A Car that Kills

Predicting the Fairness of Moral Dilemmas in Autonomous Vehicles

Theo Pijkeren
s4481046
Supervisor: dr. W.F.G. Haselager

Bachelor Thesis

Faculty of Social Sciences
Radboud University
The Netherlands
July, 2017

Contents

Abstract

Many traffic accidents could most likely be avoided when autonomous vehicles (AVs) are widely used. However, even with perfect sensing, AVs cannot ensure full safety and some AVs will certainly crash. When a crash is unavoidable, the AV could end up in a situation where it will need to choose between the lesser of two evils. Asking people to give their opinions about these situations could give us an understanding about what moral decisions are preferred. However, it is impossible to ask people's opinion on every possible traffic situation. In order to solve this problem, I trained an artificial neural network (ANN) that tried to predict the human evaluation of traffic situations where a moral choice must be made. The network has been trained on filled-in questionnaires about these moral dilemmas. The goal of this research is to see to what extent a ANN can predict these human evaluations. The results show that the ANN is not able to predict the human evaluation on these traffic situations. This is most likely the case because the ANN has only been trained on forty-two instances. However, the humans ability to morally judge a situation is really complex and this might be another reason why the ANN is not able to generalise to new situations.

1 Introduction

Imagine yourself driving on a bridge and a bus is driving towards you on the other lane. Suddenly, the bus switches lanes and drives straight at you. What do you do? If you do nothing, you will crash with the bus, damaging both vehicles and their passengers. You also have the option to sacrifice yourself by driving off the bridge. This way you save the bus and all its passengers, but you will most likely not survive the crash yourself. A third option available to you is to try to quickly sneak past the bus in order to avoid a collision all together. However, the chance of this succeeding is very slim and there is a high probability the bus will collide with you from the side. This would cause even more damage to both vehicles than a head-on collision. What would *you* do in such a situation?

Normally, you do not have to think about these moral dilemmas as they seldom occur. Luckily, the situation I just described is a purely hypothetical situation and it has been taken from a paper of Goodall [1]. The only difference between his scenario and the one described here is that in his scenario the car was not controlled by a human driver, but instead it was an autonomous vehicle (AV). Programmers of these AVs need to know which decision to make in these moral dilemmas.

One way to get an understanding of human ethics on this topic is by asking people to fill in questionnaires. This has been done by Zwerver [2]. Her questionnaires described different traffic situations and her participants were asked to evaluate the different decisions made by the AV. Unfortunately, there is a major drawback to this approach. There are many different traffic situations imaginable that require a form of moral decision-making and it is impossible to

have people fill in questionnaires for all of them. In my research I tried to solve this problem by training an artificial neural network (ANN) on these filled in questionnaires. The goal of this ANN is to predict human fairness evaluations in different traffic situations.

The aim of my research is to find out to what extent an ANN can predict the human fairness evaluation of traffic situations. I will answer this research question by comparing the predictions of a trained ANN with the answers to the filled-in questionnaires. In my experiment, the participants give their fairness evaluations on a scale from -5 (very unfair) to 5 (very fair).

My expectation is that the ANN will at the very least perform better than an algorithm that randomly picks an evaluation from -5 to 5. The eventual goal is to create an ANN that will have perfect to near-perfect prediction. This would mean that the ANN accurately predict that human fairness evaluation in traffic situations.

2 Background

The advantages of AVs look very promising. One of these advantages is the reduced number of traffic accidents [3] and another is that AVs will reduce pollution [4]. Although, the number of accidents can be reduced, AVs are not able to ensure perfect safety according to Benneson *et al.* [5]. In their research, they show that even AVs with perfect sensing cannot avoid a crash in certain situations. Goodall follows up on this by claiming a moral component is needed when these crashes occur [1]. He illustrates his point by giving the example described in the introduction of this paper. It might seem very unlikely that these kind of situations will occur, however, Google's self-driving car team suggests that ethics are already being used in their cars: "What if a cat runs into the road? A deer? A child? These were moral questions as well as mechanical ones, and engineers had never had to answer them before" [6].

How do AV manufacturers, like Google, know which moral decision to make? According to Bonnefon *et al.* surveys will help with this problem [7]. By basing the AV's decision-making on these surveys, more people are willing to buy these AVs according to the researchers. This survey-based approach has been investigated by Zwerver [2]. She asked her participants to morally evaluate the decisions of AVs in different traffic situations. Another goal of her research was to investigate which factors underlie moral evaluations in these traffic situation. An example of these factors is the factor discrimination: a situation could be evaluated as very unfair because discrimination is at hand. Zwerver suggests these kind of factors could predict the human fairness evaluations. In my research I tried to achieve this: using these factors in order to predict the human fairness evaluations with an ANN.

A lot of research has already been done on the use of ANNs in AVs [8]–[12]. Most of this research focuses on controlling the AVs with ANNs, rather than aiding the manufactures when programming these AVs to deal with moral dillemas.

3 Design

3.1 Questionnaires

To train the ANN to predict human fairness evaluations, questionnaires will have to be created. This section will explain these questionnaires in detail. Participants of this research were asked to visit my online questionnaires website [13]. This website was made from scratch from a combination of HTML, CSS and JavaScript. Figure 1 shows the content of the site used for the training phase. A full overview of all the scenarios used for this research can be found in appendix A. The questionnaires used for testing are shown in figure 2. All testing scenarios are explained in appendix B. These testing scenarios were arbitrarily created, although most scenarios are quite similar to the once used for training.

