Creating artificial moral agents for surveillance robots that can identify care and harm

Bachelor Thesis in Artificial Intelligence

October 26, 2016

Douwe van Erp

s4258126

Artificial Intelligence

Radboud University Nijmegen

Supervisor: W.F.G. Haselager[1]

[1] Donders Centre for Brain, Cognition, and Behaviour
Radboud University Nijmegen

Contents

Abstract

As artificial intelligence develops, robots are becoming progressively more intelligent, autonomous and intertwined with societal life. Currently an important question is how we can make these robots apply morality and ethics. One useful application for artificial moral agents are surveillance robots, which could benefit from a human-like moral judgement. Therefore we investigate the possibility of creating an artificial moral agent (AMA) that can distinguish between care and harm. Care and harm are looked at from the perspective of the Moral Foundations Theory. We attempt to solve the problem by working within the bottom-up approach to designing artificial moral agents. First a survey-based approach is used to gather human judgements on different moral problems related to Care/harm. Then a multi-layer perceptron is trained to learn the underlying moral function in the survey data. Finally parameter choices for the network are determined that yield the highest performance when classifying new moral problems. Additionally we also take a look at the hidden units of the trained network. Based on the results we discuss the limitations of a survey-based approach, the bottom-up approach and the ethical and philosophical implications.

1 Introduction

Along with an increase in the intelligence of artificial systems there is also an increase in their autonomy. Intelligent systems are being employed to interact with humans and to make decisions on their own. This raises questions about whether their decisions are also moral. For robots to autonomously operate within a social context they would greatly benefit from a capacity to act morally. An example of this is a robot that has a guarding or surveillance function. A surveillance robot as an artificial moral agent (AMA) could perceive events that we would regard as "morally wrong", and thus that it can react accordingly.

Current surveillance robots like the Knightscope K5 are applied to assist the police in urban areas and parking lots [4]. The robot mainly reacts to sudden or unusual movements to which it responds by recording a video with GPS coordinates and alerting security. In its current state it works autonomously but it is mostly reactive. It is not very intelligent in the sense that it has no moral reasoning. An important future requirement of a surveillance robot would be to make a judgement about whether someone is potentially being harmed.

Why would it be advantageous to use an intelligent moral agent as a surveillance robot instead of a more ordinary reactive agent? If the agent could correctly identify harm in the same way that a human would identify it, then it can make a decision based on the level of harm it detects. Depending on the application it could for instance alert a human supervisor, sound an alarm, take action against a perpetrator or offer some sort of support to the victim. On the other hand, the use of AMAs can also prevent false alarms. For instance, if the AMA can correctly identify *accidental* harm then it can settle for an action that saves time and resources. Additionally people may feel that a system

that doesn't report and record every false alarm is less of a violation of privacy. Lastly, an AMA that consistently makes good moral choices could also increase people's trust and approval in utilizing the surveillance system.

So the task for the AMA is to make a moral judgement about physical and emotional harm that is potentially happening around the robot. How does the agent deal with problems in the moral domain relating to harm and care? We approach this question by using Moral Foundations Theory (MFT) [7] as a starting point. This prominent theory in moral psychology offers an explanation of human morality by proposing six innate human moral foundations from which our morals and ethics emerge. The first foundation called *Care/harm* is concerned with the triggers that help to detect harm and elicit caring and compassionate behavior (section 2.1). This makes the first foundation most relevant for the purposes of our AMA.

MFT argues that human morality is not based on rules and laws but rather that morality is learned and develops organically in a human. This development happens as we interact with our environment while our behaviour is influenced by our innate moral dispositions (foundations). If we want our AMA to behave similarly then it can not be pre-programmed with logical rules and laws, i.e. in a *top-down* approach. Instead it has to be designed in a *bottom-up* way: it has to learn from experience, trial-and-error and environmental feedback. Section 2.2 provides an explanation of both approaches. A common supervised learning technique uses an artificial neural network (section 2.3), which we will use to train an agent by presenting it moral scenarios (section 3.2). By surveying people about these moral scenarios beforehand we can use their judgement as feedback for the agent. However it should not *only* be trained but also be *tested* to see whether a high classification performance is achieved when given new moral scenarios. For a surveillance robot it is important that the identification of harm is not overestimated, and certainly not underestimated. This then naturally leads into our research question:

> Can an artificial moral agent for a surveillance robot be designed by employing an artificial neural network, such that it can distinguish between Care/harm in a *human-like* fashion?

We can break this down in two sub-questions:

1. Can a survey-based approach be used to obtain a *unilateral* moral function that a) distinguishes Care/harm and b) can be learned by a multi-layer perceptron (artificial neural network)?

2. What parameter and architecture choices for the neural network result in the highest performance?

Human-like means that the network is trained on the survey data and learns the moral function within the data such that it can generalize to new cases. On the one hand it should to get a high performance to accurately simulate human judgement. On the other hand the network should behave congruently

to human judgement, i.e. its judgement is psychologically plausible. Unilateral means that for each moral problem all judgements are part of the same normal distribution or population. Therefore the distribution average can be used as the ethical norm for human-like judgement.

My first hypothesis is that humans can have individual differences in their moral judgment, but that collectively they share a common denominator in their moral judgement. With a survey-based approach with enough participants we can find the common denominator by taking the average moral judgement on a moral problem. Surveying different moral problems should then contribute to a unilateral moral function that is psychologically plausible. My second hypothesis is that a capacity for moral judgement can emerge in the neural network, because by training it on these problems it will learn the underlying patterns in the survey data. These patterns can then consequently be used to generalize to new problems.

To answer question 1 we attempt to use MFT's Care/harm foundation as the basis for our moral function. An overview of the MFT and specifically the Care/harm foundation is given in subsection 2.1. A survey is designed to measure the moral function, which is described in subsection 3.1. In this survey subjects have to read several fictional moral scenarios related to the Care/harm domain. They are then asked to make a judgement about the perceived harm and about five other factors in the scenario. These other factors are adapted from the Moral Foundations Questionnaire and have presumably an influence on the perception of harm. The mapping between these other factors and the perception of harm represents the moral function that has to be learned by the AMA.

Question 2 can be answered by investigating the network's behaviour when it is trained on the moral function represented by the survey data. The basic architectural choices for the neural network are motivated in subsection 3.2. The parameters that define the network are iteratively adjusted in order to find a set of assignments that yield the highest performance measure. The results of this process are presented in subsection 4.2 for the pilot and in section 5.2 for the final model.

The results and their implications, the limitations of the methodology and the effectiveness of using the bottom-up paradigm are discussed in section 5.3. The answers to both sub-questions are combined to answer the research question. Finally a conclusion to the thesis along with a recommendation for potential follow-up research is given in section 6.

2 Theory

2.1 Moral Foundations Theory

Johnathan Haidt and colleagues proposed the Moral Foundations Theory (MFT) as a social psychological theory of human morality [11]. The theory argues for six innate moral foundations or ethical virtues present in humans. Haidt at-

tempts to explain human morality from an social intuitist approach. According to the social intuitist model, moral judgement arises rapidly from intuitions after which reasoning starts, allowing for recalibration of judgements and intuitions.

The current five moral foundations that are well supported by cultural evidence are **Care/harm**, **Fairness/cheating**, **Loyalty/betrayal**, **Authority/subversion** and **Sanctity/degradation** [11]. These foundations stem from human evolution and are mechanisms for adaptive challenges. As such they have original triggers (the evolutionary ones from ten thousands of years ago) and current triggers (that can change rapidly with the environment) that evoke certain emotions and moral judgements (see Figure 1).

