Toward Probabilistic Natural Logic for Syllogistic Reasoning

MSc Thesis *(Afstudeerscriptie)*

written by

Fangzhou Zhai
(born November 20th, 1989 in Changchun, China)

under the supervision of **Dr Jakub Szymanik** and **Dr Ivan Titov**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
August 31st, 2015	Dr Maria Aloni
	Prof Dr Johan van Benthem
	Prof Dr Ing Robert van Rooij
	Dr Jakub Szymanik
	Dr Ivan Titov
	Dr Willem Zuidema

Institute for Logic, Language and Computation

Abstract

Logic emerged as the discipline of reasoning and its syllogistic fragment investigates one of the most fundamental aspect of human reasoning. However, empirical studies have shown that human inference differs from what is characterized by traditional logical validity. In order to better characterize the patterns of human reasoning, psychologists and philosophers have proposed a number of theories of syllogistic reasoning. We contribute to this endeavor by proposing a model based on natural logic with empirically weighted inference rules. Following the mental logic tradition, our basic assumptions are, firstly, natural language sentences are the mental representation of reasoning; secondly, inference rules are among the basic mental operations of reasoning; thirdly, subjects make guesses that depend on a few heuristics. We implemented the model and trained it with the experimental data. The model was able to make around 95% correct predictions and, as far as we can see from the data we have access to, it outperformed all other syllogistic theories. We further discuss the psychological plausibility of the model and the possibilities of extending the model to cover larger fragments of natural language.

Contents

1 Introduction **3**

2 Background **5**
- 2.1 Logic as A Theory of Reasoning: Challenges and Responses . . . 5
  - 2.1.1 The Criticisms . . . 5
  - 2.1.2 The Responses . . . 6
  - 2.1.3 The Probabilistic Basis of Reasoning . . . 6
- 2.2 The Mental Logic Proposal . . . 7
- 2.3 The Natural Logic Program . . . 8
  - 2.3.1 Monotonicity . . . 8
  - 2.3.2 The Natural Logic Program . . . 8
- 2.4 Syllogistic Reasoning . . . 8
- 2.5 Experimental Investigations Of Syllogistic Reasoning . . . 10
- 2.6 The Meta-Analysis of Theories of Syllogisms . . . 12
- 2.7 Theories of the Syllogisms . . . 13
  - 2.7.1 The Atmosphere Hypothesis . . . 13
  - 2.7.2 The Illicit Conversion . . . 13
  - 2.7.3 The Mental Model Theory . . . 14
  - 2.7.4 The Probability Heuristic Model . . . 15
  - 2.7.5 The Natural Logic Approach . . . 17
  - 2.7.6 The Mental Logic System . . . 17

3 Motivations **18**
- 3.1 Mental Representation . . . 18
- 3.2 Probability on the Arena: Reasoning as a Stochastic Process . . . 18
- 3.3 Mistakes or "Realistic Validity" . . . 19
- 3.4 The Difficulty of Reasoning . . . 19

4 A Generative Model For Syllogistic Reasoning **20**
- 4.1 Bayesian Generative Model . . . 20
- 4.2 A Generative Model: Version 1 . . . 20
  - 4.2.1 Theoretical Assumptions . . . 20
  - 4.2.2 Model Definition . . . 21
- 4.3 Data . . . 26
- 4.4 Training . . . 26
- 4.5 Evaluation . . . 27
  - 4.5.1 The Khemlani and Johnson-Laird (2012) Method . . . 27

4.5.2 The Entropy Based Measurements . . . 28
4.5.3 The Approval Rate . . . 29
4.6 Version 1: Training Result and Discussion . . . 30
4.6.1 Training Result . . . 30
4.6.2 Summary . . . 31
4.7 The Generative Model: Version 2 . . . 32
4.7.1 Model Definition . . . 32
4.7.2 Training Results . . . 33
4.7.3 Discussion . . . 34
4.7.4 Summary . . . 34
4.8 The Generative Model: the Complete Version . . . 35
4.8.1 Theoretical Assumptions . . . 35
4.8.2 Model Definition . . . 36
4.8.3 Training Results and Discussions . . . 40
4.8.4 Summary . . . 43

5 Further Discussion **44**
5.1 Psychological Plausibility . . . 44
5.2 A Parallel Comparison to Other Theories of the Syllogisms . . . 45
5.3 Towards a Uniform Theory of Reasoning . . . 46

6 Conclusion **48**
6.1 Summary and Future Work . . . 48
6.2 Coda: Theory of Reasoning, a Mosaic? . . . 48

A The Predictions Of the Model With the Experimental Data **54**
A.1 Predictions: Version 1 . . . 54
A.2 Predictions: Version 2 . . . 57
A.3 Predictions: Version 3 . . . 60
A.4 Predictions of Other Theories of the Syllogisms . . . 63

