OSLOMET

Master's Thesis

Master's Program in Behavioral Science - Specialisation in Concepts and Applications

March 2023

Human versus computer responding

Simulating stimulus equivalence experiments using Enhanced Equivalence Projective Simulation

Name: Sarah Fjeld Oueslati
Course code: MALK5000

30 ECT

Faculty of Health Sciences
OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Acknowledgement

I would like to express my gratitude for the patient supervision given by Professor Erik Arntzen during my work on the thesis. I would also like to extend a thank you to post doctor at UIB and co-author of the simulation code used in this thesis, Asieh Abolpour Mofrad, for guiding me in running the simulated experiment correctly. Furthermore, I would like to express my gratitude to my colleagues at Mindshift, for their encouragement, flexibility, and willingness to be experimented on as well as letting me turn one of their meeting rooms into a temporary lab. Last, but not least, a thanks to my family and friends for cheering me on and believing I could make it to the finish line.

Human versus computer responding -

Simulating stimulus equivalence experiments using Enhanced Equivalence Projective Simulation

Abstract

The purpose of this study was to explore if it is possible to simulate participant responding in a stimulus equivalence experiment, using the enhanced equivalence projective simulation model from Mofrad et al. (2021). Five human participants and 45 simulated participants, characterized by different model parameter values, were exposed to the same matching to sample procedure. In both experiments, the participants in average did not attain the criterion of 90% correct trials during equivalence test. The mean human participant equivalence test score was 84%, while the mean simulated equivalence test score was 82%. Human participants passed the baseline and symmetry tests with a mean score of 97%, against a mean score among simulated participants of 89%. The simulated participants, when using the same parameter values, were homogenous in terms of their response patterns, while the human participants exhibited a larger variation. A suggestion for further work, groups of simulated participants with different sets of parameter values, could perhaps yield inter-group responding similar to the observed responding in human experiments. A practical application of these findings is to explore how simulations and human experiments can be compared at a larger scale. A possible solution could be to add the simulation code to the MTS software that is used in the human experiments, so that simulations are automatically run and compared to human responding.

Keywords: Stimulus equivalence, simulation, connectionism, artificial neural network, enhanced equivalence projective simulation, machine learning

Introduction

The aim of this paper is to understand whether results from a stimulus equivalence experiment with human participants can be simulated successfully using a computer model based on an algorithm called enhanced equivalence projective simulation. This research problem was introduced in the paper "Equivalence Projective Simulation as a Framework for Modeling Formation of Stimulus Equivalence Classes" (Mofrad et al., 2021).

Stimulus equivalence

Stimulus equivalence is relationship between stimuli that emerge without directly reinforced training. These relationships are called emergent relations and emerge as a consequence of the conditional discrimination procedure (Sidman, 1994, 2000). Unlike other stimuli classes that can form, stimulus equivalence classes are not formed based on topographic similarities or stimulus generalization, and they do not necessarily share the same behavioral functions (Green & Saunders, 1998).

Murray Sidman and William Tailby introduced the different characteristics of stimulus equivalence in their paper from 1982. To be able to call a conditional relation an equivalence relation, the relation must, according to Sidman, be shown to be reflexive, symmetrical, and transitive (Sidman, 1992). To this end, researchers use conditional discrimination procedures. A conditional discrimination procedure is defined by the relationship between discriminative stimulus and conditional stimulus, in a four-term contingency. The behavior is produced only in the presence of the stimuli in a given context. A matching to sample protocol is an example of a conditional discrimination procedure used to demonstrate stimulus equivalence (Sidman, 1992; Sidman & Tailby, 1982). The relation between two stimuli is reflexive if both stimuli have the

same relationship to themselves as to each other. For the relationship to be called symmetrical, sample stimuli and comparison stimuli must be functionally interchangeable. Furthermore, the relationship must be transitive, if one has a conditional relationship between a sample stimulus A and a comparison stimulus B and between a sample stimulus B and a comparison stimulus C, then a condition arises between A and C. If a relation has formed between C and A, i.e., the transitive relation is symmetrical, then global equivalence has been achieved (Sidman et al., 1974; Sidman & Tailby, 1982).

To pass the global equivalence test, the symmetry and transitivity tests must also be passed. Research has shown that the definition and the tests hold, and where inconsistent results have been found, they can be explained by weaknesses in the experimental design (Sidman, 1992).

In a matching to sample procedure where the goal is to demonstrate stimulus equivalence, we often use abstract (symbolic) stimuli. That is, stimuli are not meaningful i.e., they do not belong to an established stimulus class in the participants in advance. If they do, previous learning history associated with the stimulus will affect the results of the experiment, and it will be difficult to conclude about the lack or achievement of equivalence.

Simulating emergent relations using machine learning

Artificial neural networks (ANN)

An artificial neural network (ANN) is a type of machine learning algorithm that is inspired by the biological brain and how learning happens (could happen) in human/animal brains (Goodfellow et al., 2016, p. 13). ANNs consist of layers of interconnected "neurons," which process and transmit information. In a biological neural network the neurons receive

stimulus, change them via synaptic weights, combine them and produce a single response (output) dissimilar to the combination, typically a prediction or classification based on the input. (Guresen & Kayakutlu, 2011). ANNs are trained using large amounts of data and an algorithm that adjusts the connection weights between neurons based on the input data. Haykin (1999) defines an ANN as "a massively parallel combination of simple processing units which can acquire knowledge from environment through a learning process and store the knowledge in its connections."

The performance of an ANN is based on finding the right set of weights, i.e., the strength of the connections between the processing units. The network uses algorithms to calculate the right weights for different tasks. Initially the weights are set at a random value, then the algorithm adjusts the weights gradually in the direction needed to approach the correct output values seen during training. The algorithm repeats this approximation towards the correct output for a given number of times, or until it correctly produces correct output for all given inputs in the training data set. A successful training will also entail that the network can use what it has learned from the training data set to find correct new outputs for a new set of input values, analogous to stimulus generalization in humans (Ninness et al., 2018).

Connectionist models

Parallel to the research and application of ANNs in the fields of natural sciences/computer science, the research within connectionist networks within the behavioural science has been focused on the use of these methods to simulate emergent relations. This field of study is called connectionism and had its early beginning in the late nineteen eighties and the early nineties. Connectionism is a field within cognitive science that seek to explain cognitive

processes using artificial neural networks, inspired by the biological neural networks part of the human brain (Bechtel & Abrahamsen, 1991).

When training an ANN to predict class identity for a given stimulus, there are a few components in the training protocol that differs from the training protocol used in human participant experiments. In a typical human experiment, the participant is exposed to abstract symbols. In a ANN training, these stimuli is expressed as a series of binary activation units that are either on (1) or off (0) (Ninness et al., 2018).

The input layer consists of the processing units in the network that accept input stimuli, while the output layer consists of processing units outputting values from the network. Between the input and the output layer we find one or more layer, the hidden layer(s). These layers only interact with the input or the output layer or each other, the way they interact is by means of synaptic connections or weights. These weights enable the network to learn the relationship between the input and output stimuli (Ninness et al., 2018).

As the ANN is repeatedly exposed to the different input strings of ones and zeroes and uses feed forward back propagation to adjust the weights, it recognizes unique patterns in the input stimuli, just as human participants learn the relationship between the different abstract stimuli after being exposed to the MTS training protocols several times (Ninness et al., 2018).

To assess if the ANN has successfully acquired stimulus equivalence, we introduce the generalization test as a new set of input strings. If the ANN successfully predicts the correct output, the ANN has acquired the equivalence relations, just as the human participants is said to have acquired equivalence when correctly matching the not trained emergent relations during the test phase of MTS (Ninness et al., 2018).

There are mainly two variations of ANNS employed in the connectionist stimulus equivalence research, RELNET and EVA (Ninness et al., 2018). RELNET, short for Network for Relational Responding, was introduced by Barnes and Hampson (1993) in their paper "Stimulus equivalence and connectionism: Implications for behavior analysis and cognitive science" in 1993. Here Barnes & Hampson maintain that the stimulus equivalence research area within behaviour analysis and the connectionist area within the cognitive sciences could benefit from taking a connectionist approach to the stimulus equivalence research. In this paper they seek to point out an intersection between connectionism and stimulus equivalence by successfully simulating emergent equivalence relations using a connectionist network/ artificial neural network. In this case it was the study by Steele and Hayes (1991), where the researchers demonstrated contextual control of derived stimulus relations in human responding to nonarbitrary stimuli. Using a RELNET model, Barnes and Hampson (1993) demonstrated that they could successfully run simulations of the Steele and Hayes study (1991) and end up with similar results to the results of the human participants in the original study. It was a starting point for later papers that uses RELNETs to simulate stimulus equivalence and transfer of function. In the last three decades there are several research papers published on how we can simulate emergent stimulus relations in humans using RELNET (e.g., Lyddy & Barnes-Holmes, 2007; Lyddy et al., 2001; Tovar & Chávez, 2012; Tovar & Westermann, 2017; Vernucio & Debert, 2016).

A special feature of the RELNET is that the input layer also includes so-called sample marking duplicators, three elements that informs the network how identify the contextual stimuli. This unique among ANNS. Thus, RELNET does not acquire the contextual relations on its own,

it is rather pre-wired to acquire them due to the sample marking duplicators. This is remarked as a possible flaw with RELNET (Barnes & Hampson, 1993). The models employed by Barnes and Hampton in 1993 and the subsequent studies Lyddy et al. (2001), Lyddy and Barnes-Holmes (2007), Tovar and Chávez (2012), and Vernucio and Debert (2016) employed versions of the previously described RELNET connectionist model.