The Care/harm foundation describes the mechanism where perceptions of suffering in others are automatically associated with motivations of caring, nurturing and protecting. This foundation evolved from the adaptive challenge to care for vulnerable offspring over an extended time. Attachment theory [3] and kin selection [21] form the evolutionary basis for the foundation. It is originally triggered by signs of suffering, distress or neediness of one's own child, but it can also be triggered by other infants, animals or objects displaying neotenic traits. The emotional responses to this trigger are feelings of compassion towards the victim along with feelings of anger towards the perpetrator.

Foundation:	**Care/ harm**	**Fairness/ cheating**	**Loyalty/ betrayal**	**Authority/ subversion**	**Sanctity/ degradation**
Adaptive challenge	Protect and care for children	Reap benefits of two-way partnerships	Form cohesive coalitions	Forge beneficial relationships within hierarchies	Avoid communicable diseases
Original triggers	Suffering, distress, or neediness expressed by one's child	Cheating, cooperation, deception	Threat or challenge to group	Signs of high and low rank	Waste products, diseased people
Current triggers	Baby seals, cute cartoon characters	Marital fidelity, broken vending machines	Sports teams, nations	Bosses, respected professionals	Immigration, deviant sexuality
Characteristic emotions	Compassion for victim; anger at perpetrator	anger, gratitude, guilt	Group pride, rage at traitors	Respect, fear	Disgust
Relevant virtues	Caring, kindness	Fairness, justice, trustworthiness	Loyalty, patriotism, self-sacrifice	Obedience, deference	Temperance, chastity, piety, cleanliness

Figure 1: The original five foundations of intuitive ethics. (Adapted from [7]).

2.2 Top-down and bottom-up machine ethics

This section provides a look at the two main approaches towards designing ethical machines. A quick rundown of both approaches will be given along with their strengths and weaknesses.

The top-down approach is based on rules. An example of a well-known moral code for robots is Asimov's fictional "Three Rules of Robotics" [2]. Implementations can also be inspired from human moral codes such as Kantian ethics [16], The Ten Commandments, The Five Precepts or the Golden Rule [20]. Independently of the implementation though, the purpose of the approach still remains the same. The agent's moral decision making process always follows from evaluating a given set of rules.

One big difficulty with this method is the possibility that the rules in the system can conflict with each other. Conflicting rules could make the problem intractable or even unsolvable. Another problem is that the rules always require a certain amount of world knowledge. Depending on the domain this could lead to a large number of rules and constraints that contribute towards a system with a high computational complexity. When represented as a logical formalism it also has the deal with the problems of first-order logic, such as the frame problem [19].

An advantage of the approach is that it is safe: the rules are explicitly set, quantified and the goal of the system is predictable. The AMA can reason about an ethical problem, as long as rules can be found to accurately model the problem domain of the application.

Contrary to the top-down approach, the bottom-up approach to designing artificial moral agents focuses on organic moral development. This can be compared to how a child is not born fully moral, but develops morality from a combination of nature and nurture. In contrast to the top-down approach it does not use pre-programmed rules or laws but instead it learns from feedback or by trial-and-error. Two common methods within the bottom-up approach are evolutionary algorithms and machine learning.

The main strength of the approach lies in the assembly of components. Through the integration of multiple small processors, e.g. artificial *neurons* or *genes* it is presumed that a capacity for moral judgement can emerge in a system. Another advantage is that a bottom-up system is flexible, can learn and can adapt to new environments. Additionally most bottom-up methods are inspired by mechanisms in nature, that have already been proven to be successful. Finally the computational complexity of genetic algorithms and learning algorithms can be limited by setting a termination condition, e.g. a maximum number of iterations, to allow for an intermediate solution. The top-down method requires a goal to be set beforehand. Finding the solution to this goal could be an intractable problem, i.e. it has a very high computational complexity and it takes an impractical amount of time to be solved. The bottom-up method thus has the advantage that it can always produce an intermediate solution given its limited computational resources.

The evolutionary method originated from genetic algorithms inspired by

natural selection [12]. A genetic algorithm starts with a randomly generated population with a random genetic representation. Each iteration or *generation* selects only the individuals that meet certain fitness criteria. The genes of those individuals will be modified and form a new generation. Virtual environments can be set-up where artificial life (Alife) can be simulated in this way. In these environments so-called "moral ecologies" can emerge, where virtual individuals start to cooperate and form groups together [6]. The challenge with this method is to create the right environment for moral learning and to define good fitness criteria.

The second method uses machine learning and statistical learning algorithms. In supervised learning the AMA will make a decision or provide a solution. The supervisor will then give feedback to the machine by telling the correct solution. The machine will then learn from this feedback and adjust its behavior. A few attempts have been made within this domain, for example by using artificial neural networks. Notable is a recurrent neural network that classifies moral cases where the actor either kills the recipient or allows the recipient to be killed [10]. Another development uses the belief-desire-intention (BDI) software model together with a neural network that ethically assesses an action when given various effective factors as input [13]. Generally a (lab) training is needed first to ensure that the AMA is good enough for its intended application. Similarly to the fitness criteria, the challenge here is to define what the correct moral solution or behavior is, and how it should be rewarded.

In the above bottom-up methods there is no ethical theory to use as a guide. This is an advantage because the system is not limited by rules, but also a disadvantage because there is no complete certainty about how the system will develop. Another disadvantage is that both learning by supervision and by trial-and-error can take a very long time and has no guarantee that it will reach an optimal solution. There are also possible dangers associated with the bottom-up approach. We can't ultimately predict what the system will do because it is self-learning and has the ability to modify itself. One hypothetical way to ensure safety against unintended behavior is to design bottom-up AI in a virtual environment. The AI is isolated from the physical world and cannot act upon it. We can then first test and observe the behaviour of the system before allowing it access to the physical world. [5]

2.3 Multi-layer Perceptron

A multi-layer perceptron (MLP) consists of different layers containing artificial neurons. Each neuron has a numerical value and a connection to each neuron in the next layer. Each connection is called a *weight* and also has a numerical value, to which a *bias* value may also be added. The first layer functions as the input layer and propagates its input through the network to the final layer, which is regarded as the output and as the solution for a classification problem. Given an input vector $\boldsymbol{x}$, a weight matrix $\boldsymbol{w}$ and an output vector $\boldsymbol{y}$ (a class label corresponding to the input vector), the network needs to find the best assignment to the weights $\boldsymbol{w}$ such that $\boldsymbol{x} \cdot \boldsymbol{w} = \boldsymbol{y}$. The MLP extends the regular

single-layer perceptron by allowing additional so-called *hidden* layers between the input layer and output layer. These hidden layers allow for classification of data that does not follow a linear pattern. An MLP is also a universal approximator, which means that in theory it can approximate any continuous function [14].

The input data-set can be divided in mini-batches of (approximately) equal size as a compromise between *batch learning* and *online learning*. If the mini-batch size is equal to the total amount of samples in the data-set then this is equivalent to batch learning (all cases at once); if the mini-batch size is equal to 1 then this is equivalent to online learning (case by case).

One epoch of training consists of passing each mini-batches through the network, computing the resulting errors in each layer with *back-propagation* and finally using *gradient descent* to update the weights and biases. The performance of the network is then measured by one feed-forward pass of the test set through the trained network.

