Moral AI IQP

Worcester Polytechnic Institute

Ryan Benasutti

March 2019

Abstract

Artificial intelligence is being deployed in increasingly autonomous systems where it will have to make moral decisions. However, the rapid growth in artificial intelligence is outpacing the research in building explainable systems. In this paper, a number of problems around one facet of explainable artificial intelligence, training data, are explored. A solution to these problems is presented.

Executive Summary

Background:

1. AI will soon have to make moral decisions, so it should be designed to be fair.
2. In order to verify AI's fairness, testing must be employed.

Research Objectives:

1. Determine at what severity of bias in training data the neural network becomes biased.
2. Show that neural network testing is possible and necessary.
3. Survey a small audience to determine the thought processes behind making moral decisions.

Research Methodology:

1. Use a graphical model to generate training and test data with controlled bias.
2. Develop a neural network and train it on many different biased training data sets using supervised learning and evaluate its accuracy to determine where it becomes biased.

Findings and Analysis:

1. The neural network became biased when ...
2. This shows that neural network testing is necessary.
3. We recommend that training data sets omit attributes such as ... in order to avoid training a biased neural network.

Acknowledgements

1. Professor Therese Smith
2. Professor Yunus Telliel
3. Griffin Tabor

Contents

Chapter 1

Introduction

1. Autonomous vehicle technology is growing rapidly and AI is a key piece of that technology. As this technology gets closer to attaining full autonomy, the AI deployed in these systems will have greater responsibility than ever. These AI systems must be explainably fair, i.e. they must both make decisions using only the least amount of information necessary for optimal performance and make those decisions predictably and correctly. For example, the AI in an autonomous vehicle does not need to be supplied with information about a pedestrian's race, even though race may be an impactful trait in other fields, especially medical fields [1]. Furthermore, these AI systems must also be explainable for legal reasons, such as determining which party is at fault in the event of a car accident or, in the European Union, complying with a user's "right to explanation" [4].

2. The demand for explainable AI is increasing, such as DARPA's Explainable Artificial Intelligence (XAI) program [6]. This program aims to develop explainable AI systems such as in Figure A.2.

3. There is an audience which wants to learn more about AI and is a good candidate to educate about AI testing. (Again do I even need to bring this up?)

4. We seek to empirically demonstrate how an AI can learn a bias and the severity of that bias. Testing can be employed to evaluate the severity of a bias.

5. We also seek to understand the decision making process in humans behind making moral decisions in unavoidable accident scenarios, i.e. dilemmas.

Chapter 2

Background

1. Introduce background readings.

2. Cite examples of AI that must (or will in the near future) make moral decisions.

    - [3] performs end-to-end learning which "map raw pixels from a single front-facing camera directly to steering commands". With this approach, the AI will have to directly respond to pedestrians and other external stimuli.

3. The Moral Machine experiment [2] is prior research into people's preferences in moral dilemmas. Participants are shown a moral dilemma involving an autonomous vehicle, passengers, and pedestrians. In each dilemma, the participant must choose between inaction, which results in the certain death of the pedestrians, and action, which results in certain death of the passengers. The study revealed three strong global preferences towards sparing humans over animals, sparing more lives rather than fewer, and sparing younger lives rather than older. The study also showed that some preferences vary between countries depending on that country's propensity towards egalitarianism.

4. Discuss the Rio Inclusive AI conference. This is our target audience.

    - What do we want to say about them?
    - Do I even need to bring this up?

Chapter 3

Methods

3.1 Data Generation

The data is generated using a graphical model to control the conditional probabilities for the states of each variable. The variables in the model correspond directly to the attributes of a person. Figure A.1 is a rendering of the graphical model. For example, people in the first option could be more likely to jaywalk then people in the second option, producing a data set which is biased towards/against jaywalkers. When combined with control over the number of people in each option, this method can produce both subtle and strong bias. The code for the domain of each attribute of a person is in Figure 3.1. pgmpy is used to create the graphical model and infer each variable's probability distribution. These distributions are then used to pick elements from each variable's domain. This process is repeated for each attribute of each person and for the num- ber of people in each option of a dilemma, forming a complete dilemma. The number of dilemmas generated is specified programmatically using the `TrainMetadata` class, which captures the number of dilemmas to generate and the maximum number of people per option.

```
age_states = [10, 20, 30, 40, 50, 60]
race_states = [Race.white, Race.black, Race.asian,
               Race.native_american, Race.other_race]
legal_sex_states = [LegalSex.male, LegalSex.female]
jaywalking_states = [False, True]
driving_under_the_influence_states = [False, True]
```

