Linköping Studies in Science and Technology
Thesis No. 2264

Empirical Studies in Machine Psychology

Robert Johansson

LINKÖPING UNIVERSITY

Division of Human Centered Systems
Department of Computer and Information Science (IDA)
Linköping University, SE-581 83 Linköping, Sweden

Linköping 2023

Add a description of the front page here, if necessary.

ISBN 978-91-7929-505-9
ISSN 0345-7524

Printed by LiU-Tryck 2023

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Populärvetenskaplig sammanfattning

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Preface

Text for preface.

Acknowledgements

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Contents

CHAPTER 1

Introduction

CHAPTER 2

Background and Preliminaries

In this chapter, I provide the necessary background for understanding the results from the thesis. First, I provide an introduction to the field of Artificial General Intelligence (AGI). Then, I describe a particular AGI system, NARS, that will be used as the subject of study. After that, I describe OpenNARS for Applications (ONA), the particular NARS implementations that I have used in all studies. Finally, I introduce Learning Psychology, that will be a foundational framework, that clarifies the subject matter of the empirical results in this thesis.

2.1 Artificial General Intelligence

The birth of Artificial Intelligence (AI), as we know it today, is by many agreed on to have been in 1955, when John McCarthy decided to organize a workshop on the topic of building thinking machines. In the invite letter, McCarthy wrote:

We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that ***every aspect of learning*** *or any other feature of intelligence can in principle be so* ***precisely described that a machine can be made to simulate it****. An attempt will be made to find how to make machines* ***use language, form abstractions and concepts, solve kinds of problems now reserved for humans****, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. (Boldface added)*

The reason why I mention this workshop invitation here, is to highlight a few facts. First, AI clearly aimed from the start to target ***every aspect of learning*** (put differently: all types of learning). Second, already at this stage, there was a call for precise definitions of terms (with the aim for this learning to be possible to study with machines). Third,

2.2 NARS

2.3 OpenNARS for Applications

2.4 Learning Psychology

Learning psychology, can be seen as the study of *ontogenetic adaptation* [1]. That is, the adaptation of an individual organism to its environment during its lifetime. In line with this, learning in itself can be defined as *changes in the behavior of an organism that are the result of regularities in the environment of that organism* [1].

CHAPTER 3

Operant Conditioning

Given the perspective on AGI taken in this thesis, that we are defining AGI as all types of learning that a human can do, we will use this chapter to investigate one particular form of learning, operant conditioning.

3.1 Operant Conditioning

Operant conditioning is a type of learning that can be observed in most species, from fruit flies [2], invertebrates [3], to rats [4] and human beings [5].

3.2 OpenNARS for Applications

This chapter uses a sensorimotor version of the NARS system OpenNARS for Applications (ONA) [6]. More specifically, this was a version of ONA compiled with the parameter `SEMANTIC_INFERENCE_NAL_LEVEL` was set to `0`, which means that only NAL layers 6–8 were available. This means that the system could only do sensorimotor inference (procedural and temporal reasoning), but no semantic inference (declarative reasoning). The NARS rules are described in Listing 3.1.

```
// Rule 1: Projecting to current time
{Event a.} |- Event a. Truth_Projection

// Rule 2: Forming a temporal sequence of two events
{Event a., Event b.} |- Event (a &/ b). Truth_Intersection

// Rule 3: Forming an implication between two events
```

```
{Event a., Event b.} |- Implication <a =/> b>. Truth_Eternalize(
    Truth_Induction)

// Rule 4: Revision of an implication
{Implication <a =/> b>., <a =/> b>.} |- Implication <a =/> b>.
    Truth_Revision

// Rule 5: Subgoal derivation v1
{Event b!, Implication <a =/> b>.} |- Event a! Truth_Deduction

// Rule 6: Subgoal derivation v2
{Event (a &/ b)!, Event a.} |- Event b! Truth_Deduction

// Rule 7: Revision or Choice (dependent on evidental overlap)
{Event a!, Event a!} |- Event a! Truth_Revision or Choice

// Rule 8: Temporal deduction
{Event a., Implication <a =/> b>.} |- Event b.  Truth_Deduction
```

Listing 3.1. NARS Rules from Layer 7 and 8

3.3 Method

ONA was configured with the following settings.

```
*babblingops=2
*motorbabbling=0.2
*setopname 1 ^left
*setopname 2 ^right
*volume=100
```

This means that the system was configured to have two operators `^left` and `^right`, and an initial chance of motor babbling set to 20%. The `volume` parameter was set to a maximum value of 100, for the entire log file of the NARS reasoning process to be available for inspection.