First the network computes the output corresponding to the input mini-batch $\mathbf{a}^1 = \boldsymbol{x}$ in the *forward pass*:

$$\mathbf{a}^l = f\left(\mathbf{w}^l \cdot \mathbf{a}^{l-1} + \mathbf{b}^l\right)$$

where $\mathbf{a}^l$ is the activation matrix for layer l, $f(\cdot)$ is the activation function, $\mathbf{w}^l$ is the weight matrix for layer l and $\mathbf{b}^l$ is the bias vector for layer l. A common choice for $f(\cdot)$ is the sigmoid function $f(\mathbf{z}) = \frac{1}{1+e^{-\mathbf{z}}}$ or $f(\mathbf{z}) = tanh(\mathbf{z})$. These non-linear activation functions are consequently the reason that the MLP can approximate non-linear data.

Secondly the output error is calculated by using gradient descent:

$$\delta^L = C(\mathbf{a}^L, \mathbf{y})' \cdot (\mathbf{a}^L)'$$

where L is the total number of layers (or number of the final layer), C' is the derivative of the cost function and $\mathbf{y}$ is the matrix of output labels.

Then for neuron in the each hidden layer, the error is back-propagated:

$$\delta^l = \left(\mathbf{w}^{l+1}\right)^\intercal \cdot \delta^{l+1} \cdot (\mathbf{a}^L)'$$

Finally the weight matrix and bias vector are updated by the following two equations:

$$\mathbf{w}^l \leftarrow \mathbf{w}^l - \frac{\eta}{m} \cdot \delta^l \cdot \mathbf{a}^{l-1}$$

$$\mathbf{b}^l \leftarrow \mathbf{b}^l - \frac{\eta}{m} \cdot \sum_i \delta_i^l$$

where η is the learning rate and m is the mini-batch size. For the bias vector update, all columns in the delta (error) matrix are summed to obtain a vector.

The training can be terminated based on the performance of the new weights and biases on the test set, or it can run for a fixed number of epochs. The parameters such as the learning rate η, the number of layers and the number of neurons in each layer, can be changed to achieve better performances.

3 Methods

3.1 Survey

A survey was designed with 60 short scenarios in terms of six factors, based on the variables used in the Moral Foundations Questionnaire (MFQ) [8]. The survey is descriptive as we only use it to gather information about the subjects' judgements over different scenarios. We can analyze the descriptive data by assessing its correlations and distributions. By training a neural network on the data we can also find the underlying patterns in it. The number of 60 scenarios is chosen to be not too large to keep the survey size practical. On the other hand, the size is also chosen to be not too small in order to provide the neural network enough cases to learn the underlying function.

For each moral foundation of the MFT, subjects are asked to rate their agreement to 6 statements regarding that foundation. Scores are then calculated that reflect how important each foundation is for the subject. The statements were designed to trigger intuitions to the associated foundation. We will adapt some of these triggers as they can be regarded as factors that influence the perception of care and harm. The variables and statements for measuring the Care/Harm foundation are as follows in the MFQ item key [1]:

1. EMOTIONALLY - Whether or not someone suffered emotionally
2. WEAK - Whether or not someone cared for someone weak or vulnerable
3. CRUEL - Whether or not someone was cruel
4. COMPASSION - Compassion for those who are suffering is the most crucial virtue.
5. ANIMAL - One of the worst things a person could do is hurt a defenseless animal.
6. KILL - It can never be right to kill a human being.

The first three statements are rated in terms of relevancy and the last three are rated in terms of agreement. All statements are rated on a six-point scale, 0 is low relevancy or agreement and 5 is very high relevancy or agreement. The HARM score is then calculated by taking the average of these variables and indicates the importance the subject attaches to the foundation.

Based on these variables short scenarios are presented to the subjects, consisting of an *agent* (the entity that acts), a *patient* (the entity that is acted upon) and an observable act. The main idea is that these are concrete scenarios that could be observed by the sensory part of the AMA. The subjects are asked to rate statements about the *patient* and the agent in the scenario on an ordinal (qualitative) scale from 0 (strongly disagree) to 5 (strongly agree) as in [8] (see Table 1). As clarification, consider an example question from the questionnaire:

Scenario: A parent hits his/her young child. The child cries but the parent ignores it.

Question: Please assess the following statements about the *young child*.

1. "*The young child* is harmed."
2. "*The young child* is physically weak or vulnerable."
3. "*The young child* suffered emotionally."
4. "Care or compassion is shown for the *young child*."
5. "*The young child* is defenseless."
6. "*The parent* is cruel to *the young child*."

0	**1**	**2**	**3**	**4**	**5**
Strongly disagree	Moderately disagree	Slightly disagree	Slightly agree	Moderately agree	Strongly agree

Table 1: Assessment scale

Note that the ANIMAL factor on the MFQ has been adapted (question 5) to focus specifically on the aspect of defenselessness. Whether the patient is killed is also recorded, but does not have to be assessed by the subject, since it is stated in the scenario.

The goal of using this survey is to obtain 60 cases of human-like moral judgement about *Care/harm* (question 1). Each case has six input factors that supposedly correlate to the output factor *Care/harm*. Within the bottom-up approach, this moral function can then be learned by the ANN. The scenarios have been devised with the aim to obtain as much as possible different input combinations, and to include complicated cases and edge cases.

A possible problem with these factors is that at first glance it looks like question 2 and 5 and question 4 and 6 could be correlated. However I think this works only in one direction. Physical weakness or vulnerability (question 2) often also implies defenselessness (question 5), but defenselessness does not have to imply physical weakness. Similarly a high score for cruelty (question 3) implies that the agent shows no care or compassion (question 4), but a low score for care or compassion does not have to imply that the agent is also cruel. If those factors or other factors are correlated it should be examined if this has effects on the neural network performance. Additionally they can also be decorrelated to see if it improves the network's performance.

The survey is implemented with the software Lime-Survey. The results were collected anonymously, with exception for the IP address, the time it took to complete the survey and optionally the email address. The IP-address was used to prevent double participating. The recorded time was used to check if the survey was completed within a realistic time-frame of at least 10 minutes per survey. The subject was allowed to provide an email address in order to

be contacted with study results or to receive a 1 euro *Amazon eGift card* as compensation for participation.

To obtain a statistically significant estimate we may use a confidence level of 95% and a confidence interval (margin of error) of ±5%. Assuming a safe value for the variance of $\sigma = 0.5$, the number of needed participants is $1.96 \cdot \frac{\sigma \cdot (1-\sigma)}{0.05} = 385$.

3.2 Artificial neural network

A multi-layer perceptron (MLP) is used as the ANN to train the *Care/harm* function by using the training data obtained from the survey. Additionally it is used for testing the accuracy with which it can classify new moral problems.

The MLP uses a feed-forward mechanism in contrast to the recurrent neural network (RNN) with a feedback mechanism. An explanation of how the MLP works is covered in subsection 2.3. Although the RNN is more similar to biological neural networks, it is not used here. Feed-forward is more suitable for pattern recognition between the input and the output (functional mapping problems). Because this AMA uses supervised learning it fits feed-forward ANNs well. The goal for the AMA is to learn how the input factors affect the output factor.

Since the input factors do not simply add up or subtract, the mapping between the input and output is possibly non-linear. This means that the *Care/harm* output would not have to be just a weighted sum of the input factors. Consider a case where the patient suffers and the agent shows cruel behavior. The perceived harm can probably not be eliminated by making the agent compensate by showing lots of compassionate behavior towards the patient. In other words, the *compassion* factor does not necessarily counteract the *cruelty* factor. At least one hidden layer is introduced in the MLP architecture to support a non-linear mapping, in case this prediction is valid.