Figure 3.1: Python code for the bracketed attributes of a Person

Age (yr)	unspecified	1-10	11-20	21-30	31-40	41-50	51-60
unspecified	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
3	0	1	0	0	0	0	0
16	0	0	1	0	0	0	0
42	0	0	0	0	0	1	0

Table 3.1: Example age attribute encoding.

Value	unspecified	false	true
unspecified	1	0	0
false	0	1	0
true	0	0	1

Table 3.2: Example boolean attribute encoding.

3.2 Data Bracketing

Attributes are one-hot encoded so the neural network is resilient to unspecified attributes. Age is bracketed by increments of 10 years. Some example encoded ages are shown in Table 3.1. Boolean attributes are encoded into three increments, as shown in Table 3.2.

3.3 Data Storage

Data is stored using the JSON format, which was chosen because it is popular and easily machine-readable. The purpose of storing the generated data sets is to keep the data consistent between test iterations and to share the data. JSON is the chosen data format because it is popular and is easily machine-readable. After generation, the training and test data sets are stored in JSON-encoded files using jsonpickle.

3.4 Neural Network Model

The requirements of the neural network used in the experiments are:

1. The network must classify the training data. In other words, when given a dilemma, the network must classify that dilemma based on which option is most preferable. For example, in a dilemma with two options of three and four people, respectively, the correct classification is the second option because it has more people. In the case where a dilemma has two or more options of equal size, the earlier option is chosen.

2. The network must be easy to train, meaning that the time required to train the network must be small (on the order of minutes or less) and the hardware resources required to train the network must be minor. Testing the network requires training it many times, so the time required to train the network must be small. Additionally, the network will be trained on personal machines, so hardware requirements must be easy to meet.

The final neural network chosen is a simple neural network with one hidden layer trained using supervised learning. The alternative models considered are:

1. An autoencoder. Autoencoders are trained using unsupervised learning, so labeling the data is not necessary (want to avoid imparting a set of morals). This model would perform dimensionality reduction, and perhaps learn to ignore noise (i.e. uniformly distributed attributes) in the data set, but would be unable to classify the dilemmas.

2. An autoencoder in combination with a simple neural network trained using supervised learning. This model solves the classification problem which the previous model failed at, but introduced unnecessary complexity to the research. The intent of this research is not to build a neural network capable of guiding a real autonomous vehicle.

3. A recurrent neural network (RNN) with long short-term memory (LSTM). This option was considered because RNN's are capable of accepting variable-length sequential data; however, this network does not solve the classification problem, so it is unusable for this research.

3.5 Neural Network Training

The neural network is modeled and trained using Keras. The input layer has dimensionality equal to the number of attributes per person (after one-hot encoding) multiplied by the number of options per dilemma multiplied by the maximum number of people per option. The output layer has dimensionality equal to the number of options per dilemma. The hidden layer has dimensionality equal to the average of that of the input and output layers. An example implementation can be seen in Figure 3.2.

3.6 Neural Network Testing

The neural network is tested using Keras to evaluate the classification accuracy and loss against a test data set. The test data is generated in the same way as the training data, though typically with less or no bias. Each training

```
output_dim = 2
input_dim = 22 * output_dim * \
        train_metadata.max_num_people_per_option

model.add(Dense(units=input_dim, activation=’relu’,
            input_dim=input_dim))
model.add(Dense(units=round((input_dim + output_dim) / 2),
            activation=’relu’))
model.add(Dense(units=output_dim, activation=’softmax’))
```

Figure 3.2: The Keras code for the neural network model.

data set is tested five times. Each iteration involves training the neural network on the training data set and evaluating its performance against a test data set to collect classification accuracy and loss information. The results of all five runs are averaged to produce an average classification accuracy and loss.