Another feature of the RELNET network is that it includes a training set that is meant to compensate for human experience outside the experiment setting. In Ninness et al. (2018), the researchers introduce a new type of ANN, called EVA. With the EVA (emergent virtual analytics) the researchers abandon this compensatory training. and the sample marking duplicators. In this paper, the researchers present the results from replicating the study by Tovar and Chávez (2012) and replicated by Vernucio and Debert (2016) using an evolved connectionist model they call EVA (emergent virtual analytics). They go on to show that this network is able to yield the same performance (i.e. acquire derived stimulus relations) as the RELNET employed in the Tovar and Chávez (2012) and Vernucio and Debert (2016) studies. Ninness et al. (2018) maintains that ANNs "are capable of performing in ways that are very similar to those seen among human participants" (p. 141). The EVA algorithm continues to of interest within connectionist research (Ninness et al., 2019; Ninness et al., 2021).

Enhanced Equivalence Projective Simulation (EEPS)

Projective Simulation (PS) is a new machine learning model that incorporates principles from physics, and was introduced in 2012 (Briegel & De las Cuevas). It is a reinforcement learning algorithm that can perceive stimuli, execute actions, and learn through trial and error. PS has a neural network structure that serves as its physical basis, with a memory system called Episodic

& Compositional Memory (ECM). ECM is a directed and weighted network of clips, each representing a remembered percept, action, or sequence. The recall of memories in PS is understood as a dynamic replay of an excitation pattern, leading to episodic sequences. The learning program in PS is updated by adjusting connection weights and adding new clips through Bayesian rules and interactions with the environment (Melnikov et al., 2017).

In the paper “Equivalence Projective Simulation as a Framework for Modeling Formation of Stimulus Equivalence Classes” (Mofrad et al., 2020), the authors simulate nine different MTS protocols using a EPS algorithm, among others the same protocol used in the Sidman and Tailby experiment (1982). Based on the results of these experiments, the authors concluded that they work as a proof of concept that the EPS algorithm can be successful in modelling different aspects of human responding during MTS experiments given more study, for instance by tuning of parameters to model specific behaviour (Mofrad et al., 2020).

In Mofrad et al. (2021), the paper from which the algorithm employed in the simulated experiments described in this paper, the build upon their findings in the 2020 paper, and introduces Enhanced Equivalence Projective Simulation (EEPS). The original EPS model assumes that relations between elements in a trial are generated on demand during testing and updated during training. However, the assumption is changed so that these relations are formed at the end of the training phase, resulting in a noisy version of the agent's memory network. By using a denoising method, the network can be cleaned up to better display information about equivalence class formation. This is called network enhancement, and the aim is to identify the most reliable connections in the network and eliminate weak or false links, resulting in a more robust representation of the biological processes underlying the network (Wang et al., 2018). The

resulting denoised network can be used in testing and to evaluate the agent's performance on equivalence tests.

The h-values, also known as projective simulation values, are a critical component of the Enhanced Equivalence Projective Simulation (EEPS) algorithm. They serve as an indicator of the equivalence between nodes in a network, where higher h-values signify a stronger equivalence. Compared to Mofrad et al. (2020), the 2021 paper introduces new updates to memory retrieval during testing, control over symmetry and transitivity relations, improved network enhancement, fewer parameters to fine-tune, a new method of deriving relations using a diffusion model using the updated network as a cognitive map, as well as reducing computation during testing (Mofrad et al., 2021).

The research problem

The goal of this study is to further understand if it is possible to simulate human responding in stimulus equivalence experiments accurately, using computers. If it is possible, it could entail a reduction of research resources, as experiments run on computers are done in seconds instead of hours. Another advantage of computer simulation is the possibility of complete experimental control. With human participants in the lab, there are always factors from outside the laboratory that can or cannot influence the performance in the experiment, while with simulated participants the setting is under experimental control and participants are not prone to physiological error sources such as hunger, fatigue etc. (Ninness et al., 2018). This is also pointed out by Lyddy and Barnes-Holmes (2007), were they observe that computational modeling could be one alternative source of understanding how the emergent stimulus relations are formed, due to the fact that most sources of error from human experiments can be ruled out (Lyddy & Barnes-Holmes,

2007). The long-term goal for this line of research is that computer simulated experiments can provide a more complete and unified understanding of underlying mechanisms of stimulus equivalence.

General method

Design

The overall design was a between-group design, one group consisting of five human participants (experiment 1) and nine groups each consisting of five participants simulated by an EEPS-model (experiment 2). Experiment 1 followed an AB design. All groups started by completing a training phase where baseline relations were trained. Then they completed the testing phase.

Procedure

The protocol was simultaneous matching-to sample (MTS). All participants were presented with one sample stimulus and three comparison stimuli. The measured response was the participant's choice of comparison stimulus they believed matched the sample stimulus. The procedure consisted of two phases. Establishing and maintaining baseline relations (training) and testing for emergent relations (testing).

Mastery Criteria

The mastery criterion was set to 90% correct responses. To progress to the next phase of the experiment the participants had to select the correct comparison stimulus in at least 90% of

the trials in a block. The criterion was used consistently throughout the training phase, when establishing and maintaining baseline relations.

Experiment 1

Method

Participants

The participants ranged in age from 30-55, three male and two female. Prior to starting the experiment each participant was given a document to read, stating the background for the project, describing the experiment situation, and presenting the researchers responsible for the experiment. Furthermore, the document described the ethical considerations and how the data collection in the experiment was anonymous so that no individual data can be traced back to a specific participant. It also stated the terms of withdrawing the consent given to use the data that was produced during the experiment. They were also informed that a debriefing session would be given after completion of the experiment.

The consent form was part of the computer program used for the experiment procedure and had to be accepted before the participants could start the experiment. Each participant was also given a document regarding the treatment of contact information in research experiments at the institute of behaviour science at OsloMet University, prior to starting the experiment.

Setting

The experiments were performed in a quiet meeting room, with dimmed lights, to ensure the participants ability to concentrate on the task.

Apparatus

The experiment was performed using a custom MTS software, on a 15 inches Hewlett Packard laptop computer with Windows 8 (64 bit). The software controlled the presentation of all the stimuli and recorded the results of the experiments as excel files. An external mouse connected to the laptop was provided for the participants to use, to click on the stimuli presented on the screen during the experiment. During the experiment, the custom software placed a sample stimulus in the middle of the computer screen, while placing the three comparison stimuli randomly in three of the four corners of the screen.

Instructions

All participants were presented with the following instructions on the screen prior to starting the experiment: "A symbol will appear in the centre of the screen. You have to click on this with the mouse. Three other symbols will appear. Select one of these by clicking with the mouse. If you select the one defined as correct it will say "good", "great", etc. on the screen. If you press incorrectly, "error" will appear on the screen. Throughout the experiment, the computer will give less and less feedback on whether your choices are right or wrong, but from what you have learned you can get all the tasks right. Do your best to get most correct answers as possible. Good luck!"

Consequence thinning

When establishing the baseline relations, the participants were presented with a programmed consequence for all trials (100% of the time) informing them if their response is correct or not. When the participant met the mastery criterion, the baseline relations was defined as established. The training changed to maintaining the established baseline relations, by introducing consequence thinning. The participants were presented with a programmed

consequence three out of four trials (75%), in blocks of 18 trials. When the mastery criterion of 90% correct trials was reached, the programmed consequences was introduced for one out of two trials (50%). If the participant reached the mastery criterion in the following block the programmed consequences would go to 0 percent for the next block, else the feedback rate dropped to 25 percent. Programmed consequences were displayed on the screen for 500 milliseconds, followed by an inter-trial interval of 1000 milliseconds.

MTS training

During the training phase, six sets of conditional relations were trained in a simultaneous training protocol (table 1). The baseline relations AC and BC were presented in blocks consisting of 18 trials, where all six baseline relations were trained in random order within each block.

A set of nine abstract stimuli (figure 1) was employed. The stimuli consisted of three classes (1,2,3) with three members (A, B, C) in each class (figure 1): a. (A1, B1, C1), b. (A2, B2, C2) and c. (A3, B3, C3). The training structure was MTO (many to one) where the participants were trained to choose stimulus C1 in the presence of A1 and C1 in the presence of B1, and so on. The sample presentation was simultaneous MTS (SMTS) which means that selection and comparison stimuli were simultaneously presented on the screen.

MTS test

After the participants met the 90 percent mastery criterion with no programmed consequences during training, they progressed to the test phase, which consisted of testing for symmetry- transitive and equivalence relations, as well as the baseline relations.

The test block contained 54 trials, where 18 trials tested baseline relations (AC, BC), 18 trials tested whether symmetry relations (CA, CB) were established during training and 18

relations were tested for stimulus equivalence (AB, AC, BA, BC, CA, CB) (table 1). During the test there were no programmed consequences, and relations were tested in random order.

The participants were defined as responding in accordance with stimulus equivalence if they reached a minimum of 90% correct responses on the equivalence test.