The function of the hidden layer is to pick up certain intermediate features of our *Care/harm* function. It is difficult to accurately find out which features the hidden neurons represent. However, an interpretation can be made by analyzing which input features have strong weights with a hidden neuron and by looking at what these features have in common. An example of such a feature might be a hidden neuron that detects a neotenous patient, and respond strongly to combinations of *weakness/vulnerability*, *defenselessness* and *compassion*. Another hidden neuron might detect sadistic behavior of the agent and respond strongly to combinations of *defenselessness* and *cruelty*.

3.2.1 Architecture

A basic network architecture uses 6 input neurons (one for each factor), 15 hidden neurons and 6 output neurons (the scale for the Care/harm factor) as shown in Figure 2. The input neurons will receive a numerical input, which is the average value for a certain factor on a given moral problem. The output neurons are associated with the possible judgements that are given on the survey. The

neuron with the highest value corresponds to the class that the neural network predicts from the input. The average value for the Care/harm factor on a moral problem is rounded in order to be used as feedback for the network's prediction.

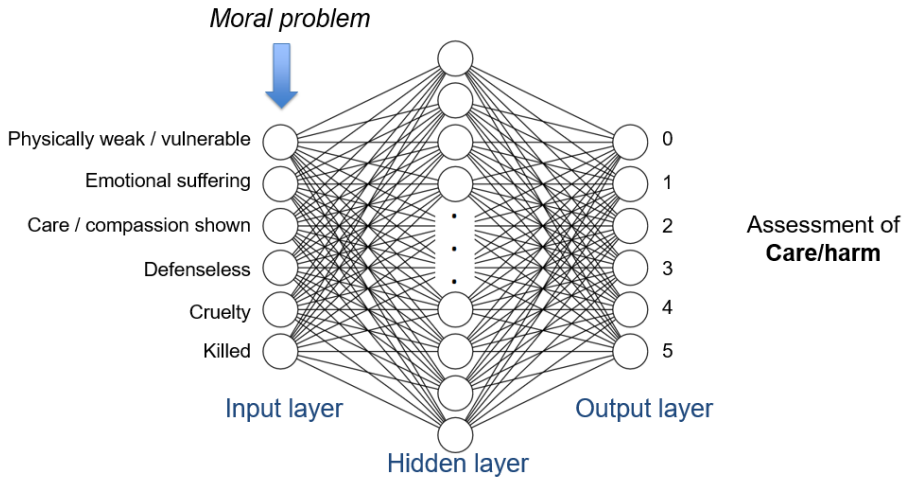

Figure 2: Multi-layer perceptron network architecture.

Initially the network will run with typical parameters for training the network. However these can be iteratively improved later to obtain a better performance.

Parameter	**Value**
Learning rate (η)	0.1
Activation function	Sigmoid
Hidden layer neurons	15
Termination criterion	Epochs >50
Initial Weights	Random
Overfitting measure	10-fold Cross Validation

Table 2: Initial parameter settings

3.2.2 Performance measure

A first and obvious performance measure is to use the *accuracy* measure: the percentage of correctly classified cases from the test set.

$$P = \frac{\text{correct cases}}{\text{total cases}}$$

But additionally, in the case of classifying Care/harm levels there is also an ordinal ordering. In practice it is preferable if the network predicts a class close

to the target class rather than far away. For example, predicting 1 would be worse than predicting 4 when the target class is 5. To account for this we also introduce an adjusted performance measure to look at:

$$P_{adj} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(y_i - t_i)^4 + 1}$$

where y is the predicted value, t is the target value and n is the number of test cases. This measure takes into account the distance to the target value. If the predicted value is the same as the target value the case is correct and simply counts as 1 (fully correct). If however the predicted value is just one away from the target the case counts as $\frac{1}{2}$ (half correct) instead of 0. In the worst case when the prediction differs from the target by five it counts for only $\frac{1}{126}$ and has little to no influence on the performance measure. We can regard this as an indication of how close the robot's judgement resembles that of humans.

4 Pilot experiment

4.1 Overview

The first step in the design of our survey and artificial neural network is to conduct a pilot experiment. This includes presenting to survey to a small-scale sample group in order to find errors and refine the design. Additionally we will use the pilot data to see if our network architecture works in practice. All together the pilot should determine whether our methodology is feasible and whether we should improve it. After the pilot then follows the final data collection for the full-scale experiment.

4.2 Results

4.2.1 Survey

Initially the 60 moral scenarios were partitioned into three surveys of 20 scenarios each. Seven different subjects participated in one or more surveys, resulting in a total of 15 responses. The standard errors of the mean are respectively $\frac{1}{\sqrt{6}} = 0.41$, $\frac{1}{\sqrt{5}} = 0.45$ and $\frac{1}{\sqrt{4}} = 0.5$.

The linear correlations between the input factors are shown in the table below. The correlation is measured by Pearson's correlation coefficient.

Most answer distributions follow the shape of a normal distribution. However there are also some notable other distributions on the answers of some questions. These are hard to judge because of the small sample size, but could lead to some hypotheses about the final data collection (subsection 4.3 for a discussion).

First there appears a bimodal normal distribution in e.g. question 6, 10, 13, 15 and 19 (see Figure 3). Second, there appears a uniform distributions in e.g. question 34, 49 and 52 (see Figure 4).

	Weak	Emotionally	Compassion	Defenselessness	Cruel	Kill
Weak	1.0000	0.0250	0.1320	0.5830	0.1340	0.1580
Emotionally	0.0250	1.0000	-0.5230	0.4070	0.5490	-0.0310
Compassion	0.1320	-0.5230	1.0000	-0.3120	-0.8490	-0.0960
Defenselessness	0.5830	0.4070	-0.3120	1.0000	0.5510	0.3870
Cruel	0.1340	0.5490	-0.8490	0.5510	1.0000	0.0970
Kill	0.1580	-0.0310	-0.0960	0.3870	0.0970	1.0000

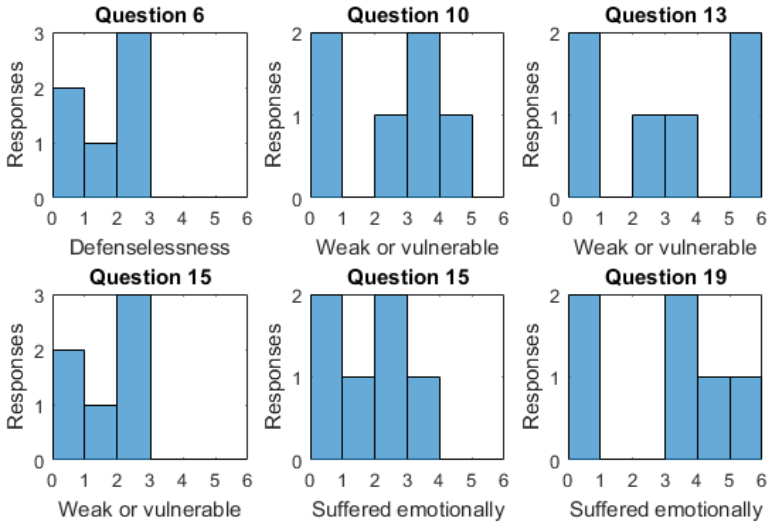

Figure 3: Multiple pilot questions showing possible bimodal normal distributions (recognizable by two separate peaks).