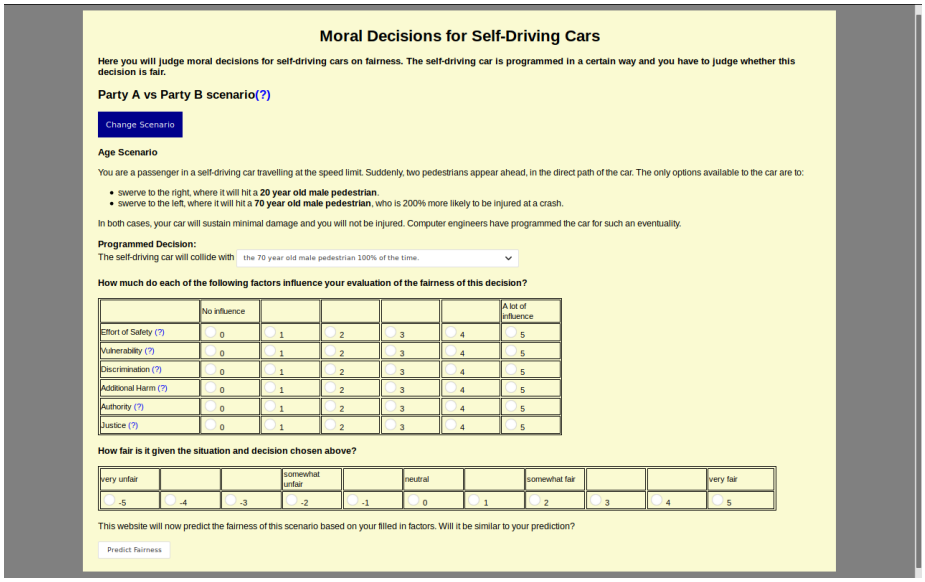

Figure 1 – The web-based questionnaire used for the experiment

3.1.1 Scenario Description

First, a scenario containing a moral dilemma was presented to the participants. In total there are ten different scenarios. Each scenario is part of a scenario group which indicates the type of scenario. The different scenario groups are explained in table 1. An example of one of the scenarios is the age scenario: in this scenario an AV is driving at speed limit and suddenly two pedestrians appear in ahead of the car. Unfortunately, the car is unable to stop in time to avoid collision. The only options still available to the car are either to swerve right and collide with a 20 year old male pedestrian or to swerve left and hit a 70

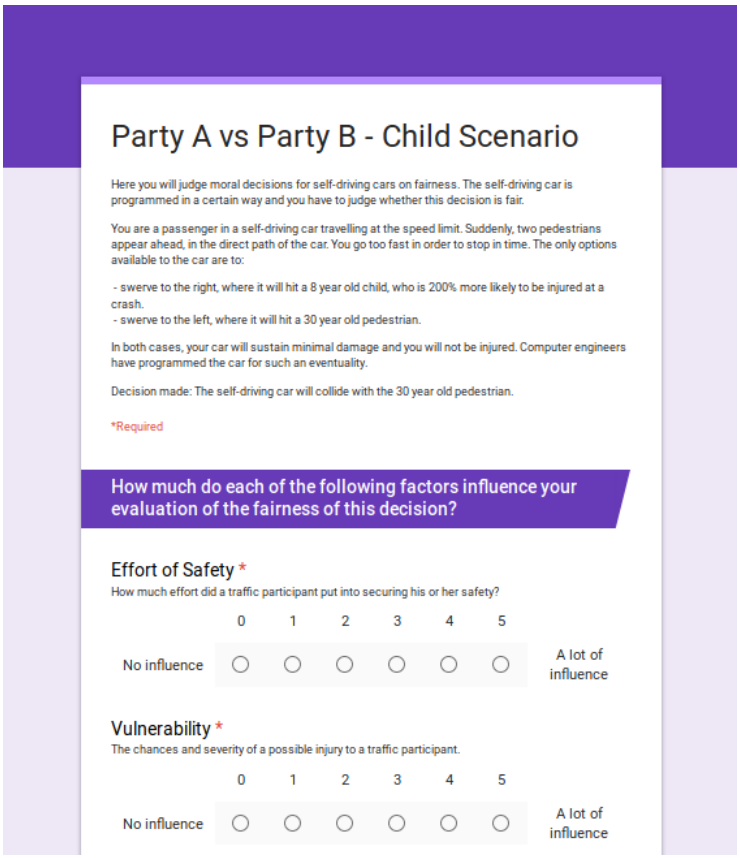

Figure 2 – The questionnaire used for testing

year old male pedestrian.After such a situation was described the programmed decisions were stated. The programmed decision dictates what the AV will do in this particular scenario. An example of this in the scenario described above is to always collide with the older pedestrian. All possible programmed decisions are described in appendix A.

When participants visited the website, a random scenario was shown to the participants. This was done to ensure the scenarios were distributed equally over all participants. Participants were able to change the scenario or decision if they wanted to evaluate another situation.

3.1.2 Fairness Evaluation

Now the participants are asked to evaluate the programmed decision on fairness. Participants rate the decision on a scale from -5 to 5 corresponding to a very unfair and very fair decision respectively. Besides this fairness evaluation, participants were also asked to give their reason for this particular evaluation. This

Scenario group	Description
Party A vs Party B	In the scenarios of this group a choice has to be made between two different parties.
Advantage vs. Risk	In the advantage vs. Risk scenarios a decision has to be made whether to drive below speed-limit or take a detour.
Occupant vs. Party A	A choice has to made between colliding with a wall or party A.