Results

Establishing baseline

When we examine the response pattern for each participant, we find that three (18552, 18554 and 18555) of the five participants consistently increase the number of correct trials for each block, of which two had the fewest number of trials to meet the mastery criterion during training. The mean number of trials to establish the baseline relations was 159. There is a relatively large variation in responding within the group, with a standard deviation of 162 trials, a value slightly higher than the mean. The participant with the largest number of trials had 432 trials distributed over 24 blocks. On the other end we had one participant with 36 trials, less than 10% of the maximum observed number of trials (figure 3).

The group average number of errors is 53, which provides a mean error rate of 33 percent. The error rate was calculated by dividing the mean number of errors by the mean number of trials. The error rate ranged over an interval of 11 percent to 37 percent in this phase of the experiment (figure 4).

Maintaining baseline

In this phase of the experiment the participants have reached the mastery criterion of 90 percent correct baseline relations. In the maintaining phase of the training the feedback was

thinned (consequence thinning). The mean number of trials needed to reach the mastery criterion in one block, with no feedback, was 61, with 72 being the largest number of trials observed. The minimum trials observed was 54, the minimum number of trials possible in this phase, as each participant at least had to finish one block of 18 trials for each level of programmed consequence (75,50,0) (table 1).

Equivalence class formation

As a group, the participants in experiment 1 did not attain the criterion of 90% correct trials during the equivalence test, with an average score of 84% (figure 4), while they obtained a score of 97% when testing the baseline and symmetry relations. Three out of the five participants (60%) formed equivalence classes (table 4). The two participants that did not reach the equivalence formation criterion earned an equivalence test score of 0,5 and 0,84 respectively.

Four out of five participants (80%) formed symmetry relations. The one participant not showing symmetry relation mastery (1853) had a correct response rate 89%, one percent point below the test criterion of 90. Likewise, four out of our five participants maintained their baseline relations. The one participant (1853) not meeting the baseline test criterion scored 89% (table 2).

Error patterns

Within MTS-protocol there are three different types of errors; random, experimenter defined (correct responses) and participant defined (systematic incorrect responses). Participants 18551 and 18553 had a strong tendency towards choosing C1 in the presence of B2. This tendency is not observable in the response patterns of other participants. On the other hand,

18554 seems to have errors spread out relatively even over the different sample stimuli (figure 5).

Discussion

The goal of this experiment was to measure how five human participants performed in terms of learning rate (number of trials) and formation equivalence relations. The results show large variations between the participants both in terms of class formation and number of trails. It is interesting to observe that the participant with the greatest number of trials in both the establishing and maintaining phase, and subsequently the highest error rate, was one of the three participants that formed equivalence classes, with an overall score on baseline, symmetry, and equivalence of 98%. On the other hand, the participant with the fewest trails (36) in the establishing phase of the training, did not form equivalence relations according to our criterion of 90%, scoring only 50% in the equivalence test.

In terms of the variation in participant responding during training, findings are somewhat in keep with earlier findings from matching to sample experiments with similar procedural variables. When comparing experiment 1 to a high impact experimental study where the protocol is like the one used in experiment 1, for instance Sidman & Tailby (1982), simulated in Mofrad et al (2020), we find larger variation in responding our experiment. This could be explained by the characteristics of participants, in the Sidman & Tailby study children of similar age, and in experiment 1 adults ranging from 30 to mid-fifties with variation in education and profession.

In Sidman & Tailby (1982) 75% of the participants formed equivalence relations. Given this, as well as the protocol used in experiment 1, there would be reason to expect that a higher

share of participants would form equivalence relations. The training response pattern for participant 1854 could be interpreted as a mastery criterion set too low, resulting in baseline relations not being properly established before testing started (Arntzen, 2012).

Experiment 2

In this experiment the goal is to simulate nine groups of five participants with an EEPS algorithm. The simulation code is downloaded from a GitHub repository (Mofrad et al., 2021).

Method

Participants

The different participants are characterized by a set of adjustable parameters in the algorithm. Changing these parameters will change the simulated participant's responding. There are six parameters that can be adjusted in the simulation interface. A key feature in the EEPS is the directed network enhancement (DNE) method, that is proven to be appropriate when using EEPS to model emergent symmetry and transitivity relations (Mofrad et al., 2021, p. 503). For this experiment DNE is set as a default for all the simulated participants. Furthermore, the gamma damping parameter (γ), that controls learning decay, is set at default value $\gamma = 0{,}001$ for all participants, considering this parameter analogous to the maintaining baseline phase in the training protocol in experiment 1. The value of $\gamma = 0{,}001$ was used in Mofrad et al. 2021 for all experiments except the one that studies the effect of changing this parameter (experiment 4). Lastly, K is a positive value that controls the symmetry relation. K was set to one throughout the experiment, which means that the network is symmetric at the end of the training phase (Mofrad et al., 2021, p. 493).

Groups of five participants were simulated with the same set of parameter conditions in each group. Nine different parameter conditions were used (table 3). I refer to the source paper (Mofrad et al., 2021) for detailed descriptions of the mathematics behind the EEPS algorithm and its adjustable parameters.

Apparatus

The simulation ran in Python 3.9.10 on a Lenovo ThinkPad computer, 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz , 32,0 GB RAM, Windows 10 64x.

Procedure

Instructions

Although the simulated participant did not receive instructions like the human participants, they are in fact receiving strict instructions from the experimenter through the python source code downloaded from GitHub (Mofrad et al., 2021). When performing machine instructed (coded) experiments it is important to understand how even slight changes in the source code can influence the results. In this experiment only one alteration was made to the source code used in Mofrad et al. (2021), which consisted of adjusting the coded protocol to match the protocol used in experiment 1.

MTS training

The protocol was simultaneous matching-to sample (MTS), identical to the procedure used in experiment 1. However, the training is done without the phase aimed at maintaining baseline using consequence thinning. The gamma damping parameter could possibly be interpreted as analogous to this phase of training, as it directly affects the simulated participant's ability to remember baseline relations during the test phase, but this remains untested.

Like experiment 1, the training is set up in blocks with 18 trials per block. Both the order of comparison stimuli and the order of trials in the block were randomly presented to the simulated agent.

MTS test

The same test protocol was used for both experiments (table 1)

Results

Establishing baseline

On average (for all 45 participants) it took 149 trials to establish baseline, with a standard deviation of 6; less than half the mean value. The standard deviation varied across groups, where α=0,001 and 0,05 and βh = 0,01 stood out with no variation in number of trials among the five participants in each group. On the other end, βt=3 had the highest within group standard deviation with 29 trials (table 4).

The maximum number of trials during training was 216. Almost one in three participants used 216 trials to establish baseline. On the other end we had five participants with 54 trials. These participants belong to the α = 0,01 group, with zero variation in number of trials across the five participants in the group (figure 6).

α conditions

This parameter controls how the network enhancement affects the trained network before the testing starts. Smaller values of α indicates strong baseline relations, as well as the symmetry and reflexivity relations after training, while higher values of α value reinforces the transitive and equivalence relations (Mofrad et al., 2021, p. 494). The 15 participants simulated with variation

in the α-condition exhibited small variation within each parameter value group (figure 7), with only α=0,1 having any variation in number of trials to establish baseline. For all 15 participants having three different values of α, the min-max interval is [54-90], with a standard deviation of 13,9 (table 4).

β_h conditions

This parameter controls how the model converts h-values to probabilities during training. The h-values are adjusted continuously as the EEPS algorithm progresses and are used to steer network exploration and the merging of nodes. These values are instrumental in ensuring the EEPS algorithm reaches a solution that is both correct and efficient. The parameter is also used for the network enhancement, by generating its input matrix. The higher this parameter is, the higher the probability if choosing the connection with the largest h-value (Mofrad et al., 2021, p. 494). Again, there is one parameter value (βh=0,01) exhibiting no inter-group variation in number of trials (figure 8). For the βh parameter groups, the number of trials increased compared to the number of trials in the α-groups. The min-max interval is [144,216], with a standard deviation of 30,0 (table 4).

βt conditions

The βt-parameter controls how the participant perform in a trial during testing. By decreasing the value of this parameter, the probability of forming any of the tested relations decreases as well (Mofrad et al., 2021, p. 510). For these groups the inter-group variation in number of trials is higher than for the other six groups (figure 9). For βt = 3 we find the largest inter-group standard deviation of all the nine groups. For all 15 participants in the βt-groups, the min-max interval was [144,216] with a standard deviation of 22,5 (table 4).

Equivalence class formation

As a group, the participants in experiment 2 did not attain the criterion of 90% correct trials during test with an average score of 82%, while they obtained a score of 89% when testing the baseline and symmetry relations. 18 out of the 45 participants (40%) formed equivalence classes. 32 out of 45 participants (70%) formed symmetry relations. 34 out of the 45 participants maintained their baseline relations (76%), (figure 10).

α conditions

Across the three α groups, one in three formed equivalence relations, with three out of five forming equivalence relations with α = 0,05. The variation in response patterns during equivalence testing was low, with a standard deviation of 0,03 point (out of the [0-1] test score scale) (table 4 and figure 11).

βh conditions

Across the three βh groups, there is large variation, having one parameter yielding no equivalence formation (βh = 0,01), while for the two other parameter values (βh =0,05, βh =0,1) all participants formed equivalence relations. An equal pattern was observed with both the baseline and symmetry relations tested, for βh=0,01 all the participants scored well below the criterion for relation formation (90%), while all participants in the other two groups scored well over 90% (table 4 and figure 12).