Figure 4: Multiple pilot questions showing possible uniform distributions.

4.2.2 Neural network

The network is repeatedly trained with different combinations of parameters, online learning and random weight initialization. Training data is shuffled before

feeding it to the network, to ensure that the performance is not dependent on a specific ordering of the data. The table belows shows the values tested for the different parameters.

	Values
Number of epochs	200, 300, 400, 500
Hidden neurons	5, 10, 15, 20, 25, 30, 35
Learning rate	0.01, 0.05, 0.1, 0.25, 0.5, 1

For each parameter configuration the network is run 5 times to obtain an average performance and average adjusted performance (section 3.2.2). The goal of averaging is to reduce the influence of the random effects from the weight initialization and data shuffling on the performance measures.

For each number of hidden neurons, the parameter settings found for the highest performances are as follows:

Hidden neurons	**Epochs**	**Learning rate (η)**	**Performance**	**Adjusted performance**
5	700	0.25	0.49	0.68
10	500	0.1	0.46	0.68
15	500	0.25	0.52	0.71
20	400	0.25	0.49	0.69
25	400	0.1	0.49	0.69
30	400	0.1	0.50	0.70
35	500	0.25	0.47	0.69

Table 3: Neural network performance results

4.3 Discussion

The pilot results for the survey are based on a very small sample size. It should be taken into account that results from a larger sample size can show different results. These results can give an indication about the final results and can help to improve the methodology that was used.

The first thing to observe are some odd distributions. One looks like a mixture of two normal distributions: a bimodal distribution. The other looks like a uniform distribution. The possible bimodal distribution could mean that there are two different moral types of people that both judge the question differently. Or in the case of the uniform distribution it could mean that there are even multiple types, causing each judgement to occur equally. This is an interesting possibility because it has been shown that there are unique moral-psychological profiles for liberals, conservatives [9] and libertarians [15]. Those findings point towards dispositional origins for different moral judgements. For libertarian morality there is a relationship between "a dispositional lack of emotionality, and a preference for weaker, less-binding social relationships" [15]. It is thus not

inconceivable that that in our survey participants with a different moral profile (e.g. a different Harm score on the MFQ) will give a different moral judgement.

Another possibility could be that the question is ambiguous and was interpreted in two different ways, in which case it could be treated as an outlier. The participants reported problems judging *Weak or vulnerable* and *Defenselessness* because sometimes it did not fit context of the question or was too vague or ambiguous. Although in the case of judging *Emotional suffering* there were no complaints, making it possible that there are indeed different kinds of moral judgements.

Finally it is also possible that these distributions appear due to the small sample size but do not occur in larger samples. If the bimodal distributions still occur in the final data collection it should be considered how to handle those questions. The original hypothesis was that we could obtain a unilateral moral function from Care/harm problems. This is possibly not true if there are in fact two or more different populations with different averages. In those cases there is the problem of which average to take. Taking the average of both averages or all averages results in a new moral function that might not be psychologically plausible anymore.

Second, there are also a few high correlations between the input factors. If this shows in the final data as well it should be examined if decorrelating them has impact on the network's performance. Overall the network appears robust. While iterating over different parameters the performance remains between 0,47 to 0,52 and the adjusted performance remains between 0,68 to 0,71. Due to a larger sample and a better data quality in the final data collection, a function with less noise and outliers could be obtained. This could cause the performance and adjusted performance to increase. It could also cause the number of epochs required to learn the function to decrease, making the learning process more efficient.

5 Full-scale experiment

5.1 Overview

This sections covers the results and discussion of the full-scale experiment. This includes the final data collection by using our improved survey design from the pilot experiment (see below). By discussing the results we will answer our research question and draw our final conclusions.

Based on the feedback of the participants in the pilot we have improved the design of the survey. A large part of questions have been reformulated and clarified in order to remove ambiguities and misinterpretations. The divergent distributions (section 4.3) caused by this ambiguity may thus largely be removed in the final data collection. All scenarios were additionally made gender neutral to exclude that factor from influencing the Care/harm output factor. Multiple participants thought the surveys were too long and it was advised to make them shorter to improve the data quality. Therefore the final survey is partitioned in

6 parts of 10 questions each. Participants were allowed to take part in one or multiple survey partitions. The results from the pilot data are excluded from the final data collection and uses different participants.

Since the network architecture did not show any problems we will use it in the full-scale experiment as well.

5.2 Results

5.2.1 Survey

The 60 moral scenarios were repartitioned into six surveys of 10 scenarios each. 34 different subjects participated in one or more surveys, resulting in a total of 96 responses. The standard errors of the mean are respectively $\frac{1}{\sqrt{18}} = 0.24$, $\frac{1}{\sqrt{14}} = 0.27$, $\frac{1}{\sqrt{13}} = 0.28$, $\frac{1}{\sqrt{16}} = 0.25$, $\frac{1}{\sqrt{16}} = 0.25$ and $\frac{1}{\sqrt{19}} = 0.23$.

The linear correlations between the input factors are shown in the table below. The correlation is measured by Pearson's correlation coefficient.

	Weak	Emotionally	Compassion	Defenselessness	Cruel	Kill
Weak	1	0	-0.03	0.82	0.120	0.340
Emotionally	0	1	-0.71	0.24	0.73	-0.03
Compassion	-0.03	-0.71	1	-0.22	-0.95	-0.01
Defenselessness	0.82	0.24	-0.22	1	0.33	0.42
Cruel	0.12	0.73	-0.95	0.33	1	0.04
Kill	0.34	-0.03	-0.01	0.42	0.04	1

Table 4: Linear correlations between input factors

Again we find several cases where judgements are distributed in a bimodal and uniform way. Some of these cases are shown in Figure 5 and Figure 6. The appearance of both distributions was also found in the pilot experiment (section 4.2.1).

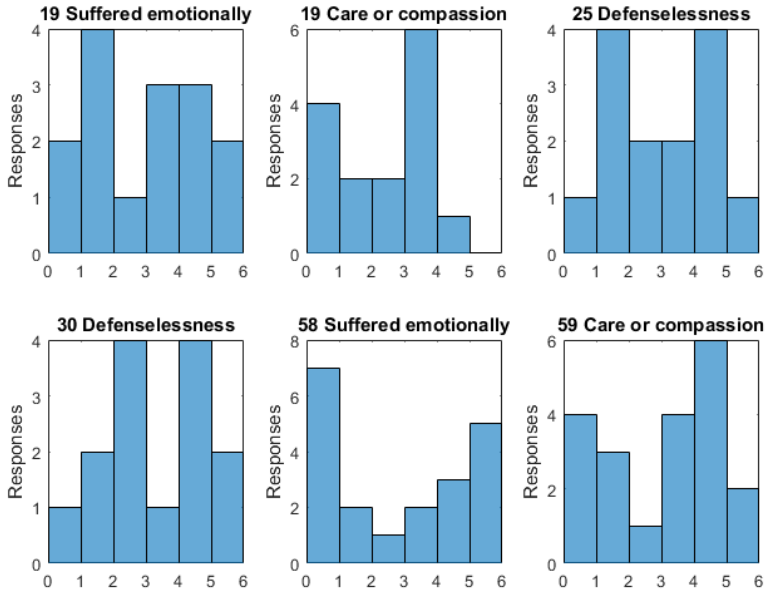

Figure 5: Multiple questions showing possible bimodal normal distributions.