The experiment consisted of three phases: Baseline assessment, Training (with feedback), and Testing (without feedback). In all phases, training and testing were done in blocks of trials. Given the experimental design, only two possible trials were possible. Either, that A1 was to the left, and A2 to the right, or, that A2 was to the left, and A1 to the right. A block contained twelve trials, with the two possible trials each presented six times each in random order.

A typical task was given to ONA in the form of temporal statements:

```
<A1 --> [left]>. :|:
<A2 --> [right]>. :|:
G! :|:
```

Given the task description of "always select A1, independent of location", the experimental design encoded a `^left` response by ONA when `<A1 --> [left]>` as "Correct", and similar for ONA responding `^right` when `<A1 --> [right]>`. Importantly, this "Correct" is from the perspective of the experiment. A response from ONA is always "Correct" from the perspective of the system itself.

The dependent variable of the experiment was set to the rate of correct responding per block of 12 trials. A detailed description of the three phases follow.

1. **Baseline.** The baseline consisted of three blocks. No feedback was given during this phase. This phase functioned to establish a baseline probability of responding correct according to the experimental design. It was expected that ONA would respond correctly by chance in 50% of the trials.

2. **Training.** The system was trained on a set of three blocks. Feedback was given as `G. :|:` and `G. :|: {0.0 0.9}` when the system was correct and incorrect, respectively.

3. **Testing.** In the final phase, the system was tested without feedback in three blocks. Testing was conducted with the same contingencies that was presented during training.

3.4 Results

During baseline, the amount of correct trials ranged between 30 and 60% during the three blocks, indicating that no learning took place (no changes in behavior was due to regularities in the environment). In the training, NARS was 100% correct already in the second out of three training blocks. In the final phase, where NARS was tested without feedback, the system responded 100% correct in all three blocks. Given this, the results in total indicate that the changes in behavior from baseline to testing, were due to the regularities between the system's own behavior and other events, as presented during the training phase. The results are illustrated in Figure 3.1.

3.5 Discussion

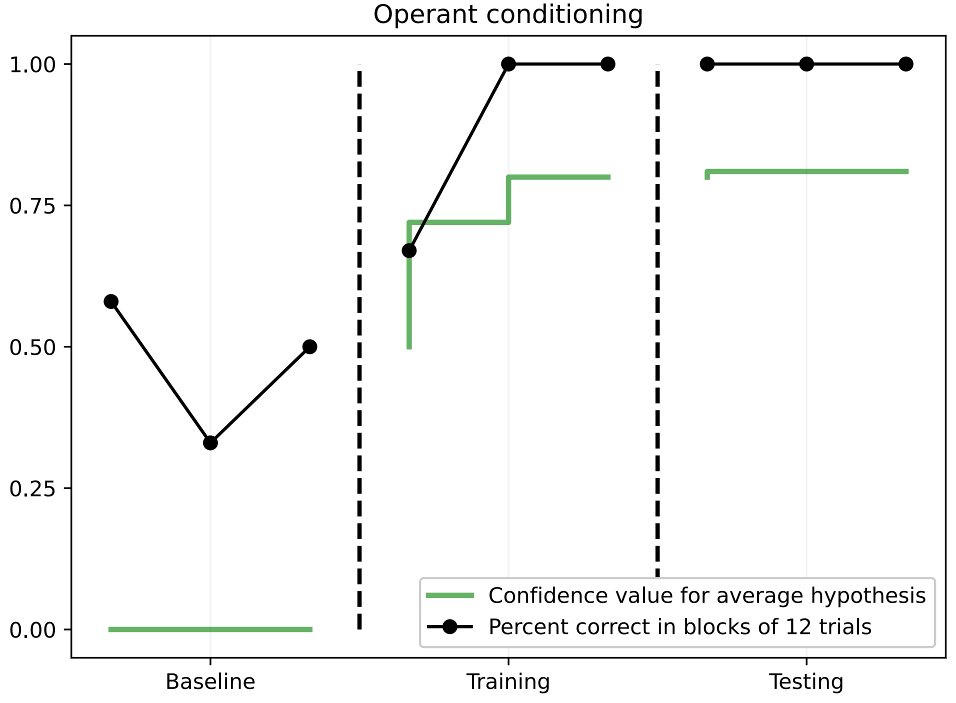

Figure 3.1. Learning in the form of operant conditioning. Dots illustrate the percent of correct in blocks of 12 trials. The solid line shows the NARS confidence value for specific hypotheses.