βt conditions

As for the βh groups, there is large variation in the βt groups, having two parameter values yielding no equivalence formation (βt = 3, βt =4), while for the βh = 5, three out of five participants formed equivalence relations. While the standard deviation for participants in βh

groups was higher (0,22), the standard deviation of 0,15 for the βt groups was more evenly distributed (table 4 and figure 13). This indicates a higher inter-group variation for the βt conditions than the βh conditions.

Discussion

We observe that within groups of simulated participants with the same parameter values, there was little variation in terms of number of trials to establish baseline. Parameter βt exhibits the largest variation between participants. Going from βt=3 to βt =4 decreased the standard deviation from 29 to 26 trials, while going from 4 to 5 decreased the standard variation to seven. The inverse pattern can be observed for βh, increasing the value for βh increases the standard deviation.

In terms of equivalence formation, the variation in responding between the nine groups is relatively high, with three groups having no participants forming equivalence relations, while there were two groups where all participants attained equivalence.

It is interesting to note that the parameter βh shows extreme values when set to 0.01, with no participants attaining a minimum of 0,9 score during equivalence testing. This underlines the comment from Mofrad et al. (2021), saying that it is extremely important for the model that this parameter value is chosen correctly. On the other hand, we observe that all participants in the groups with βh =0,5 and βh =1 formed equivalence relations, with test scores on equivalence higher than 95% across all 10 participants. These parameter values also yielded a larger variation in terms of trials needed to establish baseline relations during training.

The results for the parameter βh also highlight another important take away. One parameter value for βh was tested in the α group (βh=0,1, α = 0,05). As all other parameters were the same, the results from this α group can also be seen as an additional βh group testing the value βh=0,1. With βh=0,1 and α = 0,05 the mean number of trials was 72, which was lower than the mean number of trials for the lower parameter value βh = 0,05 (mean=205) and also lower than the higher parameter value βh= 1 (mean=166). In other words, the relationship between this parameter and number of trials is non-linear. For a mathematician this is obvious by taking a look at the algorithm. For researchers within behavior science on the other hand, this is not necessarily obvious, but it is beneficial for this line of research that we are aware of the mathematical complexity of the model we are working with. This avoids wasted resources and mistakes when trying to tune the model parameters to produce simulations that resemble human responding.

The results from experiment 2 are in line with the findings of the first connectionist study using an ANN (RELNET) to simulate human stimulus equivalence formations (Barnes & Hampson, 1993) in terms of proving that simulations can form equivalence relations comparable to humans. However, this paper is not comparable to experiment 2 in terms of training and testing protocols, nor is the same algorithm used. Similarly, the results from experiments 2 replicates Ninness et al. (2018) in terms of showing that it is possible for an ANN (EVA) to form equivalence relations, though not comparable in terms of protocol and algorithm. Furthermore, the variation in responding between the simulated participants are not reported in either paper.

In Ninness et al., (2019) they compare a human experiment against a simulated experiment, while both the protocol and algorithm from that of experiment 2. By comparing the

graphs in figure 6 and figure 8 visual analysis shows that the EVA algorithm used in the experiment produces similar response patterns between human and simulated participants, (Ninnes et al, 2019). This is not in line with the findings in this experiment, where we observe small variations in responding within the simulated participants, while the opposite is observed in the human participants.

General discussion

In experiment 1 the mean number of trials to establish the baseline relations was 158 while the mean for experiment 2 is 149. While the max-min interval for experiment 1 was [36,432], the same interval for experiment 2 was much smaller, being [54-216]. This is reflected in the difference in standard deviation between the experiments, being 162 in experiment 1 and only 61 in experiment 2 (figure 14).

When comparing the participants in experiment 1 with any of the nine groups of participants simulated in experiment 2, it becomes evident that the simulated participants, when using the same parameter values, are homogenous in terms of their responding. On the other hand, we find similarities in responding between experiment 1 and experiment 2 when comparing the relative test results from the for experiment 1 with the relative results for all the participants in experiment 2. In both experiment 1 and 2, the participants in average did not attain the criterion of 90% correct trials during equivalence test. For experiment 1 the mean score was 84% while the mean score in experiment 2 was 82%. While the participants in experiment 1 passed the baseline and symmetry test with score of 97% in average, the participants in

experiment 2 scored 89%, which is slightly less than the required 90%. Still the results are too far apart (the difference is not tested for statistical significance, though for further studies it would be recommended). Furthermore, the pattern of scoring similarly in baseline and symmetry relations (in average) is observable in both experiments.

While 60% formed equivalence classes in experiment 1, only 40% formed equivalence classes in experiment 2. However, some of the parameter values used in Experiment 2 caused extreme values, with all participants scoring less than 80% on all relations tested. When excluding these extreme values (βh =0,01, βt=3) the percentage of participants forming equivalence relations increases from 40% to 51%.

Experiment 2 demonstrated response patterns similar to human responding when looking at mean results, in keeping with Mofrad et al. (2020). When looking at inter-group participant responding however, Experiment 2 demonstrated how the variation for simulated participants with the same parameter value is much lower than we find in groups of participants in human stimulus equivalence experiments. Furthermore, experiment 2 demonstrated how the model is sensitive to the parameter values, where some settings yielded a result dissimilar to human responding, while other combinations of parameter values yield responding in keep with human responding in stimulus equivalence experiments.

Limitations

One possible weakness with Experiment 1 is that it was performed in a meeting room in the participants office environment. It is possible that work related stress and/or time constraints may have affected the participants performance. Another weakness is that the experiments were not performed at the same time for each participant, due to practical constraints.

Given the response patterns of participant 1854, experiment 1 might have benefitted from a higher mastery criterion, i.e., 95-100% (Arntzen, 2012). Furthermore, the heterogeneity of the participants in experiment 1 makes it is difficult to find earlier studies to compare the results with, if participant characteristics as well as training and testing characteristics has to be comparable. For the research question in this paper it does not signify, or rather, it could well be considered an advantage. As the aim is to explore whether human responding in equivalence experiments can be simulated accurately in general, the diversity of the participants is beneficial, assuming that the diversity implies that a larger portion of the natural variation found in human responding in the total population is represented.

In experiment 1 the protocol included a maintaining baseline phase. This is not replicated in the simulation; an assumption has been made that holding the gamma damping parameter constant throughout all simulations will "replace" the maintaining baseline phase done in the human experiment. This is a weakness because this assumption may not hold.

Although the results of these experiments indicate, like the equivalence projective simulation and connectionist studies have done before, that machine learning algorithms successfully simulate human responding during stimulus equivalence experiments, there are some caveats. With the RELNET studies there was a need for a pre-programming of the network to "compensate" for human learning history in order for the findings to hold (Lyddy & Barnes-Holmes, 2007; Lyddy et al., 2001; Tovar & Chávez, 2012; Vernucio & Debert, 2016). Furthermore, with the EVA studies, there are indications in the paper that the parameter tuning of the ANN-algorithm is done post human experiments, using the human responding as a guide for finding the correct parameter values (Ninness et al., 2019). In this experiment they add

another hidden layer to the EVA algorithm and introduces mastery criteria/stopping rule that stop training after reaching a MSE of 0,0006 or 100 training epochs (Ninness et al., 2019, p. 346). The reasoning behind this MSE threshold value was that they observed a mastery level in the simulated participant analogous to the mastery level with the human participants for this configuration of the EVA algorithm. This illustrates how the responding of the human participants is used to calibrate the EVA algorithm in order for it to be comparable to the human experiment. Had they chosen their parameter values more at random, perhaps they would have gotten different results. Although the line of research requires that we understand the role of the parameter value settings of the algorithm on the simulated results, we need more data to conclude.

Further work

The code for the EEPS simulations does not provide results on a lower granularity than what was reported in experiment 2 (number of trials, time, and average mastery level) for the training phase. It would be interesting to adapt the code so that we can extract the response patterns per participant. With that data it would be valuable to analyze the training phase trial by trial, as well as error patterns per participant and stimulus, similar to the analysis done for experiment 1. The possibility to study the simulated participant's response patterns during training and testing would make it possible for us to compare machine and human responding at a lower granularity, thus making it easier to understand if a random machine simulated participant is analogous in their responding to a random human participant.

Furthermore, it would be interesting to adapt the code so that the groups of participants are modelled with different sets of parameters within each group, for example that each

simulated participant is given a random value for each of the relevant parameters, within a relevant interval (i.e., the parameter values used results in simulations that are comparable to human responding, and not values that produces response patterns not found in humans). Adapting the code as suggested would possibly produce simulations with inter-group responding more in keep with observed responding in human experiments.

Furthermore, the simulation code could be improved to model consequence thinning. In these experiments that would have allowed the training protocols for human and simulated participants to be identical, thus eliminating a weakness in this experimental setup.

Application

A practical application of these findings could be to see how simulations and human experiments can be compared at a larger scale, as there is not yet enough data to support a conclusion that using enhanced equivalence projective simulation can replicate human responding accurately in a representative way. A possible solution to this is to add the EEPS simulation code to the MTS software that is used in human experiments. Each time a human participant performs an experiment with the MTS software, a series of simulations could be run with different combinations of parameter value settings, using all the same training and testing parameters as the human participant.

Conclusion

Experiment 2 demonstrated that the EEPS model can simulate the formation of stimulus equivalence through an MTS procedure. This is in keep with the findings in Mofrad et al. 2020 and 2021, that showed that advances in mathematical modeling methods can be applied to the

behavioral sciences, continuing, and improving the connectionists work on simulating emerging relations in humans using ANNs. While Mofrad et al. (2020) demonstrated group level responding similar to group results from known human studies, the inter-group results remained unexplored.