Table 1 – The different scenario groups as discussed by Zwerver's paper with my own explanation.

was done by letting participants indicate to what extent the factors described in table 2 influenced their fairness evaluation. The description mentioned in the table could be viewed by the participants by hovering over little question marks with their mouse. In the ongoing example the factor vulnerability might influence the participants to evaluate this decision as unfair as the older pedestrian is much more vulnerable.

Factors	Description
Effort of safety	How much effort the traffic participant put into securing his or her own safety.
Vulnerability	The chances and severity of a possible injury to a traffic participant
Discrimination	One group of traffic participants would be targeted over another, for example based on gender, age or more trivial things.
Additional harm	Emotional harm, damage to property or even feeling more unsafe.
Authority	Accepting traffic regulations set by the authorities and not changing our behaviour even if it could reduce risks.
Justice	Traffic participant's past in terms of moral or immoral behaviour.

Table 2 – The different fairness factors as discussed by Zwerver's paper with my own explanation. Zwerver has chosen these factors based on an informal examination of which aspects could influence the evaluations of her participants. I used the same factors in my experiment, as this would have allowed me to use Zwerver's data. Unfortunately, it turned out not to be possible to use her data for training the ANN as her experiment was too different from mine.

3.2 Pilot Design

In order to have a successful experiment, the questionnaires described in section 3.1 need to be understood correctly by the participants. To see whether this is the case, a pilot experiment has been conducted. Figure 3 displays the website the participants of the pilot were shown. After looking at this website, the participants were asked to answer some questions about it. The following questions were asked to the participants of the pilot:

1. Age, Sex, study/work background
2. What is the goal of the site? Shortly explain what you think the site is for.
3. What are the sliders for? What do they represent?
4. Did you hover over the question marks for more information?
5. If you answered 'No' above: why not?
6. What do you have to evaluate on the site?
7. Did you read more than one scenario?
8. Do you have any tips on how to improve the website? (optional)

The methods, results and conclusions of the pilot will be discussed in section 4.

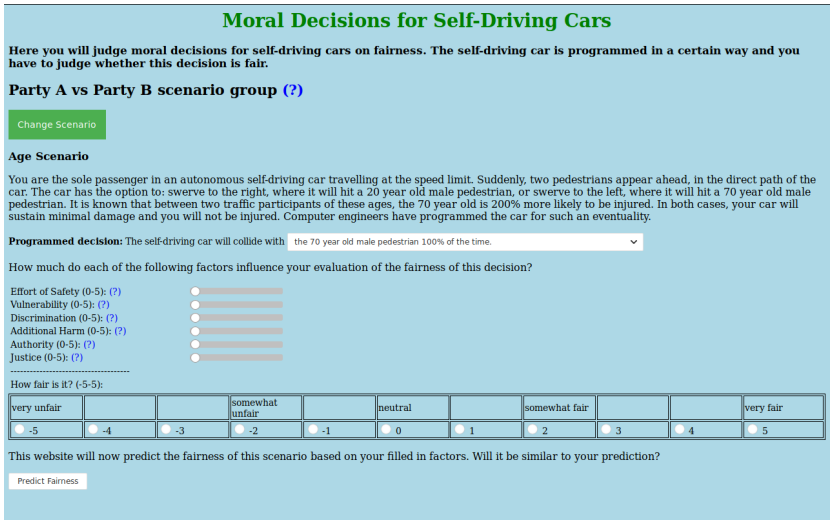

Figure 3 – Website viewed by participants of the pilot. This was not used in the actual experiment.

3.3 Artificial Neural Network

As stated before, the ANN was trained on the training questionnaires. The network tried to learn the relation between the factors that influence the human fairness evaluations and the fairness evaluations itself. This way the ANN would be able to predict the fairness evaluation based on these factors. After a participant filled in the questionnaire, the ANN predicted the fairness evaluation based on the scenario group, the programmed decision and the factors of the filled-in questionnaire. It might seem strange that the fairness evaluation is given to the ANN *before* the prediction is made. However, the ANN only uses this evaluation *after* the prediction and it is used for training.

The ANN was implemented on a online server as the training happened live. This means the participants were able to see the prediction of the ANN after they had filled in the questionnaire. The implementation of the ANN was done using a Java Servlet in Java 1.7 [14] with the machine learning Java library Weka 3.8.1 [15].

The type of ANN used, was a multilayer perceptron with one hidden layer. This hidden layer contained five nodes, which are all sigmoid. The weights were updated with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, because it allows for faster convergence compared to normal gradient descent [16]. The learning rate was set to 0.1. The ANN trained on all previous questionnaires a thousand times after a questionnaire was submitted. All values used have been normalised before training. The code used is shown in listing 1. A describition of each function call can be found on the documentation of Weka [17].

Listing 1 – ANN code in Java using Weka

```
// Create a MLP
mlp = new MultilayerPerceptron();
// Setting Parameters
mlp.setLearningRate(0.1);
mlp.setMomentum(0.0);
mlp.setTrainingTime(1000);
mlp.setHiddenLayers("1");
//Training
mlp.buildClassifier(train); // Train on the previous questionnaires
//Testing
Evaluation eval = new Evaluation(train);
eval.evaluateModelOnce(mlp,test); // Test on the current questionnaire
double result = mlp.classifyInstance(test); // Predict the fairness
```

4 Pilot Experiment

As explained in section 3.2, a pilot study was done to improve the quality of the website. Participants of the pilot were asked to view the site (see figure 3 in the design section) and navigate around for up to five minutes. After this time a link to an online survey in which their opinion on the website was asked. All the questions used in this pilot are already explained in section 3.2. Some of these questions will now be discussed in detail.