CHAPTER 4

Generalized Identity Matching

The purpose of this chapter is to investigate an instance of complex learning with NARS. More specifically, it is about investigating if NARS can do *generalized identity matching*, which is an example of a form of *relational learning*. It is also a form of learning as an effect of *metaregularities*.

4.1 Introduction

At the foundation of any intelligent system is the ability to form general concepts about the relationships among objects or other types of stimuli. These concepts are essential to a wide range of tasks. A fundamental relational concept is the identity concept, i.e., the ability to respond to the identity relationship among stimuli.

One way to study the use of the identity concept is with *identity matching* in the matching-to-sample (MTS) context. In these experiments, participants are presented with a sample stimulus and pairs of comparison stimuli and are asked to decide which comparison is identical to the sample. These experiments are thought to reveal the extent to which a subject applies the identity concept as part of the decision making. While these experiments typically are done with visual stimuli, there are no limits on the type of stimuli that can be matched regarding sensory modality.

An even more sophisticated ability, *generalized identity matching*, can be tested for using the same MTS setup. After having been trained to match by identity over a set of trials, the subject is presented with a novel sample stimulus and novel comparisons. Generalized identity matching is demonstrated if the subject can transfer the identity concept to these new stimuli. Evidence of this ability has been reported in a number of non-human species including sea lions [7], rats

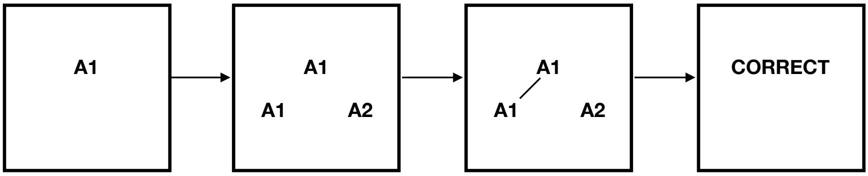

Figure 4.1. A typical trial in the identity matching-to-sample task. First, the sample is presented at the top (leftmost panel). Then, two comparison stimuli are presented (next panel). The experimental subject then indicates a choice between either the left or right option. Finally, the subject receives feedback if the choice was correct or not (rightmost panel).

[8] and pigeons [9]. A typical trial in the generalized identity matching-to-sample task is illustrated in Figure 4.1. A video of a sea lion carrying out the task can be found in [10].

While this might seem like a trivial task, it can be seen as a minimal example of general-purpose learning. In other words, being able to complete this task seems like a necessary (but not sufficient) critiera for an AGI system. Hence, this task would be interesting to demonstrate for any AGI system.

In this study, we report on a study of generalized identity matching in the AGI-system OpenNARS for Applications. The relevance for other AGI research will be discussed.

4.2 Method

4.2.1 OpenNARS for Applications

We used OpenNARS for Applications (ONA) [6], a highly effective implementation of the Non-Axiomatic Reasoning System (NARS) [11]. Importantly, the parameter `SEMANTIC_INFERENCE_NAL_LEVEL` was set to 0, which means that only NAL layers 6-8 were available. This means that the system could only do sensorimotor inference (procedural and temporal reasoning), but no semantic inference (declarative reasoning). In a way, this could be called an animal-like version of ONA.

4.2.2 Identity Match-to-sample task in NARS

The identity match-to-sample task was presented as temporal Narsese statements (as indicated by the `:|:` markers below). An arbitrary goal event `G! :|:` was presented at the end to trigger the execution of one of the two procedural operations `^left` and `^right` (through motor babbling or a decision). During training, feedback was given in the form of `G. :|:` (meaning to reinforce a correct choice) or `G. :|: {0.0 0.9}` (to indicate that the system had conducted an incorrect choice). Between each trial, 100 time steps was entered, by feeding `100` to ONA.

```
<A1 --> [sample]>. :|:
```

```
<A1 --> [left]>. :|:
<A2 --> [right]>. :|:
G! :|:
```

An explanation of the Narsese follows. The first three lines are inheritance statements with properties on the right-hand side, indicating that the stimuli ($A1$, $A2$), are either on the left, right, or are the sample.

4.2.3 Experimental setup

ONA was set to have two operators `^left` and `^right`, and an initial chance of motor babbling to 20%. The experiment consisted of four phases: Baseline assessment, Training (with feedback), Testing for identity (without feedback), and Testing for generalized identity (without feedback). In all phases, training and testing were done in blocks of trials. One trial could for example be that $A1$ was the sample and $A1$ and $A2$ were the left and right options, respectively. A block contained twelve trials, with the four trials possible with $A1$ and $A2$ as samples, each presented three times in random order.

1. **Baseline** During the baseline assessment, which was two blocks, no feedback was given. This phase was included to establish a baseline probability of responding correct. It was expected that the system would respond correctly by chance in 50% of the trials.