To be able to use results from simulated experiments in research on stimulus equivalence, we need to generalize the findings. In my opinion the only way forward is to continue to systematically compare responding between human participants and simulated participants. By gathering more comparable data, in order to understand the connection between the algorithm parameter values and human responding, we might able to verify if computer simulated experiments can provide an even better understanding of the underlying mechanisms of stimulus equivalence, by making experiments more efficient and increasing the experimental control.

References

Arntzen, E. (2012). Training and Testing Parameters in Formation of Stimulus Equivalence: Methodological Issues. European journal of behavior analysis, 13(1), 123-135. https://doi.org/10.1080/15021149.2012.11434412

Barnes, D., & Hampson, P. J. (1993). Stimulus equivalence and connectionism: Implications for behavior analysis and cognitive science. *The Psychological Record*, *43*(4), 617–638. https://doi.org/10.1007/BF03395903

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind*. Wiley-Blackwell.

Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, *2*(1), Article 400. https://doi.org/10.1038/srep00400

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Green, G., & Saunders, R. R. (1998). Stimulus Equivalence. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 229–262). Springer. https://doi.org/10.1007/978-1-4899-1947-2_8

Guresen, E., & Kayakutlu, G. (2011). Definition of artificial neural networks with comparison to other networks. *Procedia Computer Science*, *3*, 426–433. https://doi.org/10.1016/j.procs.2010.12.071

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Prentice Hall.

Lyddy, F., & Barnes-Holmes, D. (2007). Stimulus equivalence as a function of training protocol in a connectionist network. *The Journal of Speech and Language Pathology – Applied Behavior Analysis*, *2*(1), 14–24. https://doi.org/10.1037/h0100204

Lyddy, F., Barnes-Holmes, D., & Hampson, P. J. (2001). A transfer of sequence function via equivalence in a connectionist network. *The Psychological Record*, *51*(3), 409–428. https://doi.org/10.1007/BF03395406

Melnikov, A. A., Makmal, A., Dunjko, V., & Briegel, H. J. (2017). Projective simulation with generalization. *Scientific Reports*, *7*(1), Article 14430. https://doi.org/10.1038/s41598-017-14740-y

Mofrad, A. A., Yazidi, A., Hammer, H. L., & Arntzen, E. (2020). Equivalence projective simulation as a framework for modeling formation of stimulus equivalence classes. *Neural Computation*, *32*(5), 912–968. https://doi.org/10.1162/neco_a_01274

Mofrad, A. A., Yazidi, A., Mofrad, S. A., Hammer, H. L., & Arntzen, E. (2021). Enhanced equivalence projective simulation: A framework for modeling formation of stimulus equivalence classes. *Neural Computation*, *33*(2), 483–527. https://doi.org/10.1162/neco_a_01346

Ninness, C., Ninness, S. K., Rumph, M., & Lawson, D. (2018). The emergence of stimulus relations: Human and computer learning. *Perspectives on Behavior Science*, *41*(1), 121–154. https://doi.org/10.1007/s40614-017-0125-6

Ninness, C., Rehfeldt, R. A., & Ninness, S. K. (2019). Identifying accurate and inaccurate stimulus relations: Human and computer learning. *The Psychological Record*, *69*(3), 333–356. https://doi.org/10.1007/s40732-019-00337-6

Ninness, C., Yelick, A., Ninness, S. K., & Cordova, W. (2021). Predicting heuristic decisions in child welfare: A neural network exploration. *Behavior and Social Issues*, *30*(1), 194–208. https://doi.org/10.1007/s42822-021-00047-1

Sidman, M. (1992). Adventitious control by the location of comparison stimuli in conditional discriminations. *Journal of the Experimental Analysis of Behavior*, *58*(1), 173–182. https://doi.org/10.1901/jeab.1992.58-173

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Authors cooperative.

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior*, *74*(1), 127–146. https://doi.org/10.1901/jeab.2000.74-127

Sidman, M., Cresson Jr, O., & Willson-Morris, M. (1974). Acquisition of matching to sample via mediated transfer. *Journal of the Experimental Analysis of Behavior*, *22*(2), 261–273. https://doi.org/10.1901/jeab.1974.22-261

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. Matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22. https://doi.org/10.1901/jeab.1982.37-5

Steele, D., & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior*, *56*(3), 519–555. https://doi.org/10.1901/jeab.1991.56-519

Tovar, A. E., & Chávez, A. T. (2012). A connectionist model of stimulus class formation with a yes/no procedure and compound stimuli. *The Psychological Record*, *62*(4), 747–762. https://doi.org/10.1007/BF03395833

Tovar, Á. E., & Westermann, G. (2017). A neurocomputational approach to trained and transitive relations in equivalence classes. *Frontiers in Psychology*, *8*, Article 1848. https://doi.org/10.3389/fpsyg.2017.01848

Vernucio, R. R., & Debert, P. (2016). Computational simulation of equivalence class formation using the go/no-go procedure with compound stimuli. *The Psychological Record*, *66*(3), 439–449. https://doi.org/10.1007/s40732-016-0184-1

Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C. D., Batzoglou, S., & Leskovec, J. (2018). Network enhancement as a general method to denoise weighted biological networks. *Nature Communications*, *9*(1), Article 3108. https://doi.org/10.1038/s41467-018-05469-x

Table 1

Training and Test Protocol

Phase	**Type**	**Relation**	**Programmed consequences (%)**	**Minimum trials**	**Mastery criterion (%)**
MTS Training	Baseline	A1C1, A2C2, A3C3 B1C1, B2C2, B3C3	100	36	90
MTS test	Baseline	A1C1, A2C2, A3C3 B1C1, B2C2, B3C3	0	54	90
	Symmetry	C1A1, C2A2, C3A3 C1B1, C2B2, C3B3			
	Equivalence	A1B1, A2B2, A3B3 B1A1, B2A2, B3A3			

Table 2

Test Results - Equivalence Class Formation per Participant

#P	Baseline%	Symmetry%	Equivalence%	Equivalence attained
1851	100	100	94	Y
1852	100	100	94	Y
1853	94	89	83	N
1854	89	94	50	N
1855	100	100	100	Y
Mean	97	97	84	N

Table 3

The Parameter Conditions used in Experiment 2

Condition	α	K	βh	βt	γ	P#
1	0,01	1,0	0,1	4,0	0,001	1-5
2	0,05	1,0	0,1	4,0	0,001	6-10
3	0,10	1,0	0,1	4,0	0,001	11-15
4	0,05	1,0	0,01	4,0	0,001	16-20
5	0,05	1,0	0,5	4,0	0,001	21-25
6	0,05	1,0	1	4,0	0,001	26-30
7	0,05	1,0	0,1	3,0	0,001	31-35
8	0,05	1,0	0,1	4,0	0,001	36-40
9	0,05	1,0	0,1	5,0	0,001	40-45

Note. The network enhancement parameter is set to "False", that is DNS was used as network enhancement method (Mofrad et al., 2021).

Table 4

Training Results per Parameter Value Group

Parameter	Parameter value	Mean #Trials	#Trial min-max interval	#Trial standard deviation
α	0,01	54	[54 - 54]	0
α	0,05	72	[72 - 72]	0
α	0,1	86	[72 - 90]	7
βh	0,01	144	[144 - 144]	0
βh	0,5	205	[198 - 216]	9
βh	1	166	[144 - 198]	26
βt	3	194	[144 - 216]	29
βt	4	205	[162 - 216]	22
βt	5	212	[198 - 216]	7
Total		149	[54-216]	61

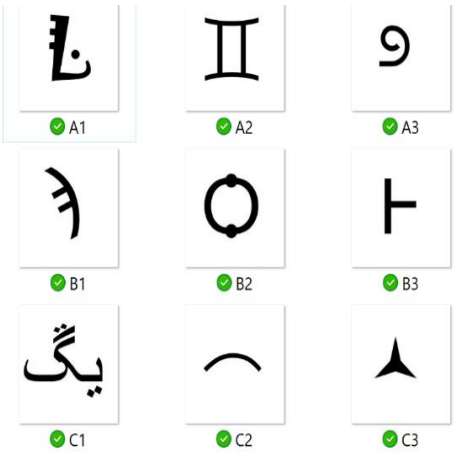
Table 5

Test Results, Equivalence Relations

Parameter	Parameter value	Mean BSL (%)	Mean SYM (%)	Mean EQ (%)	Equivalence attained (%)	EQ Standard deviation (%)
α		**93,5**	**93,2**	**88,9**	**33,3**	**3,0**
	0.01	93,2	93,0	88,1	20,0	4,4
	0.05	94,3	94,1	91,1	60,0	1,5
	0.1	92,9	92,6	87,4	20,0	2,4
βh		**86,5**	**86,4**	**79,8**	**66,7**	**22,0**
	1,00	96,3	96,2	96,1	100,0	0,0
	0.01	67,4	67,4	48,1	0,0	1,4
	0.5	95,8	95,6	95,1	100,0	0,0
βt		**86,4**	**86,0**	**78,6**	**20,0**	**15,0**
	3,00	76,3	75,9	66,8	0,0	14,3
	4,00	88,1	87,4	79,7	0,0	13,1
	5,00	94,9	94,6	89,4	60,0	6,6
Total		**88,8**	**88,5**	**82,4**	**40,0**	**16,0**

Figure 1

The set of Stimuli Used in Experiment



Note. Overview of stimuli used in the experiment. The numbers 1–3 indicate the three experimenter-defined classes, and the letters in the left column indicate the three members. The letter-number combination was not visible for the participants.