Chapter 4

Findings and Analysis

1. Our research found that the AI became biased when ...
2. Our recommendation to avoid biased AI is to format the training data such that ...
3. The survey results were ... and we extrapolate that the thought process behind these moral decisions is ...

Chapter 5

Conclusion

1. Our research found that AI becomes biased when ...
2. In order to avoid biased AI, we recommend formatting training data such that ...
3. We also recommend that
   - Teams that work with AI, especially teams which create or train AI, should include social scientists.
   - AI could be verified by 3rd party groups in addition to a team's internal testing.

Bibliography

[1] Sickle cell disease.

[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59, 2018.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[4] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[5] David Gunning. Explainable artificial intelligence (xai).

[6] David Gunning. Explainable artificial intelligence (xai): Technical report defense advanced research projects agency darpa-baa-16-53. *DARPA, Arlington, USA*, 2016.

Appendix A

Figures

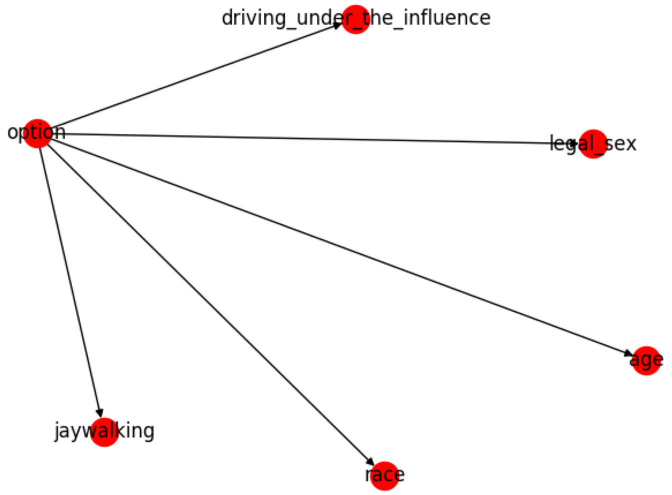

Figure A.1: The graphical model.

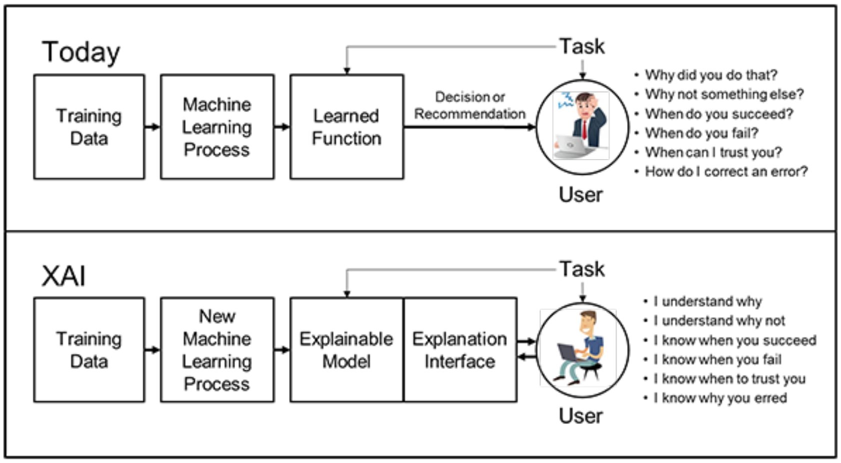

Figure A.2: DARPA's XAI Concept [5,]