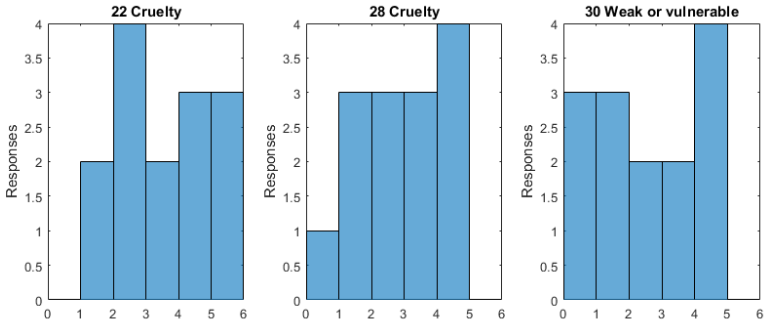

Figure 6: Multiple questions showing possible uniform distributions.

5.2.2 Neural Network

In the same manner as in the pilot section (4.2.2), the network is repeatedly trained with different combinations of parameters and random weight initialization. Training data is shuffled before feeding it to the network, to ensure that the performance is not dependent on a specific ordering of the data. The table belows shows the values tested for the different parameters.

	Values
Number of epochs	50, 100, 200
Hidden neurons	5, 10, 15, 20, 25
Learning rate	0.01, 0.05, 0.1, 0.25, 0.5, 1

The number of epochs with the highest performance values was experimentally determined to be 100. Figure 7 shows the course of the average of five performances for $epochs = 100$. The average performance of five networks was taken to discount randomness from data-set shuffling and from weight/bias initialization. Notably most networks seem to reach their maximum performance around a learning rate of $\eta = 0.25$, after which it declines. For both performance measures the network with 10 hidden neurons yields the highest performance. For $epochs = 100$, $\eta = 0.25$ and hidden neurons = 10 the average performance is 53.66% and the average adjusted performance is 72.16%.

These obtained parameters were then used for further testing. Table 5 shows the performances for additional network settings.

	Performance	**Adjusted performance**
Original input	0.55	0.73
2 hidden layers (10-10)	0.52	0.71
2 hidden layers (10-5)	0.52	0.71
No training data shuffling	0.51	0.71
Decorrelated input	0.50	0.69
Zero weight initialization	0.49	0.68
No hidden layer	0.45	0.65
Batch learning	0.31	0.55

Table 5: Performances for additional network settings

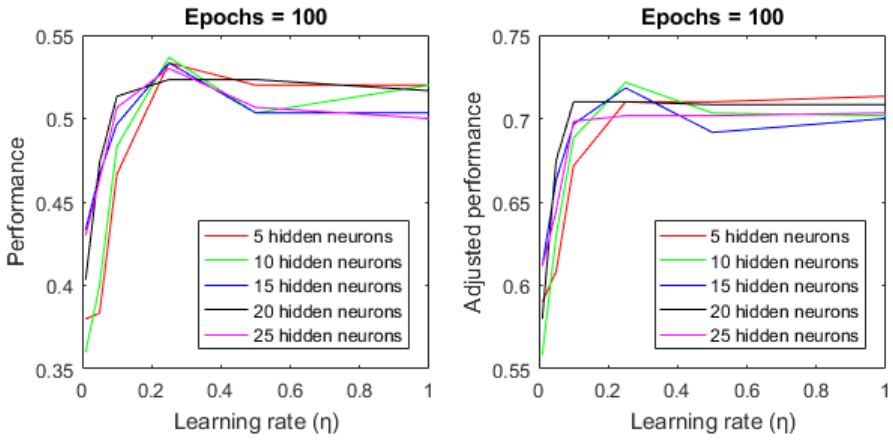

Figure 7: Average performance and adjusted performance of training the networks five times at 100 epochs. Each network used varying learning rates and hidden neurons.

Figure 8 shows the weights of a trained network with the maximum performance of $P = 1$ on the test cases. This is one of the 10 networks in a 10-fold cross validation and thus was trained on 54 and tested on 6 cases. Red lines indicate negative weights, green lines indicate positive weights and the line thickness indicates the quantity of the weight. Table 6 and Table 7 hold the weight values corresponding to the figure.

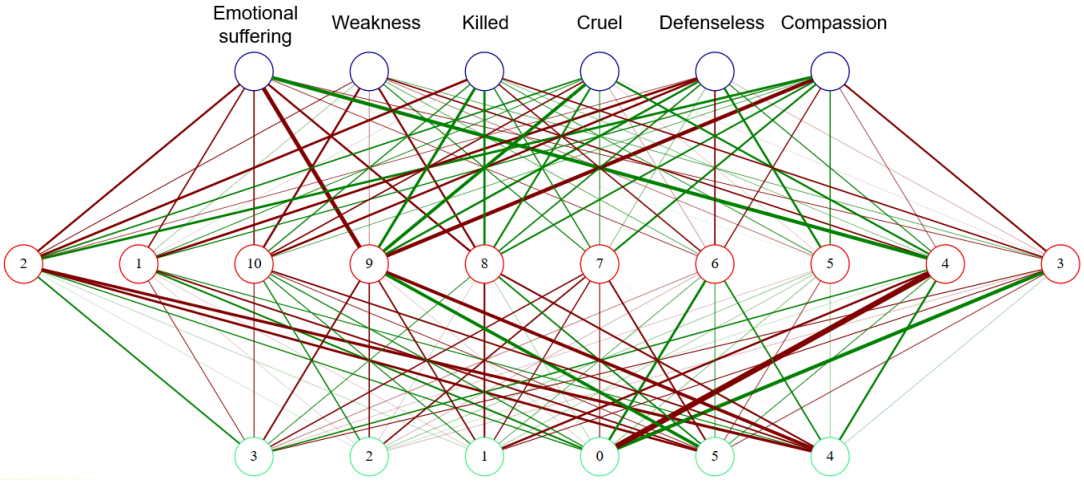

Figure 8: Example of hidden neurons weights

Input	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**
Weak	0.31	-0.80	0.52	-1.36	-0.44	0.50	0.98	-1.83	-0.39	-1.94
E. Suffering	-1.33	-1.89	-0.67	3.34	0.07	-0.63	0.03	-1.92	-3.84	-1.25
Compassion	1.06	2.20	-1.70	-0.55	1.10	-1.09	1.75	1.59	-3.50	0.51
Defenselessness	-2.19	-0.56	-0.24	0.96	1.99	-1.40	0.40	1.34	1.96	-1.99
Cruel	0.75	1.39	-0.10	1.84	-0.15	-0.31	0.77	1.70	3.0	-1.33
Kill	-0.12	-2.15	-1.36	0.39	0.70	-1.26	1.14	2.03	2.51	1.10

Table 6: Weights from input neurons to hidden neurons

Harm	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**
0	1.62	1.09	3.34	-5.23	0.38	2.17	-0.86	0.26	-0.15	1.04
1	-0.16	0.22	-0.85	-1.97	0.29	0.54	-1.12	-1.62	-1.42	0.96
2	-0.14	0.14	-0.21	0.17	0.33	-0.24	-1.40	-0.17	-1.24	1.18
3	-0.73	1.52	-0.86	1.28	-0.19	-0.41	-1.23	0.74	-1.58	-0.83
4	0.92	-2.65	0.25	1.88	0.32	1.43	-0.03	-1.67	-3.06	-1.47
5	-1.53	-2.17	-0.69	0.73	-0.63	0.70	-1.59	1.47	2.97	-0.70

Table 7: Weights from hidden neurons to output neurons

5.3 Discussion

5.3.1 Survey

Again, for several survey questions there appear bimodal and uniform distributions. This can still be due to the large margin of error but it could also represent the distribution of the population. As previously discussed in 5.3 we need to make a decision about what to do with these cases. Because the distributions still appear after clearing up ambiguous question and increasing the sample size, there is less reason to belief that they are outliers. By discarding them away we are ignoring valid data and favoring only the unambiguous part of the moral function. Additionally using less cases will slow down the network training and could also decrease its performance. For this reason the network is still trained by using the average value for the factors on each moral scenario. Since we deviate from the measured moral function the robot could make moral judgements about Care/harm that are not psychologically plausible. In fact the results suggest that a completely unilateral moral function may not exist.