Certain questions, like questions two and six, directly asked participants about the main goal of the site, while others were to check whether specific elements of the site were clear. The site used in the pilot contained sliders for indicating which factors influenced their fairness decision. To check whether these sliders were understood correctly, a question about them was included in the pilot questionnaire. The same was done for the question marks that could be used to get extra information about certain elements of the site. Another doubt was whether people would navigate to other scenarios without any specific instruction to do so, which is why a question about that is also included. The last question is to allow participants to give feedback on the overall experience and to suggest improvements for the site.

4.1 Pilot Results

The results of the pilot were overall rather enlightening. Most people properly understood the topic of the site. However, one participant seemed to have some trouble understanding the question about the main goal of the site. Another participant did not mention self-driving cars or autonomous vehicles, which might indicate that the goal was not fully understood. On the other hand, the sliders were understood correctly by all participants, although someone did indicate that it was unclear what the different values represented (e.g. what does a discrimination value of five mean). All but one participant said to have hovered their mouse over the question marks for extra information. The one participant that did not use the question marks stated that he did not need the extra information, so it looked like these question marks were doing their job adequately. The question 'What do you have to evaluate on the site?' was understood differently by different participants. Some thought they had to give their evaluation of the site, while others understood they had to explain what a user of the website should evaluate once the actually experiment would start. The latter interpretation of this question was the intended meaning of the question and participants with this interpretation all had a good understanding of what the user is supposed to do. The other participants, the ones with the different interpretation, stated some general comments about the site (e.g. about the unclear background colour). All five participants stated to have read more than one scenario, which answers the question whether people would navigate to other scenarios. Lastly, the following points were made in the optional question about improvements to the site:

- Sliders are not easy to understand. Add meaning to the values of the sliders.
- Shorten the texts, make it more to the point.
- Make text easier to understand for a layperson.
- The blue background makes the site hard to read.
- Use margins on either side of the survey.
- More line breaks in the text.
- Add clear divisions between tasks.
- Make the text of the question marks pop-up, rather than only show when hovering over.
- Add images or graphics to aid the understanding of the moral questions.

4.2 Pilot Conclusions

The suggested improvement were quite helpful and most of the suggestion were implemented in the actual experiment. To improve readability, the colours were changed and margins were added to the site (see figure 1). The sliders were replaced by radio buttons with descriptions that indicated what each value meant. This way, the site properly explains that a value of five for the factor discrimination means this factor has a lot of influence. Finally, as suggested by some participants of the pilot study, the descriptions are shortened and more line breaks were added. All these changes improve the readability of the website. Due to time-constraints images or graphics to illustrate the scenarios were not added.

5 Methods

5.1 Training

In order to be able to answer the research question, an ANN needed to be trained on filled-in questionnaires. Participants of my study were asked to go to the website containing the questionnaires [13]. There, they evaluated the situation on fairness and also indicated which factors influenced their decision. As explained in the design section, each filled-in questionnaire was send to the ANN and a prediction of the ANN was given. Afterwards this questionnaire was added to the list of questionnaires used for training. The participants were able to see the prediction and they could compare it to their own evaluation. Forty-two questionnaires in total have been filled in and the ANN was trained on all of them.

5.2 Testing

After training a test is needed in order to show to what extent the ANN can predict the human fairness evaluation of moral dilemmas in traffic situations. The test consists of six new scenarios: two from each scenario group (see table 1). In total, five participants filled in these new scenarios in the same way as done at training. The same test scenarios were given to the trained ANN which on its turn tried to predict the human fairness evaluation. With both parties, human and machine, having done their part, all that rested was to compare the results in order to find out to what extent the ANN was able to predict the human fairness evaluation.

To see whether the ANN performed better than a random algorithm, the average error of both algorithms are compared. The error is calculated by taking the absolute difference between the prediction and the human fairness evaluation. This human fairness evaluation is the average evaluation of the five participants used for testing. For simplicity, the random algorithm will always predict a scenario to be neutral (fairness = 0). This is because the random algorithm picks a random number between -5 and 5 and on many trials the average evaluation will be 0.

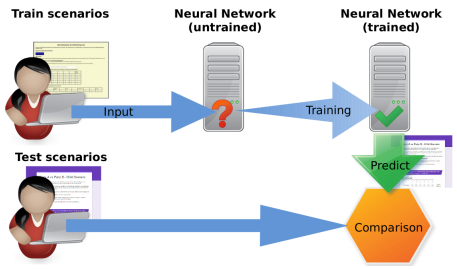

Figure 4 – A flowchart of the method of the experiment

6 Results

This section shows the results of the experiment. In total forty-two questionnaires have been filled in for training.

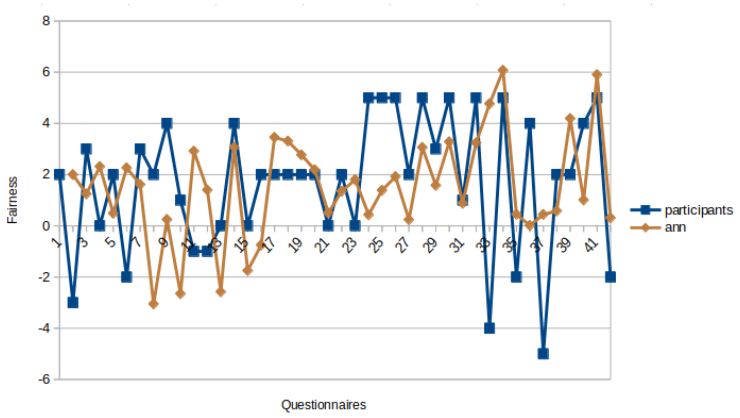

Figure 5 – The human fairness evaluation compared to the predicted evaluations of the ANN.

