural network prediction (i.e., the red line) decreases very quickly after a few generations. This is a signal that the optimization process is working towards the desired direction.

We have also run experiments where the ethical attitude of individuals (car owners/passengers) was given, rather than being randomly assigned. Not surprisingly in this case reducing the number of deaths required higher legal sanctions or social rewards.

5. Conclusion and future research

We have presented a model where AVs learn how to set their knob, i.e., what importance to give to the safety of passengers relative to the safety of pedestrians. This is obtained by having the AVs make choices and learn from the value of the outcome of such choices. The assessment of the value of the AV's choices is dependant on considering the pas-

sengers' moral attitude (their intrinsic preferences) as well as legal sanctions and social norms (extrinsic incentives). In particular, the merit of a choice has been determined by comparing the outcome of the choice and the expected outcome that would be obtained by making a different choice. An alternative model, which only takes into account the absolute outcome of a choice (the number of individuals not harmed minus those that were harmed, plus the applicable sanction and reward), has been considered. The learning takes place thanks to an evolutionary algorithm that differentially replicates the AVs making most successful choices. The results obtained show how convergence of socially valuable behaviour can be obtained by providing appropriate mechanisms for sanction and reward. In the future we aim to expand our model. For instance, we intend to endow our agents with memory — enabling them to learn probability distributions by considering their past outcomes and those of observable others—, and to model how individual ethical approaches are influenced by societal preferences. We plan to insert out agents in existing traffic simulators (such as SUMO) to test our model in a dynamic environment. This will enable us to address more complex and realistic traffic situations, involving multiple choices under resource constraints. Finally, we will investigate possible regulations of the setting of knobs, particularly in regard to liability issues (see [2]).

References

[1] J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[2] G. Contissa, F. Lagioia, and G. Sartor. The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, 25(3):365–378, 2017.

[3] P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, (5), 1967.

[4] J. C. Gerdes and S. M. Thornton. Implementable ethics for autonomous vehicles. In *Autonomes fahren*, pages 87–102. Springer, 2015.

[5] J. Gogoll and J. F. Müller. Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, Jul 2016.

[6] J. K. Gurney. Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law. *Alb. L. Rev.*, 79:183, 2015.

[7] J. Himmelreich. Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, 21(3):669–684, 2018.

[8] J. Holland. Adaptation in natural and artificial systems: an introductory analysis with application to biology. *Control and artificial intelligence*, 1975.

[9] W. Kumfer and R. Burgess. Investigation into the role of rational ethics in crashes of automated vehicles. *Transportation research record*, 2489(1):130–136, 2015.

[10] D. Leben. A rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, 19(2):107–115, 2017.

[11] P. Lin. Here's a terrible idea: Robot cars with adjustable ethics settings. *Wired*, 2014.

[12] P. Lin. *Why Ethics Matters for Autonomous Cars*, pages 69–85. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[13] J. Millar. You should have a say in your robot car's code of ethics. *WIRED. com*, 2014.

[14] J. Millar. Technology as moral proxy: Autonomy and paternalism by design. *IEEE Technology and Society Magazine*, 34(2):47–55, 2015.

[15] S. Nyholm and J. Smids. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice*, pages 1–15, 2016.

[16] A. Sandberg and H. Bradshaw-Martin. What do cars think of trolley problems: ethics for autonomous cars. *Beyond AI: Artificial Golem Intelligence*, page 12, 2013.

[17] J. J. Thomson. Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217, 1976.