Figure 2

Number of Correct Trials by Block in the Training Phase, by Participant

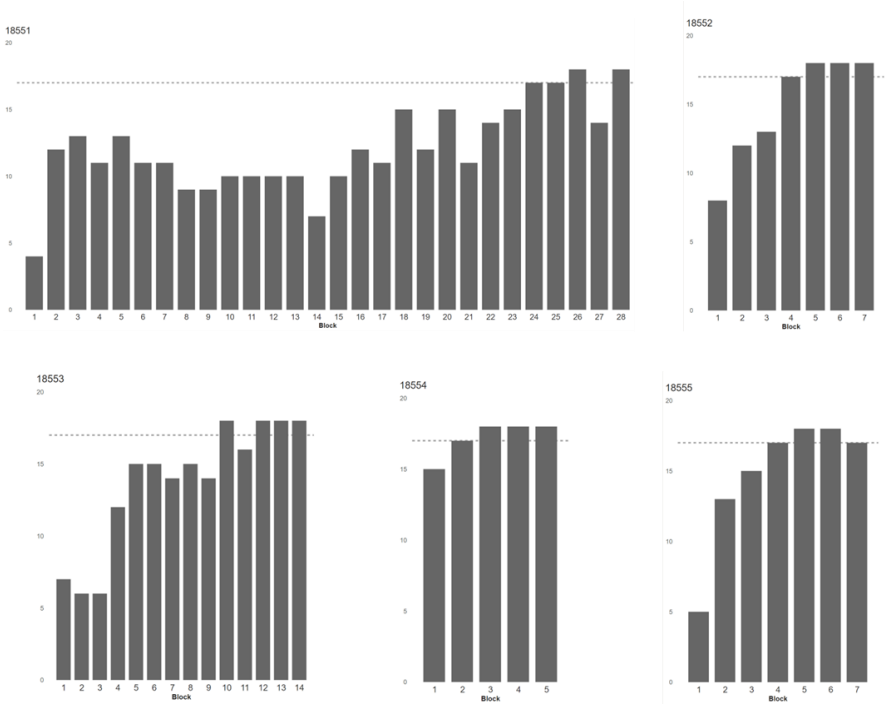
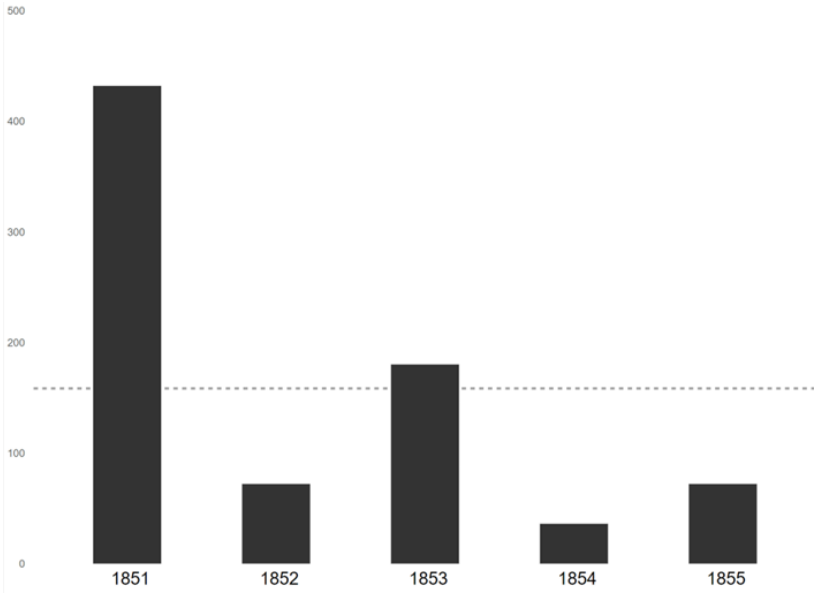


Figure 3

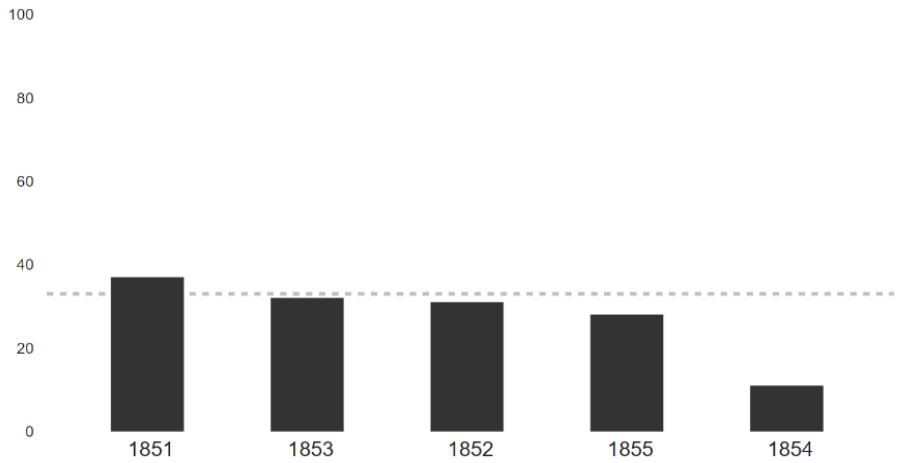
Number of Trials to Establish Baseline Relations, per Participant



Note. Group mean in grey, dashed line.

Figure 4

Error Rate When Establishing Baseline Relations, per Participant



Note. Group mean in grey, dashed line.

Figure 5

Number of Errors per Sample Stimulus, per Participant

P #: 18551 Selected

Sample	A1	C1	C2	C3	Total
A1			8	5	13
A2		8		23	31
A3		7	12		19
B1			23	19	42
B2	1	37		14	52
B3		5	4		9
C1					
C2					
Total	1	57	47	61	166

P #: 18552 Selected

Sample	A1	C1	C2	C3	Total
A1					
A2		1		3	4
A3		1	1		2
B1			2	3	5
B2		3		4	7
B3	1	1	3		5
C1					
C2					
Total	1	6	6	10	23

P #: 18553 Selected

Sample	A2	B1	C1	C2	C3	Total
A1				1	7	8
A2		1	3		10	14
A3			4	3		7
B1	2			4	9	15
B2			12		6	18
B3			2			2
C1						
C2		2				2
Total	2	3	21	8	32	66

P #: 18554 Selected

Sample	A1	A2	B1	B2	C1	C2	Total
A1				1			1
A2			3		3		6
A3							
B1		3					3
B2	2				1		3
B3					1	1	2
C1				1			1
C2							
Total	2	3	3	2	5	1	16

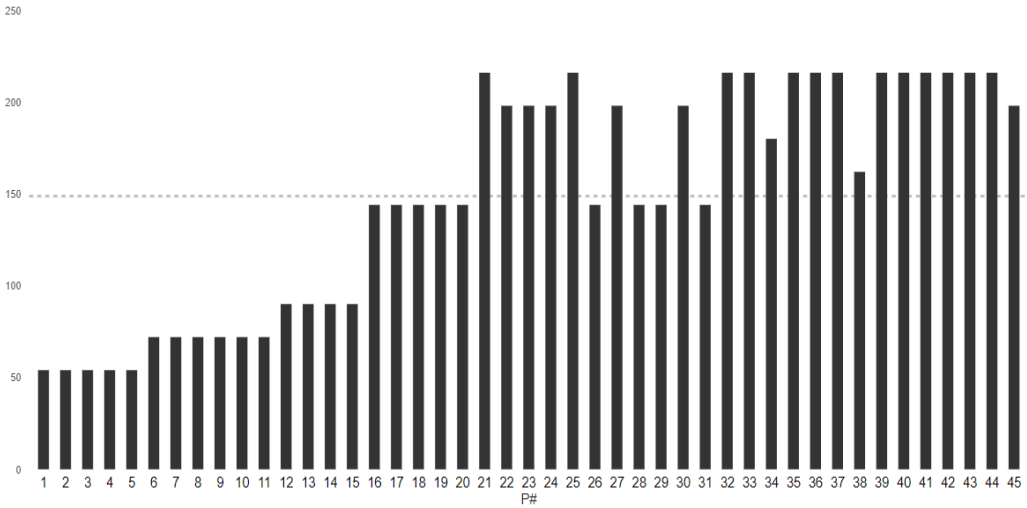
P #: 18555 Selected

Sample	C1	C2	C3	Total
A1		1	5	6
A2	1		2	3
A3	1			1
B1		2	3	5
B2	3		3	6
B3	1	1		2
C1				
C2				
Total	6	4	13	23

Note. Sample stimulus in columns and the erroneously selected stimulus in rows.