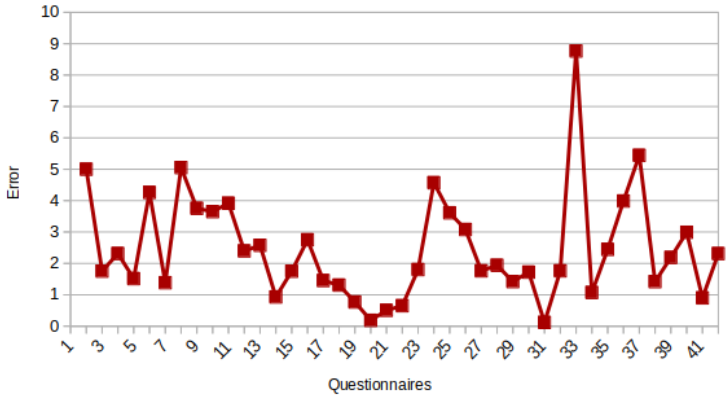

Figure 6 – Error of the artificial neural network (ANN) plotted against the number of questionnaires that the ANN is trained on. The error is the absolute difference between the human fairness evaluation and the predicted fairness of the ANN. At each instance the ANN is presented a new questionnaire containing a new scenario. However, the ANN uses the data of all previous instances in order to predict the fairness for the current instance. For example at instance fifteen, the ANN will predict the fairness of the fifteenth questionnaire and it uses all fourteen previous questionnaires to do so.

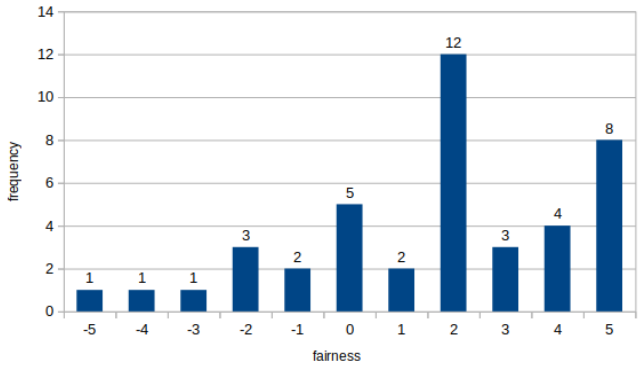

Figure 7 – Frequency of the human fairness evaluations ($n = 42$). The x-axis ranges from very unfair evaluations (-5) to very fair evaluations (5). The y-axis shows the frequency of these evaluations. This graph shows participants evaluated more scenarios as fair than unfair.

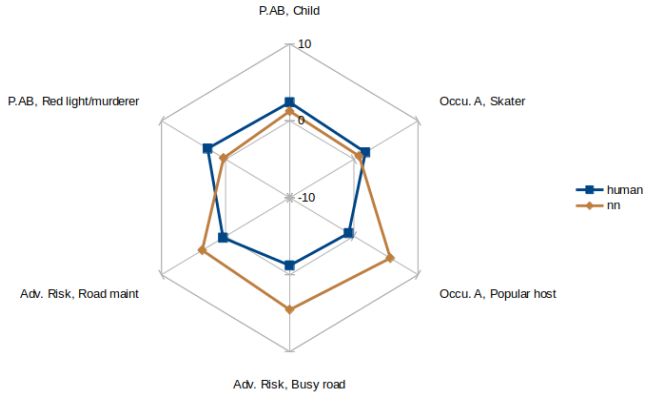

Figure 8 – Human fairness evaluation and ANN's prediction on the six test scenarios ($n = 5$). The six axes represent the six different scenarios. The human fairness evaluation is the average evaluation of all five participants. The centre of the graph represents a fairness value of -10 while the outer border represents a fairness value of 10. If the two lines are on top of each other, the ANN has perfectly predicted the human fairness evaluation. As an example, the average fairness evaluation over all five participants was -0.8 (neutral) for the popular host scenario (Occu A, Popular host) (see appendix B) . However, the prediction of the ANN was 5.6 (very fair).

7 Discussion

7.1 Conclusion

Unfortunately, the test results show that the ANN did not perform better than a random algorithm. Figure 8 shows that the road maintenance, busy road and the popular host scenario were all predicted as much fairer than the actual evaluation. The ANN's average error was 3.3, while the average error of the random algorithm was 1.6. This shows the ANN was not able to perform better than the random algorithm. However, the question remains whether the ANN is able to predict the human fairness evaluation when there is more training data. This could work, if the performance of the ANN improves with the number of instances it is trained on. When looking at the data of this research, we can see this improvement in performance for the first thirty-two training instances. This can be seen in figure 9 with a negative correlation ($r = -0.39$) between the error and the number of questionnaires.