2. **Training** Then, the system was trained on a set of six blocks. Feedback was given when the system was correct (for example matching `<A1 --> [sample]>` to `<A1 --> [right]>` by doing `^right`), and when not correct.

3. **Testing for identity** The system was then tested (without feedback) on two blocks, with the contingencies that previously had been trained. If the system was correct on all trials in all blocks, the experiment continued with the next phase.

4. **Testing for generalized identity** Finally, the system was tested (without feedback) on two blocks with trials containing novel stimuli ($X1$ and $X2$) the system hadn't seen before.

4.2.4 NARS examples from the training phase

A few example trials from the training session follows. Let's say that the system was exposed to the following NARS statements:

```
<A1 --> [sample]>. :|:
<A2 --> [left]>. :|:
<A1 --> [right]>. :|:
G! :|:
```

If it is early in the training, NARS might use motor babbling to execute the `^right` operation. Since this is considered correct in the experiment, the feedback

`G. :|:` would be given to NARS, followed by 100 time steps. Only from this single interaction, NARS would form both a specific and a general hypothesis:

```
<(((<A1 --> [sample]> &/ <A1 --> [right]>) &/ ^right) =/> G>.
// frequency: 1.00, confidence: 0.15

<(((<#1 --> [sample]> &/ <#1 --> [right]>) &/ ^right) =/> G>.
// frequency: 1.00, confidence: 0.26
```

Importantly, after this single trial, NARS would also form simpler hypothesis such as:

```
<(<A1 --> [right]> &/ ^right) =/> G>.
// frequency: 1.00, confidence: 0.21

<(<A1 --> [sample]> &/ ^right) =/> G>.
// frequency: 1.00, confidence: 0.16
```

This means, that if the same trial was to be presented again (all four possible trials will be presented three times in a block of twelve trials), NARS would respond `^right` again, but the decision being based on a simple hypothesis such as `<(<A1 --> [right]> &/ ^right) =/> G>`, since that hypothesis has the highest confidence value.

Let's say, that within the same block of 12 trials, the next trial to be presented to NARS was the following:

```
<A1 --> [sample]>. :|:
<A1 --> [left]>. :|:
<A2 --> [right]>. :|:
G! :|:
```

NARS would initially respond `^right`, with the decision being made from the simple hypothesis `<(<A1 --> [sample]> &/ ^right) =/> G>`. This would be considered wrong in the experiment, and the feedback `G. :|: {0.0 0.9}` would be given to NARS. This would lead to negative evidence for the simple hypothesis. If the same trial was presented again, NARS would then likely resort to motor babbling that could execute the `^left` operation. Over repeated trials with feedback, the simpler hypotheses would get more negative evidence, and the confidence value of the target hypotheses (specific and general) would increase.

4.3 Results

During baseline, the amount of correct trials ranged between 0 and 50% during the two blocks, indicating that no learning happened. In the training phase, NARS was 100% correct after six blocks. When being tested on the same symbols without feedback (Phase 3: identity matching), NARS was 100% correct during both blocks. Also, NARS was 100% correct on novel stimuli (Phase 4), demonstrating generalized identity matching.

Across the six training blocks, the average confidence value for the four specific target hypotheses such as

```
<((<A1 --> [sample]> &/ <A1 --> [left]>) &/ ^left) =/> G>
```

went from 0.17 to 0.86. For the general target hypotheses (please note the #1 variable which could be substituted with specific terms) such as

```
<((<#1 --> [sample]> &/ <#1 --> [left]>) &/ ^left) =/> G>
```

the average confidence value went from 0.34 to 0.96. This also confirms that the generalized hypotheses reached more evidence in total than the specialized one, which was as expected, as it is not tied to $A1$ (it could also be substituted by other stimuli).

Importantly, in the final phase (generalized identity matching), NARS made its decisions based on the general hypotheses that had developed in confidence during the training.

The results from the four phases are illustrated in Figure 4.2.

4.4 Discussion

The aim of this study was to study if a minimal version of NARS without declarative reasoning could do generalized identity matching. NARS learned this very quickly, as demonstrated by the experiments carried out in this study.