Figure 6

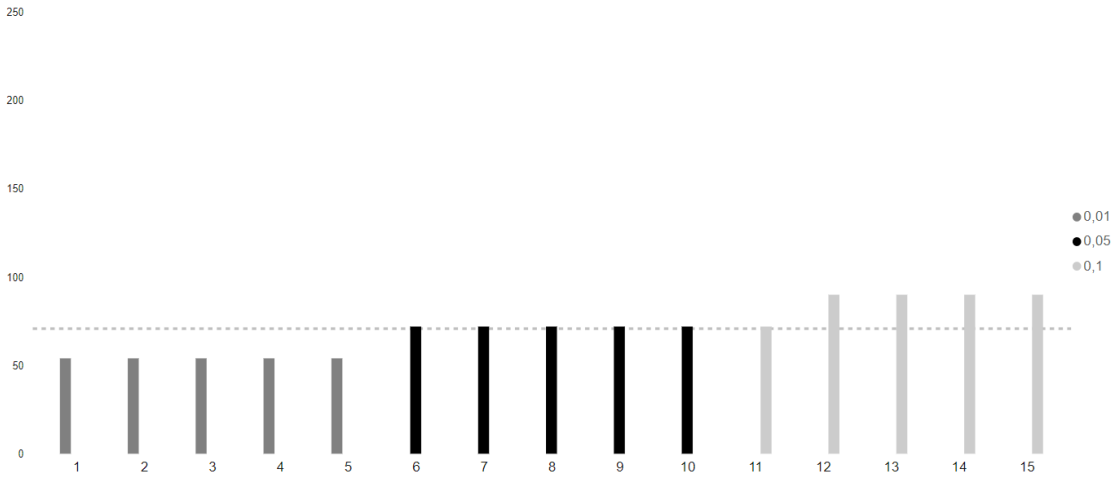
Number of Trials to Establish Baseline, per Participant



Note: Grey, dashed line depicts the mean.

Figure 7

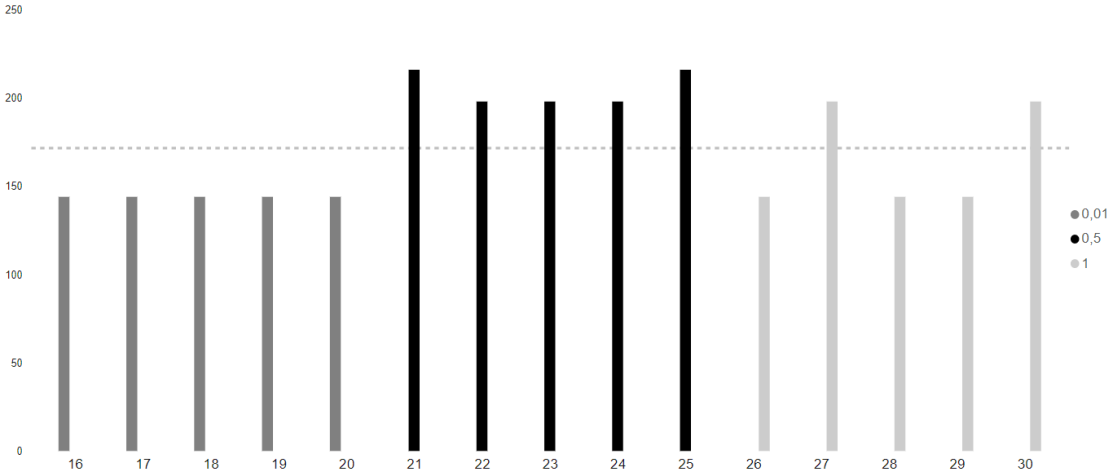
α *Conditions - Number of Trials and sum of Time per Participant, by Parameter Value*



Note. *The grey dashed line depicts the group mean.*

Figure 8

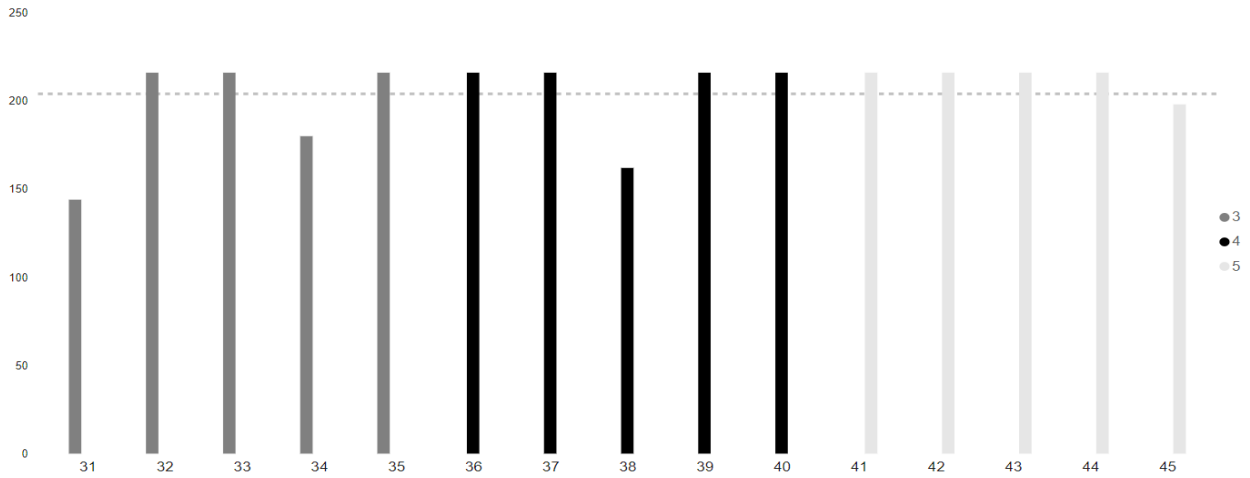
βh Conditions - Number of Trials and sum of Time per Participant, by Parameter Value



Note. Grey dashed line depicts the group mean value.

Figure 9

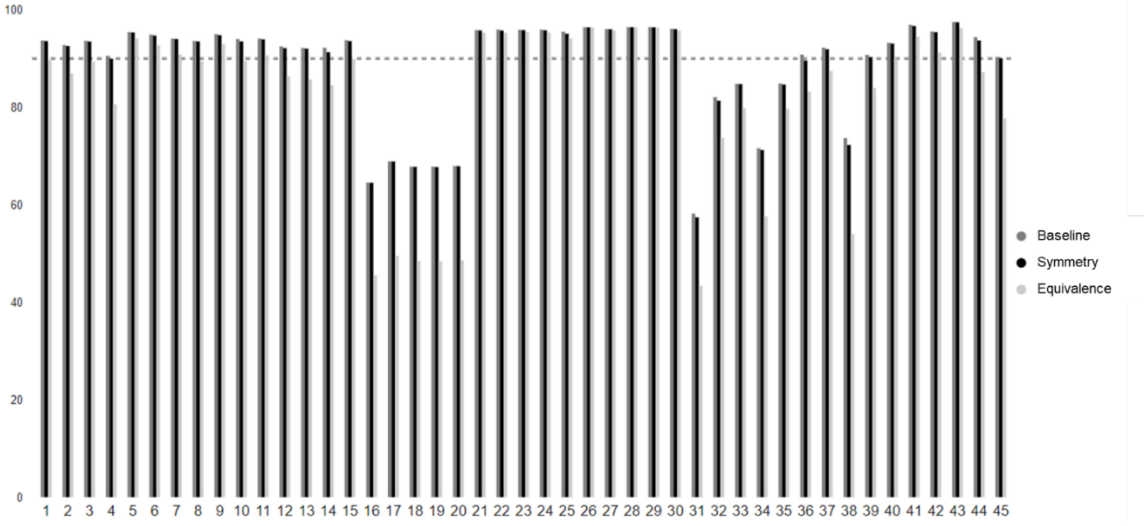
βt Conditions - Number of Trials and sum of Time per Participant, by Parameter Value



Note. Grey dashed line depicts the group mean value.

Figure 10

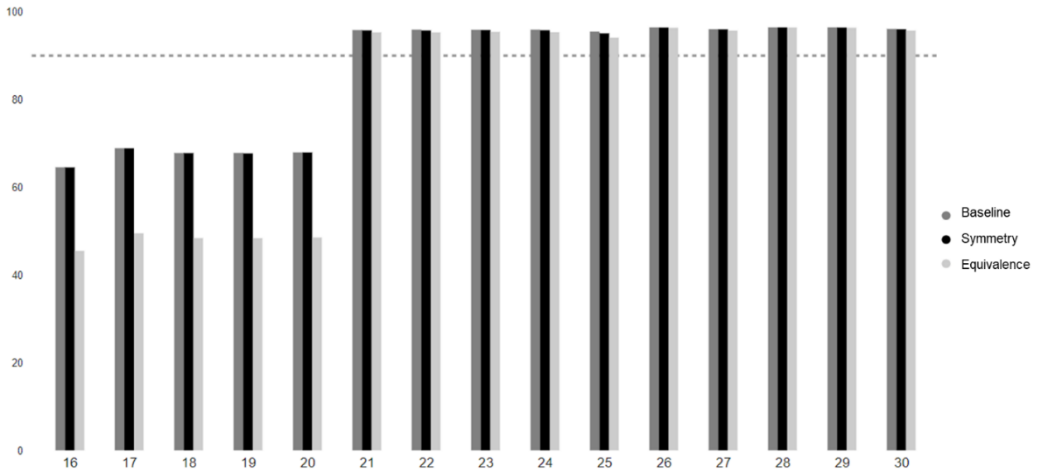
Test Results for Baseline, Symmetry and Equivalence Relations, by Participant



Note. Score of 90 (equivalence formation criterion) as a gray dashed line. See table 3 for the parameter conditions for each participant.