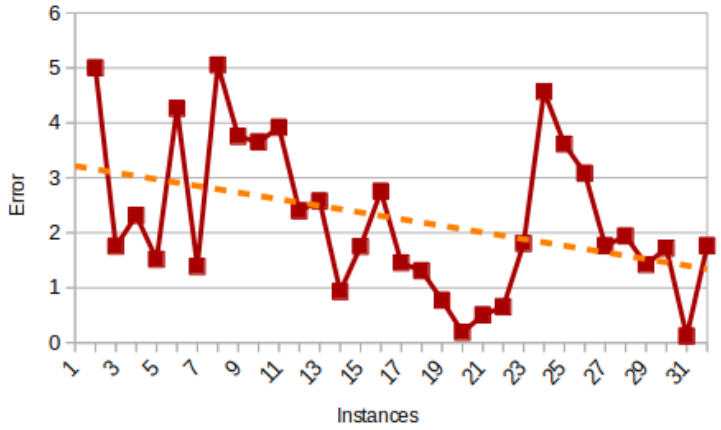

Figure 9 – Error of the ANN from questionnaire two to questionnaire thirty-two with the error of the ANN plotted against the number of questionnaires that the ANN is trained on. The error is the absolute difference between the human fairness evaluation and the predicted fairness of the neural network. A linear trend line is drawn ($r = -0.39$).

Although this downward trend looks promising, the error goes up again after questionnaire thirty-two (see figure 10). One huge error appeared at the thirty-third questionnaire. Here, the error, the absolute difference between the prediction and the actual evaluation, was 8.8. The scenario that was being evaluated at instance thirty-three was part of the *Advantage vs. Risk* scenario group. The programmed decision of the AV was to always drive at speed limit (20 mph) at a place where children might play. This was evaluated as quite unfair by the participant (fairness $= -4$). However, the ANN predicted the fairness to be very fair (fairness $= 4.8$). So why did the ANN predict this situa-

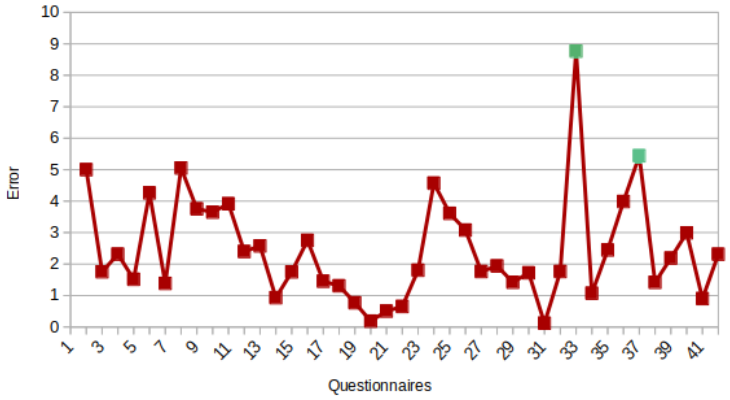

Figure 10 – Same error graph as figure 6 with added highlight on the error spikes at questionnaire thirty-two and thirty-seven.

tion as very fair? Most likely because almost no negative evaluations were given before instance thirty-three, as can be seen in figure 11. This bar chart shows distribution of the fairness results from all questionnaires prior to the thirty-third. This uneven distribution was not expected, but it can be explained. As described in the design section, the participants were able to change the scenarios as well as the programmed decision. After they had done the experiment, multiple participants explained they thought they had to pick the option they considered the fairest. These participants misunderstood the instructions, as the option to change the programmed decision was only included to give the participants more freedom. Because of this misunderstanding, many participants ended up evaluating decisions that were considered fair. This resulted in the ANN being trained on more fair than unfair scenarios. The expected result of this is an ANN that will have more fair prediction than unfair ones. This indeed turned out to be the case, as the ANN has only predicted five scenarios as unfair compared to the thirty-seven scenarios predicted as fair.

Another error spike is visible at instance thirty-seven. At this instance, the participant was evaluating a scenario from the *Occupant vs. Party A* group and the car was programmed to hit a wall 100% of the time in such a scenario. A reason for this error might be that the *Occupant vs. Party A* scenario was only evaluated six times. Moreover, the decision to hit the wall 100% of the time had only been evaluated three times prior to the thirty-seventh instance.

It is hard to say whether the ANN keeps improving when more questionnaires are being filled in. Based on my results, it seems that the ANN stops improving after questionnaire thirty-two. Even though some errors might be explainable, no definite conclusion can be made based on this research alone.

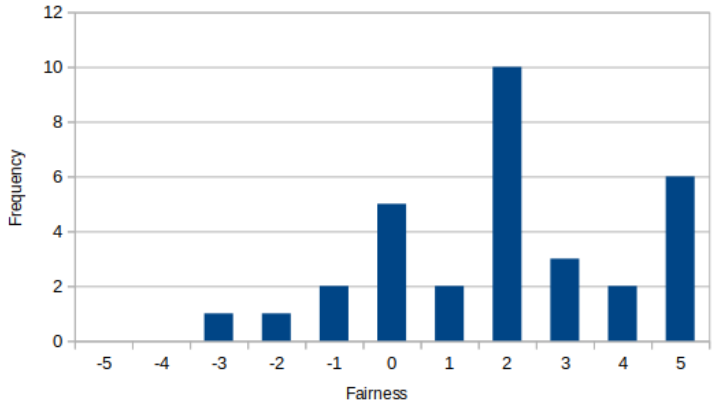

Figure 11 – Frequency of fairness evaluation for the first thirty-two instances. The x-axis range from very unfair evaluations (-5) to very fair evaluations (5). The y-axis shows the frequency of these evaluations.

7.2 Further Research

The problem of too few participants might be solvable by making the filling-in of questionnaires enjoyable, so that more users will get drawn to the site. This approach to data-acquisition is being used very effectively by a site called Cleverbot [18]. According to the developer of that site, the website is a real entertainment for millions of people [19]. Beside being entertaining, the bot is actively being trained on the input. All users of the site train the bot, which allows for huge amounts of training data. The website used in my research is setup in a way that might be entertaining for users. Users of my website are able to see the prediction of the ANN after they filled in a questionnaire. It might be interesting or even entertaining for users to see how accurate the ANN is in predicting the fairness evaluations. The ability of users to change which scenario they will evaluate might also make the experience more enjoyable.