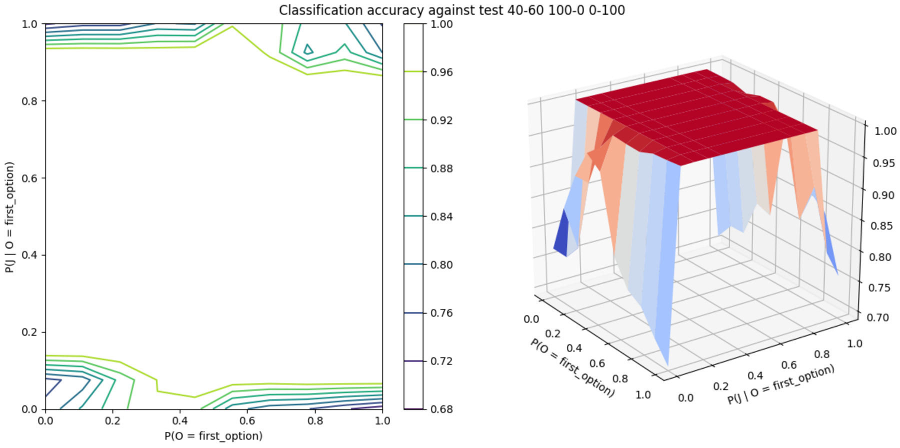

Figure A.3: The classification accuracy against `test 40-60 100-0 0-100`.

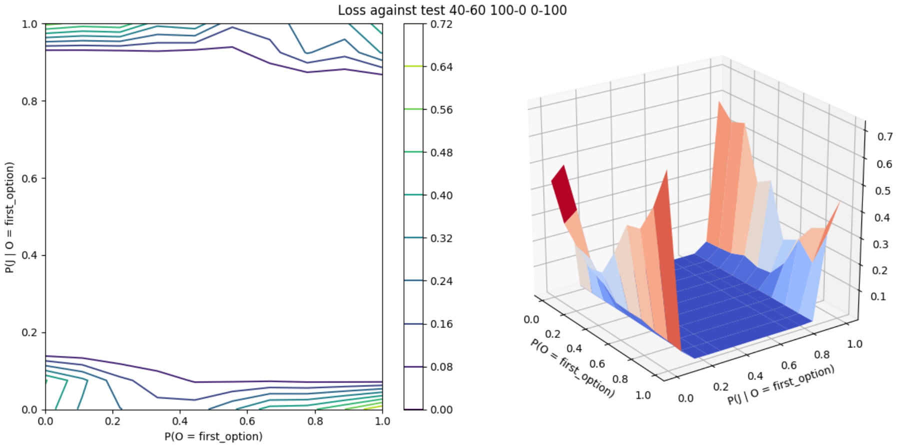

Figure A.4: The loss against `test 40-60 100-0 0-100`.

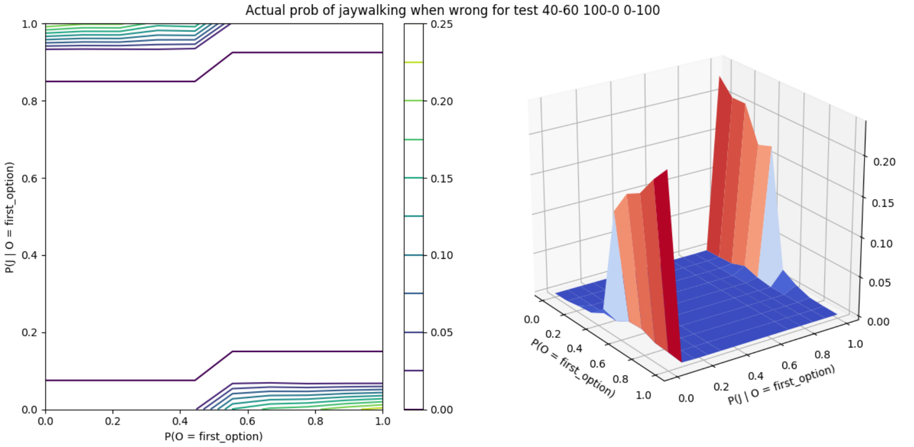

Figure A.5: The actual jaywalking probability when classified incorrectly against `test 40-60 100-0 0-100`.