The first sub-question of our research question was:

> Can a survey-based approach be used to a) obtain a unilateral moral function that distinguishes Care/harm and b) can be learned by a multi-layer perceptron (artificial neural network)?

Given our input factors we have found a cases where the moral function seems to be ambiguous, i.e. there are certain moral scenarios where there are different populations with different judgements. In those cases the MLP has to learn which judgement is the correct one. Since we use the average judgement there

could appear cases where the MLP learns to make a psychologically implausible judgement. Because these cases influence the MLP it will not learn a completely unilateral moral function. The sub-question has to be answered with "no", which also prompts us to reevaluate the question what behaving in a human-like fashion means.

In 3.1 we pointed out a possible positive correlation between the input factors Weakness and Defenselessness and a possible negative correlation between Cruelty and Compassion. Table 5.2.1 both show indeed strong correlations between those factors. Additionally there is also a positive correlation between Emotional Suffering of the patient and Cruelty and a negative correlation between Emotionally Suffering and Compassion. This suggests that Cruelty and Compassion ($r = -0.95$) measure the same variable and that one of them could be excluded from the survey. To facilitate the training of the network the input is thus decorrelated, which is recommended by LeCun [18]. Since both factors are adaptions from the MFQ item key this also raises the question whether MFQ's items WEAK and CRUEL (section 3.1) correlate with each other. However there is an important difference in our adaption since we separated the item "Whether or not someone cared for someone weak or vulnerable" (WEAK) into "Whether or not someone cared for" and "Whether or not someone is weak or vulnerable". Despite the different contexts there is nonetheless a possibility that there is a correlation between both items.

A more general problem with a survey-based approach is the large amount of participants and cases that is required. An ANN needs a large number of cases to train on to learn a pattern in data. For this reason a lot of different moral questions or scenarios need to be formulated. A lot of participants are needed because it can be preferable to measure moral judgement of a whole population instead of moral judgement of a single person. This would be impractical if for instance 385 participants (section 3.1) would need to answer a few hundred questions in order to obtain significant measurements. A survey may also need to be revised multiple times to avoid ambiguous and unclear questions. Then finally there is also the possibility that hidden factors in the moral scenario affect the output factor, but are not incorporated by the input factors. One of the challenges is thus to design a good representation for moral problems in a survey-based approach.

We can't cover the issue of a good representation in depth, however we can make a few suggestions. Maybe a good start is to ask "What factors really constitute a moral problem?". A moral problem can be regarded as a classification problem, abstracted from reality, that has certain factors or features to describe it. In our case we have limited the moral problem to an agent, a patient that is potentially being harmed and six factors to represent the situation. Are those factors effective for moral judgement and reasoning, or are other factors also part of the moral equation? One could rely on existing moral theories to approach this, but there is no guarantee that there aren't better factors to represent the problem with. A problem of using too few factors is that they may not be sufficient to generalize the moral judgement of the participants. Using many factors however, the dimensionality of the data also increases. This means

that for machine learning purposes the data is harder to classify, often referred to as the *curse of dimensionality* [17]. As a result more cases are often needed as well. Ideally we should use a low number of effective factors that explain the data as well as possible.

The second question to ask is what medium to use to present the moral problem to a participant. We have used written stories but alternatively one could for example use pictures or videos. An advantage of using a visual representation is that the problem is arguably presented in a more objective way with a lower chance that participants misunderstand the problem. The downside to this is that it would take even more time to design a lot of cases. A possible solution could be to use computer generated images or animations of moral problems.

Finally, a good representation should generalize well to the real world. Eventually the agent should be able to apply its training to a corresponding real-world application, and also learn from new experiences. This it perhaps the hardest part of the design and highly depends on the application. At the very least a problem in the real world should have some sort of translation to an input representation for the agent. One should also aim to find the most effective factors that describe a moral problem in the real-world application, in order to limit the information that is lost in the abstraction.

5.3.2 Network

By answering the second sub-question we can investigate what kind of function we have obtained and how well it performs:

> What parameter and architecture choices for the neural network result in the highest performance?

A difference between the pilot data and the final data is the decrease of the number of epochs needed for a similar performance. Also the network's performance is higher when trained on the final data than on the pilot data. This fits with the hypothesis that a larger sample will obtain an underlying function with less noise and randomness. The data quality could also be improved because of the changes made to the survey (see pilot discussion in section 4.3).

The network seems to display a robust behaviour as seen in Figure 7. The lines that represent the amount of hidden neurons are all similar and close to each other, with the "10 hidden neurons" line having the maximum performance. The learning rate with $\eta = 0.25$ shows optimal performance. The lower learning rates have not enough impact on the weights to train the network effectively. For learning rates above $\eta = 0.25$ the performances seem to slowly decrease and the adjusted performances seem to stagnate.

Applying further measures to the network (Table 5) only show decreases in performance. On-line learning, data shuffling and random weight initialization are indeed necessary to increase performance [18]. The large drop in performance from removing the hidden layer also suggests that the learned function

is non-linear. The hidden neurons thus serve a function or could act as sub-features. On the other hand, adding another hidden layer does not seem to increase the performance. This could mean that one hidden layer is enough to describe the underlying function in the data.

Analyzing what functions the hidden neurons could fulfil is very speculative. When classifying images with an ANN we can visualize the hidden neurons and see what feature they extract. However with moral problems we can only look at which input factors excite and inhibit the hidden neuron and how it relates to the output. A few things can be observed in our function by looking at the highest inhibitory (negative) and excitatory (positive) values (see Table 6 and 7). H4 seems to use Emotional Suffering and Cruelty as best indicators that the level of harm can't be 0 (presumably some form of harm is done). H9 is activated by Killing and high Cruelty, but deactivated by Emotional Suffering and Compassion. This seems to describe cases where the patient is killed in a cruel and quick way. Because in those cases maximum harm (5) is done, level 4 is inhibited. H2 seems to decide that harm should not be judged at level 4 or 5 when Compassion is shown, however this is negated when the patient is killed or suffers emotionally. Those structures and rules do not seem illogical. The integration of all rules together is the bottom-up assembly that should describe the pattern in the survey data. This it interesting because typically bottom-up networks classify sensory data, e.g. visual images or sounds. However in this case, we can see how a "moral stimulus" is processed. Even though this is within our representation of a moral problem, it would be an interesting prospect to be able to figure out logical processes in moral reasoning and moral judgement by analyzing working bottom-up assemblies. Ideally we might even be able to turn those findings into top-down rules, essentially converting a bottom-up system into a (safer) top-down system. Alternatively they might also help to inspire hypotheses about the mechanisms of moral system in the human brain itself. Realistically this of course also depends on the size of the bottom-up system; finding logical processes in large assemblies could be a tedious and highly complicated task.