Figure 11

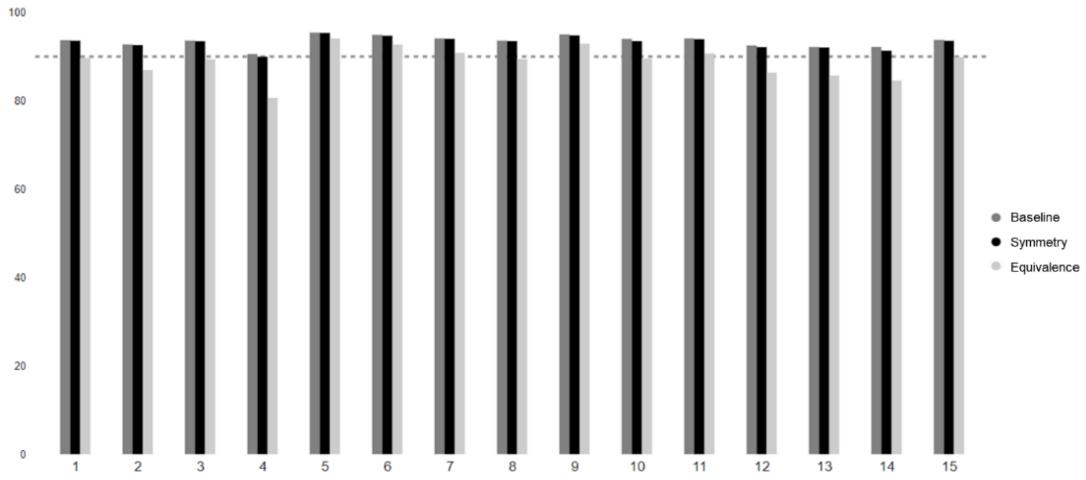
Test Results for the 15 Participants With Different α Conditions



Note. Score of 90 (equivalence formation criterion) as a gray dashed line. See table 3 for the parameter conditions for each participant.

Figure 12

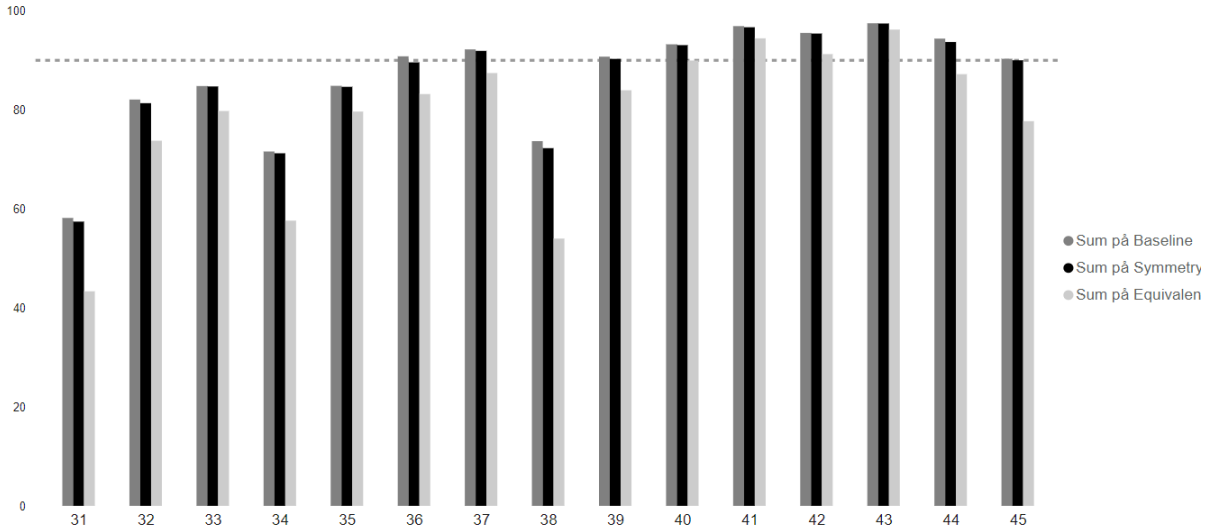
Test Results for the 15 Participants with Different βh Conditions



Note. Score of 90 (equivalence formation criterion) as a gray dashed line. See table 3 for the parameter conditions for each participant.

Figure 13

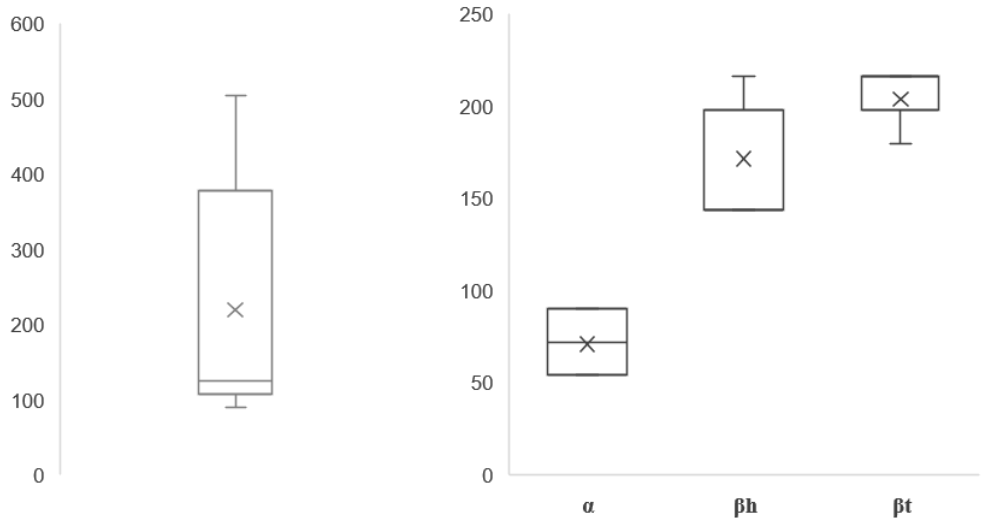
Test Results for the 15 Participants With Different βt Conditions



Note. Score of 90 (equivalence formation criterion) as a gray dashed line. See table 3 for the parameter conditions for each participant.

Figure 14

Boxplot of the Number of Trials Needed to Establish Baseline



Note. Experiment 1 (left) and Experiment 2 (right, by parameter type). The x indicates the group mean.

Etisk refleksjonsnotat

Det ble ikke innhentet personopplysninger til eksperimentet utover epostadresse, fornavn og etternavn i forbindelse med gjennomføring av eksperimentet. Alle fem deltakere ble presentert med et dokument som beskrev deres rettigheter. Dokumentet beskrev at eksperimentet ikke innbar noen form klinisk behandling eller noe fysisk ubehag av noe slag. Videre ble det beskrevet at datainnsamlingen i forsøket er anonym slik at ingen individuelle data kan spores tilbake til en bestemt deltaker. Ingen vil kunne identifisere enkeltpersoner fra studien ved en eventuell publisering eller offentliggjøring av masteroppgaven. Dokumentet oppga også vilkårene for å trekke tilbake samtykket gitt til å bruke dataene som er produsert under eksperimentet, det vil si at de kunne trekke seg fra eksperimentet når som helst under eller etter eksperimentet, uten konsekvenser. De ble også informert om at en debriefing-sesjon ville bli gitt etter at eksperimentet var fullført. De fikk kontaktinformasjonen min samt min veileders kontaktinformasjon.

I etterkant av eksperimentet ble hver deltaker vist sine resultater og ble forklart formålet med eksperimentet. Fra et etisk perspektiv var det viktig å få frem at prestasjon i seg selv ikke var viktig for forskningsspørsmålet, men prosessen.

Videre var jeg nøye med at navn og resultater ikke var koblet sammen i materiale oversendt min veileder, og at det dermed kun var meg som hadde kobling mellom responsdata og deltakernavn. Denne listen ble også slettet av meg med en gang eksperimentene var gjennomført.

I denne oppgaven er det fokus på simulering av menneskelig respondering ved hjelp av maskinlæring. Dette stiller et viktig etisk spørsmål, nemlig om de etiske konsekvensene av å prøve å simulere mennesker ved hjelp av maskiner. Det er mange som mener at menneskets

streben etter kunstig intelligens, smarte roboter, vil ha uante konsekvenser for menneskeheten. Man ser også i dag betenkeligheter rundt teknologiens utvikling, med utvikling av autonome, «selvtenkende» våpen for eksempel.

På en annen side er det alltid knyttet usikkerhet til teknologiske fremskritt, og man kan argumentere for at den klimakrisen vi står ovenfor i dag kan tilskrives industrialiseringen i de vestlige land på begynnelsen av 1800-tallet. Denne står samtidig for en forbedring av levekår for «mannen i gata» som mangler historisk sidestykke. Einsteins forskning som var banebrytende for fysikken førte også til utvikling av atombomber.

Vitenskapelig forskning som springer ut av et ønske om å bringe fagfeltet videre, i alt fra forskning på atomer til kreft, er i mine øyne etisk forsvarlig, så fremt det skjer i henhold til lover, regler og vitenskapelige prinsipper. Når vi i dag har de teknologiske rammene til å bruke datamaskiner for å forstå grunnleggende menneskelige atferdsprinsipper så mener jeg at dette er en etisk forsvarlig forskningsretning å gå i. Dette til tross for at mulighetene for at denne type forskning potensielt kan ha store, negative konsekvenser slik vi ser med for eksempel oppfinnelsen av dampmaskinen og splitting av atomer. Jeg håper at den også vil ha store positive konsekvenser som muligens vil kunne avveie for de negative.