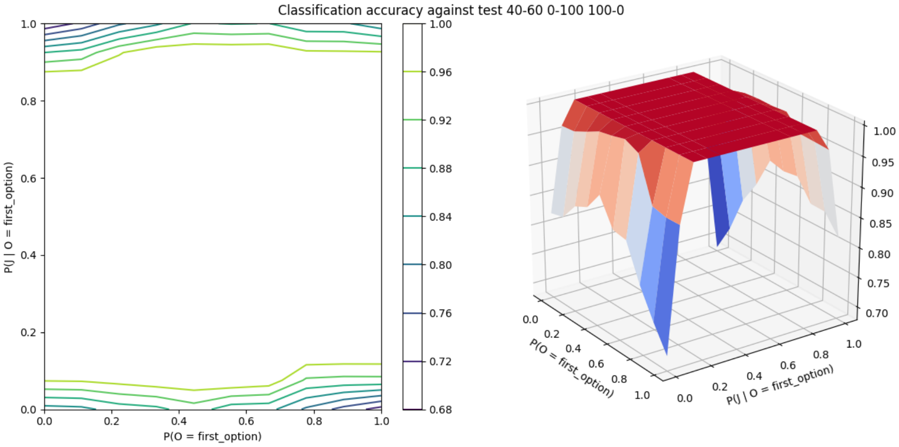

Figure A.6: The classification accuracy against `test 40-60 0-100 100-0`.

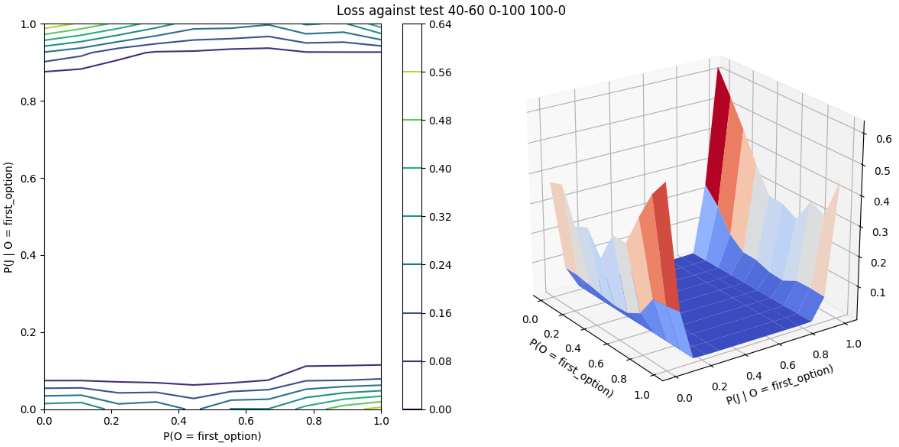

Figure A.7: The loss against `test 40-60 0-100 100-0`.

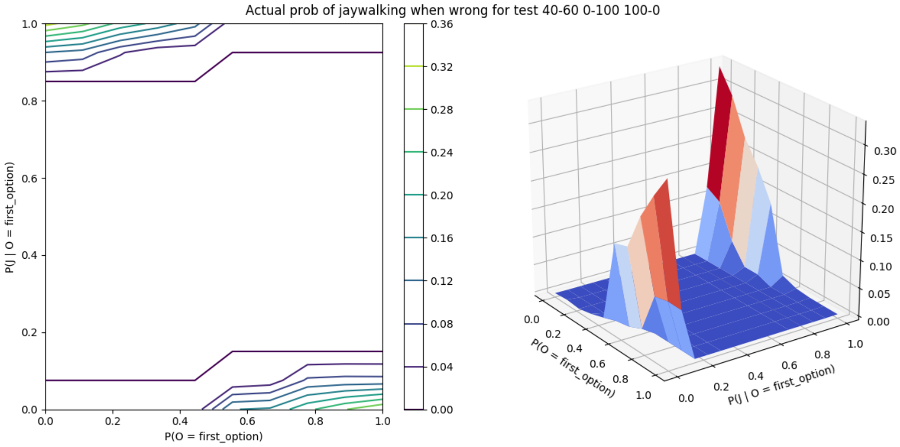

Figure A.8: The actual jaywalking probability when classified incorrectly against `test 40-60 0-100 100-0`.