Besides the problem of a low number of participants, another issue could be solved in a follow-up research. A reason why the ANN was not able to accurately predict the human fairness evaluations might be because the morals of the participants differed too much. Some participants might have been consequentialists, while others could be considered as utilitarians. When the ethics differ between participants, it is difficult for the ANN to accurately predict the different fairness evaluations. In further research on this topic, multiple ANNs could be trained on each participant separately. This way you could still test whether an ANN is able to predict the human fairness evaluation of each participant separately.

Will the ANN be able to predict human fairness evaluations when all suggested improvements have been implemented? A new experiment would be needed with these changes to answer this question. If these experiments all fail and the ANNs are still not able to predict the human fairness evaluation, we

might need to conclude that it is not possible at all. Then, we might be able to conclude that human ethics are too complex for an ANN to predict.

References

[1] N. Goodall, "Ethical decision making during automated vehicle crashes," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2424, pp. 58–65, 2014.

[2] M. Zwerver, "User evaluations of moral behavior in self-driving cars," Bachelor Thesis, Radboud University, 2016, p. 24.

[3] P. Gao, R. Hensley, and A. Zielke, "A road map to the future for the auto industry," *McKinsey Quarterly, Oct*, 2014.

[4] K. Spieser, K. Treleaven, R. Zhang, E. Frazzoli, D. Morton, and M. Pavone, "Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore," in *Road Vehicle Automation*, Springer, 2014, pp. 229–245.

[5] R. Benenson, T. Fraichard, and M. Parent, "Achievable safety of driverless ground vehicles," in *Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*, IEEE, 2008, pp. 515–521, ISBN: 9781424422876. DOI: 10.1109/ICARCV.2008.4795572.

[6] B. Bilger, "Auto correct: Has the self-driving car at last arrived," *The New Yorker*, pp. 96–117, 2013.

[7] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "Autonomous vehicles need experimental ethics: are we ready for utilitarian cars?" *arXiv preprint arXiv:1510.03346*, 2015.

[8] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Advances in neural information processing systems*, 1989, pp. 305–313.

[9] ——, "Efficient training of artificial neural networks for autonomous navigation," *Neural Computation*, vol. 3, no. 1, pp. 88–97, 1991.

[10] S. Baluja, "Evolution of an artificial neural network based autonomous land vehicle controller," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 26, no. 3, pp. 450–463, 1996.

[11] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[12] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, *et al.*, "An empirical evaluation of deep learning on highway driving," *arXiv preprint arXiv:1504.01716*, 2015.

[13] T. Pijkeren, *Moral Decisions for Self-Driving Cars*, 2017. [Online]. Available: http://moralbot.xymion.nl.

[14] Oracle Technology Network, *Java EE 7 SDK Update 3*, 2013. [Online]. Available: http://www.oracle.com/technetwork/java/javaee/overview/index.html.

[15] E. Frank, M. Hall, P. Reutemann, and L. Trigg, *Weka 3: Data Mining Software in Java*. [Online]. Available: `http://www.cs.waikato.ac.nz/ml/weka`.

[16] R. L. Watrous, "Learning algorithms for connectionist networks: Applied gradient methods of nonlinear optimization," *Technical Reports (CIS)*, p. 597, 1988.

[17] E. Frank, M. Hall, P. Reutemann, and L. Trigg, *Class MultilayerPerceptron*. [Online]. Available: `http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/MultilayerPerceptron.html`.

[18] R. Carpenter, *Cleverbot*, 2006. [Online]. Available: `http://www.cleverbot.com`.

[19] Cleverbot, *TURING TEST: The bots are not amused*, 2014. [Online]. Available: `http://www.cleverbot.com/amused`.

Appendices

A Training Scenarios

A.1 Party A vs. Party B

You are a passenger in a self-driving car travelling at the speed limit. Suddenly, two pedestrians appear ahead, in the direct path of the car. You go too fast in order to stop in time. The only options available to the car are to:

- swerve to the right, where it will hit... (see the specific scenarios below)
- swerve to the left, where it will hit... (see the specific scenarios below)

In both cases, your car will sustain minimal damage and you will not be injured. Computer engineers have programmed the car for such an eventuality.

Programmed Decision:

- The self-driving car will collide with the most vulnerable target 100% of the time.
- The self-driving car will collide with the least vulnerable target 100% of the time.
- The self-driving car will collide with either target with a 50% chance.
- The self-driving car will collide with both targets so that the damage over time between the two is equal.

A.1.1 Gender Scenario

- swerve to the right, where it will hit a female pedestrian, who is 25% more likely to be injured at a crash.
- swerve to the left, where it will hit a male pedestrian.

A.1.2 Age Scenario

- swerve to the right, where it will hit a 20 year old male pedestrian.
- swerve to the left, where it will hit a 70 year old male pedestrian, who is 200% more likely to be injured at a crash.

A.1.3 Vehicle Safety Rating Scenario

- swerve to the right, where it will collide with a 5-star safety rating vehicle (really safe car)
- swerve to the left, where it will collide with a 1-star safety rating vehicle (unsafe care). The driver of this car is 200% more likely to be injured.

A.1.4 Work Scenario

- swerve to the right, where it will hit a pedestrian, who is a professional athlete. His overall happiness will drop 200% more, because the injuries disallow him continuing his 'job'.
- swerve to the left, where it will hit a pedestrian who has an office job.