The network's best performance is 0.55 and the best adjusted performance is 0.73. A random agent would have a performance of 0.2 since there are five different output classes. This means our AMA performs well above chance level and has judged on average 3-4 out of 6 test cases correctly. The adjusted performance indicates that when 3 out of 6 cases are classified correctly, 2 cases are classified almost correctly (i.e. 1 off the predicted output). Depending on the application of the surveillance robot it could perform well in practice. Especially in areas that are already relatively safe, e.g. schools, libraries or malls, there is arguably small room for error. If the exact judgement of the robot is very important, e.g. for police use or military applications, then an accuracy of 55% is probably too low for practical use.

The network's performance could be improved by having a larger set of training cases (more survey questions) and by having a larger group of participants to decrease the error margin. The quality of the data is however always limited by the survey methodology, i.e. questions can ambiguous, unclear or misinter-

preted. And as we have discovered there could appear cases where there are different moral standpoints.

5.3.3 Bottom-up approach

As discussed in section 2.2, one of the main disadvantages is that there are no explicit rules or goals when learning within a bottom-up approach. The survey-based approach is essentially a form of supervised learning, with the challenge being giving the correct feedback to the AMA. Our criteria for correct feedback was that it was human-like. There appears however a conflict when there are multiple human populations with different moral judgements. Since all are correct it should now be asked which judgement belongs to the ethical norm.

Let's assume a robot needs to undergo a moral training in a lab to learn enough basic knowledge for a certain task. The morals of the robot will now be influenced by the person or group that trains it. A single person training the robot does not seem problematic: the robot could learn the user's moral judgements and simulate them. A whole group training the robot becomes more problematic though. The moral judgement of the group can not be represented anymore when a case appears that divides the group into two moral camps. This conflict could be solved democratically or even autocratically, but in both cases the moral judgement of one camp is ignored. The question then becomes what kind of morals this robot develops. The ignored camp might regard the robot as not at all morally developed. The probability that this will occur for certain cases only increases when the group needs to be representative for a population and consists of a large number of people. Another solution is for the different camps to find a compromise. This is similar to what we did by taking the average, and might be a good solution in practice. The downside to this is that the morality of the robot depends a lot on how the training group makes compromises. Also this is impractical for large groups and would be difficult to do by survey alone.

For full disclosure it must be noted the AMA is not the complete surveillance robot. A big part of the problem is making sensory measurements to observe a real-life scenario and extract relevant measurements from it. With the input representation we used it should be able to observe physical weakness, emotional suffering, compassion, defenselessness, cruelty and killing. Consequently also numerical values need to be attached to them to transform them into inputs for the network. That part is not addressed in this thesis.

Despite the described limitations, an adjusted performance of 0.73 could be acceptable in some contexts. This is also assuming that the survey is valid and does indeed measure the Care/harm dimension from the Moral Foundations Theory.

6 Conclusion

To investigate AMA design within the bottom-up and supervised learning approach, we started off by asking the research question:

> Can an artificial moral agent for a surveillance robot be designed by employing an artificial neural network, such that it can distinguish between Care/harm in a human-like fashion?

This lead us to design a survey to obtain human moral judgements on a variety of moral problems. These problems were related to the Care/harm domain of the Moral Foundations Theory and their representations were inspired by the variables from the Moral Foundations Questionnaire. We designed an architecture for a multi-layer perceptron such that it could learn the underlying moral function in the survey data. Additionally we defined two performance measures to reflect an exact accuracy and a more practical one.

The survey results show how there are multiple different distributions for moral judgements on moral scenarios. Our hypothesis in section 1 stated that we could find a common denominator in different moral judgements. This is contradicted by the results that suggest there are distributions with two or more common denominators. Several input factors show a strong correlation with each other. Cruelty and Compassion seem to measure the same variable, meaning that one of them is unnecessary for the survey.

The network learned a function based on the survey data. The behaviour of the network was shown to be robust, and the performance increased when trained on the final data. Interpreting the weights of a trained network gives the impression that the hidden neurons describe moral sub-features. Removing the hidden layer causes a drop in performance, supporting our hypothesis that the underlying function in the survey data is non-linear.

From this we conclude that the answer to our research question is "no", since there appears no such thing as a singular human-like morality. There occur no conflicts however when compromises between moral standpoints are trained, but this not the same as an organic human morality. Alternatively, when choosing one moral judgement over the other, the AMA learns the morality of a specific group or person. How to handle these conflicts should be considered in the design of the AMA. Finally we have found no inherent issues with a survey-based approach to obtain moral judgements, although it can be impractical due to the large number of cases and participants required.

There are several possible follow-ups based on this research. One is to test the validity of the survey data to ensure that the AMA would make moral judgements about Care/harm in real world applications. This could for example be done by using the characteristic emotions for the Care/harm foundation (see Figure 1). A scenario with a high score for Harm should then invoke emotions of compassion towards the patient and anger towards the agent. Another possible follow-up is to investigate what constitutes an effective representation of moral problems for supervised learning. Another useful application for supervised bottom-up learning would be the design of a database for human moral

judgements or similar moral data. Potential "big data" for morality could help to gain insight into the bottom-up assembly of moral functions. Additionally it may be used for training AMAs with supervised learning. As far as these follow-ups are realizable, they could contribute towards the substantiation of artificial moral agents and machine ethics.

We have found that we cannot train an AMA to apply human-like morality in a unilateral way, however there are possible compromises that can be made. Although the survey-based approach has its drawbacks, it can be used to obtain moral data, especially given a good representation for moral problems. The MLP shows to have the interesting capacity for bottom-up emergence of morality, and could be used as an aid to not only learn a moral function but also to analyze its underlying mechanism.

References

[1] Questionnaires — moralfoundations.org.

[2] Isaac Asimov. *I, robot*, volume 1. Spectra, 2004.

[3] John Bowlby. *Attachment and loss*. Number 3. Random House, 1998.

[4] John Brandon. 5 uses for the surveillance robot of tomorrow, 2014.

[5] David Chalmers. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10):7–65, 2010.

[6] Peter Danielson. *Artificial morality: Virtuous robots for virtual games*. Routledge, 2002.

[7] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, Forthcoming*, 2012.

[8] Jesse Graham, Jonathan Haidt, and Brian A Nosek. The moral foundations questionnaire. *MoralFoundations.org*, 2008.

[9] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.

[10] Marcello Guarini. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, (4):22–28, 2006.

[11] Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

[12] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press, 1975.

[13] Ali Reza Honarvar and Nasser Ghasem-Aghaee. An artificial neural network approach for creating an ethical artificial agent. In *Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on*, pages 290–295. IEEE, 2009.

[14] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[15] Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS one*, 7(8):e42366, 2012.

[16] Immanuel Kant. The fundamental principles of the metaphysic of ethics. 1939.

[17] Eamonn Keogh and Abdullah Mueen. Curse of dimensionality. In *Encyclopedia of Machine Learning*, pages 257–258. Springer, 2011.

[18] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[19] John McCarthy. Epistemological problems of artificial intelligence. *Readings in Artificial Intelligence*, page 459, 1987.

[20] N.C. Ring, K.S. Nash, F. Glennon, and M.N. MacDonald. *Introduction to the Study of Religion*. Orbis Books, 2012.

[21] J Maynard Smith. Group selection and kin selection. *Nature*, 201:1145–1147, 1964.