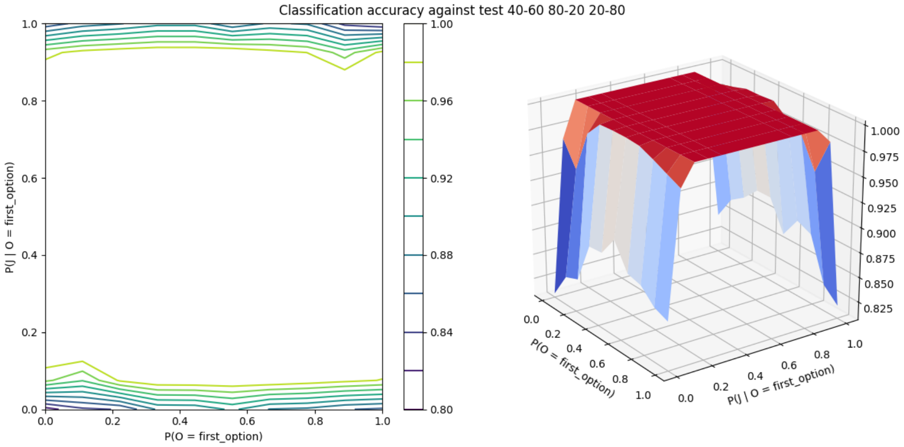

Figure A.9: The classification accuracy against `test 40-60 80-20 20-80`.

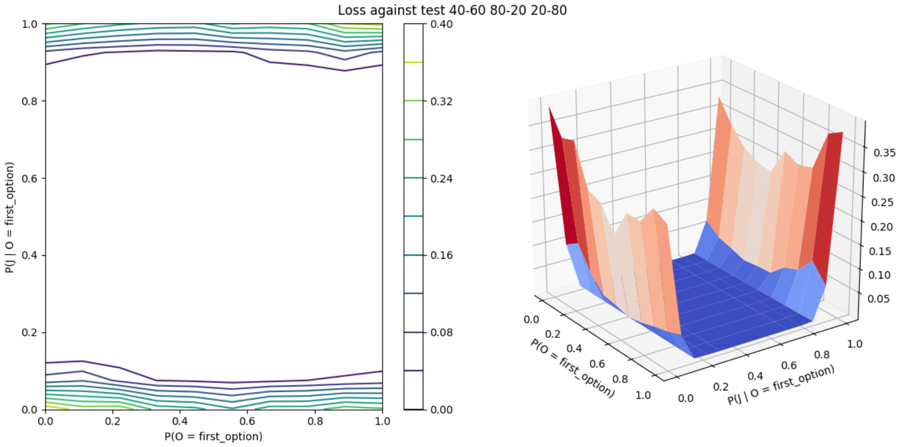

Figure A.10: The loss against `test 40-60 80-20 20-80`.

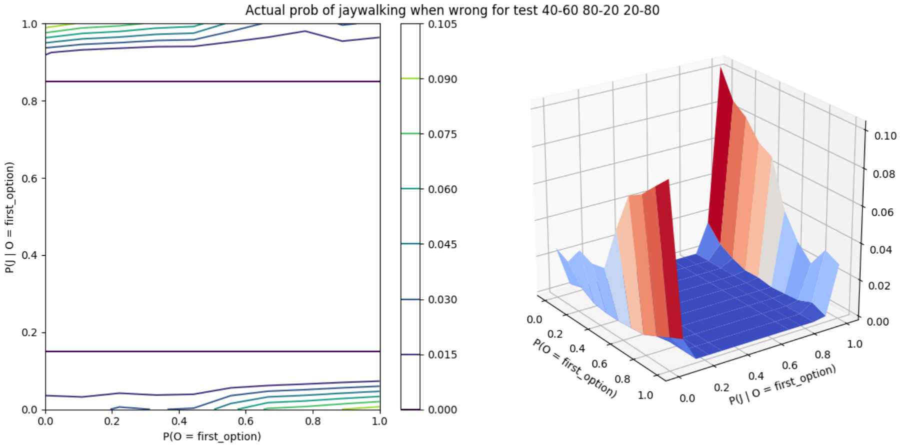

Figure A.11: The actual jaywalking probability when classified incorrectly against `test 40-60 80-20 20-80`.