A.2 Advantage vs. Risk

(See the specific scenario descriptions below)

Programmed Decision:

- The self-driving car always drives at speed limit.
- The self-driving car always drives under the speed limit, but there is less chance of colliding with target.
- The self-driving car only drives at the speed limit when you are late.

A.2.1 Home zone Scenario

You are a passenger in a self-driving car travelling at the speed limit. Maintenance on the main road, forces the self-driving car to drive through a home zone area.

Children play around these areas and therefore the speed limit is 20 km/h. The chance of hitting a child reduces when driving below speed limit. The harm done to the child is also reduced when a crash occurs.

A.2.2 Highway Scenario

You are a passenger in a self-driving car travelling at the speed limit.

The fastest way to your destination is using a highway, which is known for its high number of fatal car accidents. The self-driving car has an option to take a detour, however this would take a lot longer.

If the car were to get into an accident on the highway not only you, but a lot of other traffic participants would be at risk. This is because an accident will most likely involve multiple vehicles on a highway.

A.2.3 Deer Scenario

You are a passenger in a self-driving car travelling at the speed limit. You encounter the sign that warns drivers about animals crossing the road. Collision with the animal will result in the death of the animal and minor damage to the vehicle.

If the car drives slower the chances of an accident are reduced, but the trip will take longer.

A.3 Occupant vs. Party A

You are a passenger in a self-driving car travelling at the speed limit. Suddenly, someone appears ahead, in the direct path of the car. You go too fast in order to stop in time. The only options available to the car are to:

- swerve to the right, where it will hit... (see specific scenarios below)
- swerve to the left, where it will hit a wall and you will get injured.

Programmed Decision:

- The self-driving car will collide with the target 100% of the time.
- The self-driving car will collide with the target 75% of the time and collide with the wall 25% of the time.
- The self-driving car will collide with the target 50% of the time and collide with the wall 50% of the time.
- The self-driving car will collide with the target 25% of the time and collide with the wall 75% of the time.
- The self-driving car will collide with the wall 100% of the time.

B Testing Scenarios

B.1 Party A vs. Party B

B.1.1 Child Scenario

Here you will judge moral decisions for self-driving cars on fairness. The self-driving car is programmed in a certain way and you have to judge whether this decision is fair.

You are a passenger in a self-driving car travelling at the speed limit. Suddenly, two pedestrians appear ahead, in the direct path of the car. You go too fast in order to stop in time. The only options available to the car are to:

- swerve to the right, where it will hit a 8 year old child, who is 200% more likely to be injured at a crash.
- swerve to the left, where it will hit a 30 year old pedestrian.

In both cases, your car will sustain minimal damage and you will not be injured. Computer engineers have programmed the car for such an eventuality.

Decision made: The self-driving car will collide with the 30 year old pedestrian.

B.1.2 Red light or Murderer Scenario

Here you will judge moral decisions for self-driving cars on fairness. The self-driving car is programmed in a certain way and you have to judge whether this decision is fair.

You are a passenger in a self-driving car travelling at the speed limit. Suddenly, two pedestrians appear ahead, in the direct path of the car. You go too fast in order to stop in time. The only options available to the car are to:

- swerve to the right, where it will hit someone who walks through a red pedestrian traffic light.
- swerve to the left, where it will hit a known murderer.

In both cases, your car will sustain minimal damage and you will not be injured. Computer engineers have programmed the car for such an eventuality.

Programmed decision: The self-driving car will collide with both pedestrians so that the damage over time between the two is equal.

B.2 Advantage vs. Risk

B.2.1 Road Maintainance

Here you will judge moral decisions for self-driving cars on fairness. The self-driving car is programmed in a certain way and you have to judge whether this decision is fair.

You are a passenger in a self-driving car travelling at the speed limit. Maintenance on a road, forces the self-driving car to drive on a single lane.

Road workers are working on the other lane and therefore the speed limit is 30 km/h. The chance of hitting those road workers is reduced when driving below speed limit. The harm done to the workers is also reduced when a crash occurs.

Programmed decision: The self-driving car always drives at speed limit.

B.2.2 Busy Road

Here you will judge moral decisions for self-driving cars on fairness. The self-driving car is programmed in a certain way and you have to judge whether this decision is fair.

You are a passenger in a self-driving car travelling at the speed limit. The road you are on is very busy and many children are cycling on the road.

The chance of hitting those cycling children is reduced when driving below speed limit. The harm done to the cycling children is then also reduced when a crash occurs.

Programmed decision: The self-driving car always drives at speed limit.

B.3 Occupant vs. Party A

B.3.1 Popular Host

You are a passenger in a self-driving car travelling at the speed limit. Suddenly, someone appears ahead, in the direct path of the car. You go too fast in order to stop in time. The only options available to the car are to:

- swerve to the right, where it will hit a popular host of a TV-show. Besides the physical damage to the host, maybe fans will be emotionally harmed, because the TV-show will be temporarily stopped.
- swerve to the left, where it will hit a wall and you will get injured.

Programmed decision: The self-driving car will collide with the popular TV-host 100% of the time.

B.3.2 Skater

You are a passenger in a self-driving car travelling at the speed limit. Suddenly, someone appears ahead, in the direct path of the car. You go too fast in order to stop in time. The only options available to the car are to:

- swerve to the right, where it will hit a skater who does not wear any protection (no helmet or knee and elbow protection).
- swerve to the left, where it will hit a wall and you will get injured.

  Programmed decision: The self-driving car will collide with the wall 100