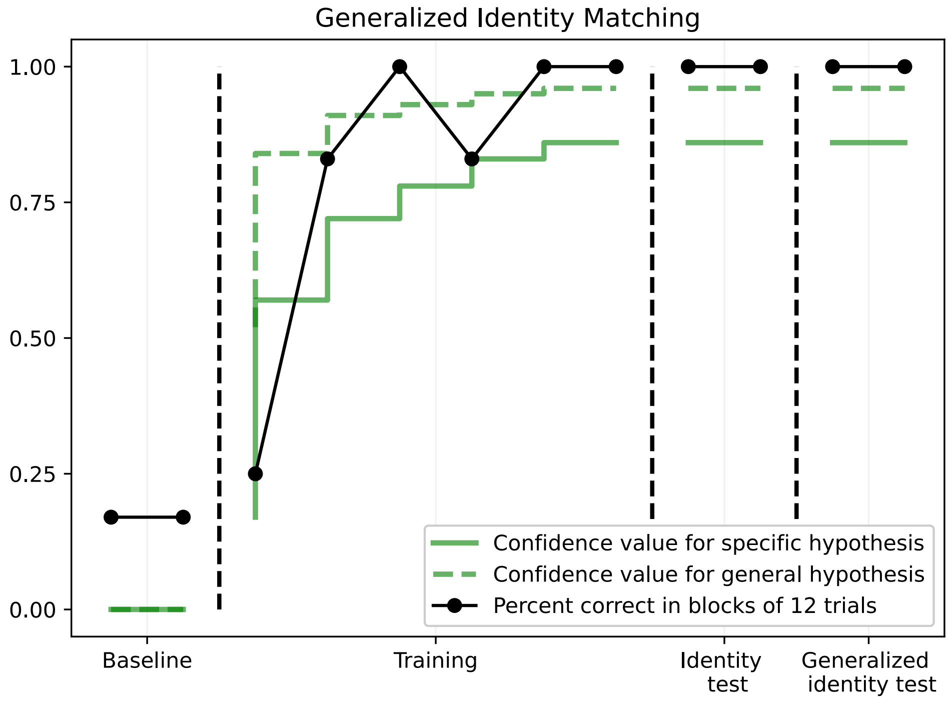

Figure 4.2. Learning generalized identity matching in the Match-to-sample task. Dots illustrate the percent of correct in blocks of 12 trials. The solid line shows the NARS confidence value for specific hypotheses (identity matching), while the dashed line illustrates the NARS confidence in general hypotheses (generalized identity matching).

CHAPTER 5

Stimulus Equivalence

In this chapter, I introduce a few experiments that all involve complex learning. Moreover, they could all be said to be related to the phenomenon called *stimulus equivalence*. This chapter involves extending ONA with a new rule, the *contingency entailment rule*. This rule is inspired by statement-level inference

5.1 Introduction

5.2 Method

5.3 Results

5.4 Discussion

CHAPTER 6

Arbitrarily Applicable Relational Responding

6.1 Introduction

6.2 Method

6.3 Results

6.4 Discussion

CHAPTER 7

Conclusions and Future Work

7.1 Conclusions

7.2 Future Work

Bibliography

[1] Jan De Houwer, Dermot Barnes-Holmes, and Agnes Moors. "What is learning? On the nature and merits of a functional definition of learning". In: *Psychonomic bulletin & review* 20 (2013), pp. 631–642.

[2] HA Mariath. "Operant conditioning in Drosophila melanogaster wild-type and learning mutants with defects in the cyclic AMP metabolism". In: *Journal of insect physiology* 31.10 (1985), pp. 779–787.

[3] Björn Brembs. "Operant conditioning in invertebrates". In: *Current opinion in neurobiology* 13.6 (2003), pp. 710–717.

[4] Burrhus Frederic Skinner. *The behavior of organisms: An experimental analysis*. BF Skinner Foundation, 1938.

[5] Jan De Houwer and Sean Hughes. *The psychology of learning: An introduction from a functional-cognitive perspective*. MIT Press, 2020.

[6] Patrick Hammer and Tony Lofthouse. "'OpenNARS for applications': architecture and control". In: *International Conference on Artificial General Intelligence*. Springer. 2020, pp. 193–204.

[7] David Kastak and Ronald J Schusterman. "Transfer of visual identity matching-to-sample in two California sea lions (Zalophus californianus)". In: *Animal Learning & Behavior* 22.4 (1994), pp. 427–435.

[8] Tracy Peña, Raymond C Pitts, and Mark Galizio. "Identity matching-to-sample with olfactory stimuli in rats". In: *Journal of the experimental analysis of behavior* 85.2 (2006), pp. 203–221.

[9] Anthony A Wright. "Concept learning and learning strategies". In: *Psychological Science* 8.2 (1997), pp. 119–123.

[10] BBC Earth. *Sea Lion Plays Snap!* 2013.

[11] Pei Wang. *Non-axiomatic logic: A model of intelligent reasoning*. World Scientific, 2013.