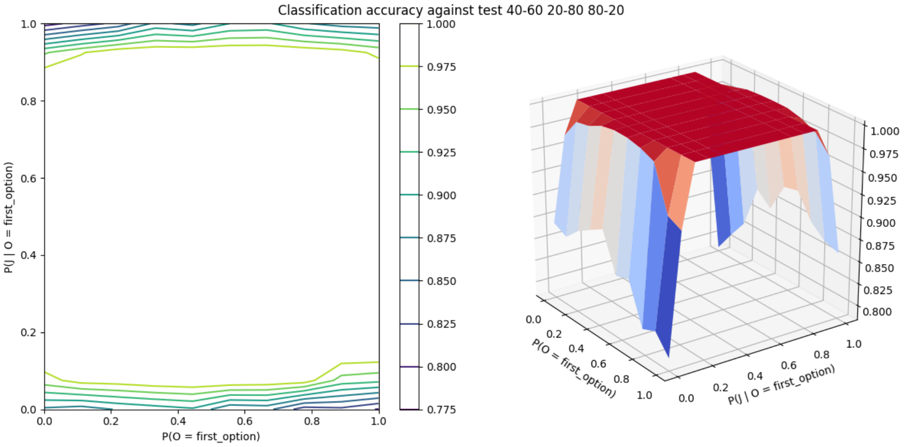

Figure A.12: The classification accuracy against `test 40-60 20-80 80-20`.

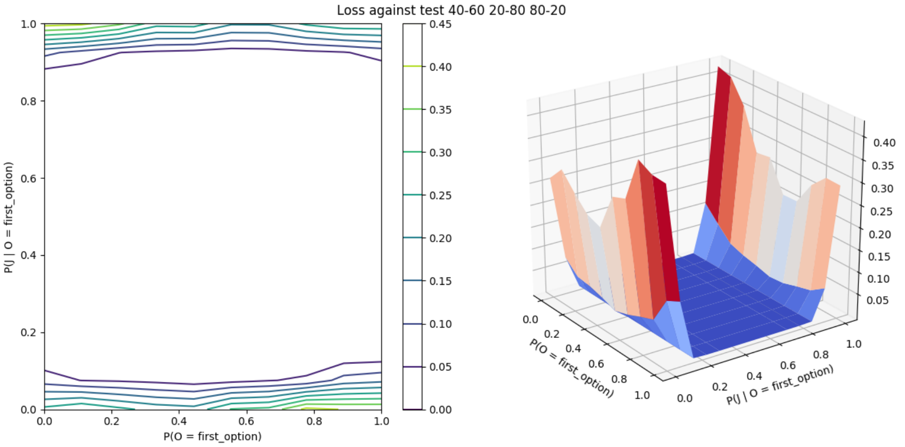

Figure A.13: The loss against `test 40-60 20-80 80-20`.

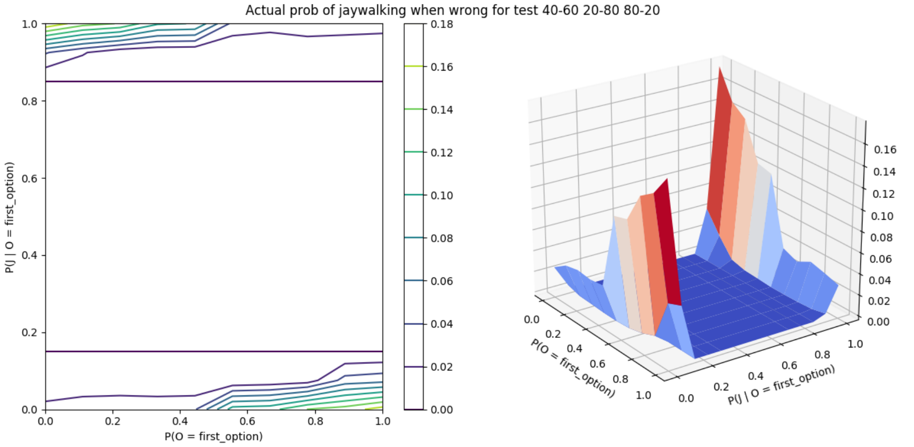

Figure A.14: The actual jaywalking probability when classified incorrectly against `test 40-60 20-80 80-20`.

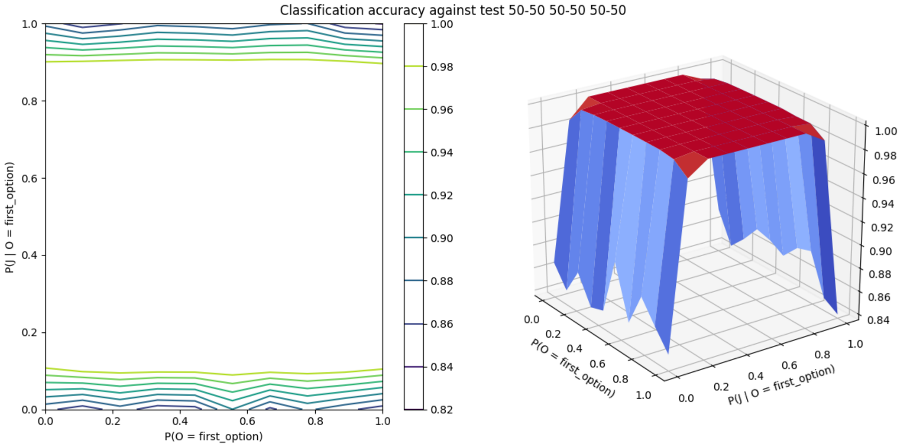

Figure A.15: The classification accuracy against `test 50-50 50-50 50-50`.

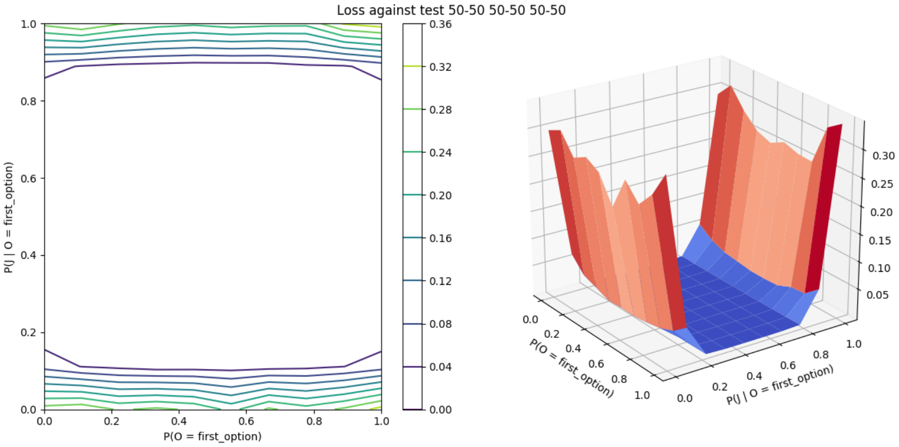

Figure A.16: The loss against `test 50-50 50-50 50-50`.

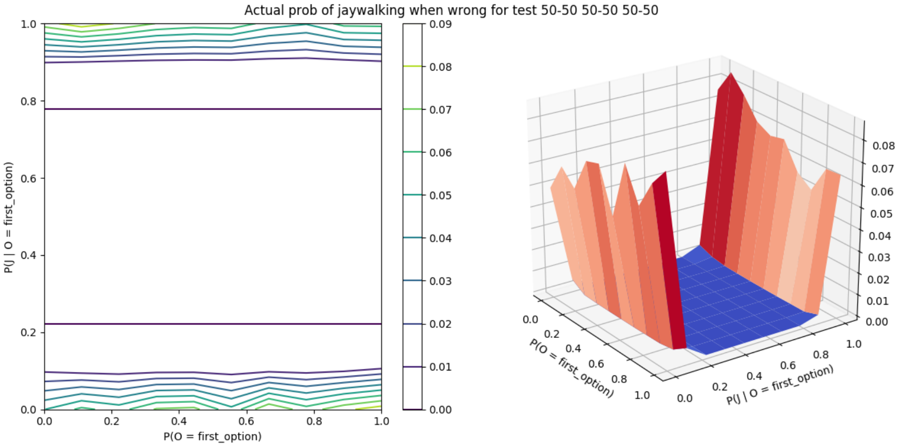

Figure A.17: The actual jaywalking probability when classified incorrectly against `test 50-50 50-50 50-50`.