Chapter 1

Introduction

This thesis is about reasoning. The psychology of reasoning tries to answer one question: how do people reason? The discipline of logic was first to suggest a solution. Aristotle proposed the syllogistic theory as an attempt of characterizing rationality. Nowadays, theory of reasoning is in the center of the investigations of many scientific disciplines, from psychology and economics to cognitive science and artificial intelligence[1]. It is observed through empirical studies that people do not reason according to traditional logical validity and new paradigms (e.g., the probabilistic validity, see Chapter 2) are being proposed. However, the syllogistic reasoning remains a central topic as it revolves around the inference within the most fundamental fragment of natural language that includes the basic quantifiers like "all", "some", and "no" and the monadic predicates. A number of syllogistic theories have been proposed. Among those Rips (1994) proposed the mental logic model assuming that formulas are the mental representations for inference, and that when reasoning, subjects generate a sequence of formulas linked by the adoptions of the specific inference rules. In turn, Geurts (2003) designed a proof system based on monotonicity which could be used to evaluate the difficulty of each syllogisms.

In this thesis we design and train a generative model for syllogistic reasoning based on a probabilistic natural logic[2]. This can be treated as a first step to to integrate the mental logic approach and the natural logic approach. The plausibility of the model lies in natural logic operating on the surface structure of natural language, which is a more reasonable candidate for the mental representation of reasoning. We assume that the procedure of reasoning consists of two types of mental events: the inferences made by the subjects, which are deliberate and precise, and the guesses, which could be less reliable but fast. Accordingly, the model consists of two parts: the inference part, which takes the form of a probabilistic natural logic (i.e., the inference rules are weighted with probabilities) and the guessing part, which leads the subject to a possible conclusion in one step depending on a few heuristics. We implemented the model, and trained it with experimental data. We evaluated the model by multiple means of evaluation. The model exhibits nice performance and outperforms all other theories whose predictions were accessible to us[3]. Besides, the training results yields interesting psychological implications.

[1] cf. Isaac et al. (2014) for a survey of logic and cognitive science. See also Van Benthem (2008); Verbrugge (2009)

[2] See also Dotlačil et al. (2014) where the authors designed probabilistic semantic automata for quantifiers whose parameters are also determined by data.

[3] However, not all the data we were able to obtain are completely reliable. See Section 4.3

The thesis is structured as follows. In Chapter 2 we introduce the background of the thesis, which includes introductions to the syllogistic fragment and the syllogistic theories. In Chapter 3 we motivate the key designs of our model. In Chapter 4 we introduce the training methods and the means of evaluations, define the three versions of our model and present their training results sequentially. We discuss a few interesting aspects of the model in Chapter 5. The final chapter includes a summary and some concluding remarks.

Chapter 2

Background

In this chapter we introduce the background of the thesis. We begin with logic and the challenges it faces as a theory of reasoning. Afterwards, we introduce the mental logic proposal and the natural logic project which are crucial to our research. Finally, we introduce the work on syllogistic reasoning, including the theories proposed for it and the experimental results.

2.1 Logic as A Theory of Reasoning: Challenges and Responses

Characterization of human reasoning remains a tough challenge. Traditional logic appears insufficient to completely describe human inference (see also Stenning and Van Lambalgen (2008)). Therefore, new paradigms have been proposed to tackle that issue.

2.1.1 The Criticisms

The Wason Selection Task

Experiments on Wason selection task (Wason (1968); Wason and Shapiro (1971)) showed the difference between human behavior and logic. In the experiments participants are shown four cards and are told that each card has a number on one side and a letter on the other. Only one side of each card is visible to the subjects. The subjects are then asked which cards they need to turn over to verify the statement that, e.g., "Each card that has a D on one side has a 3 on the other". The visible sides of the cards read D, K, 3 and 7.

According to classical logic, the answer should be D and 7. However, the most frequent responses from the subjects, in order of descending frequency, are: D and 3; D; D, 3 and 7; D and 7. The result is quite robust and reproducible, which clearly indicates that the reasoning of humans do not completely follow the prescription of classical logic.

Many studies of Wason's selection task followed that finding. Among those Griggs and Cox (1982) showed that if the cards have ages and types of drinks on their sides and the task is to verify whether "if a person is drinking alcohol, the person must be at least 19 years old", the performance of subjects will be almost perfect. Moreover, Chater and Oaksford

(1999) convincingly analyzed the experimental results on a probabilistic basis.

Monotonicity

Traditional logic is monotonic, that is, if we make a certain inference and then learn something new then the previous conclusion still stands. However, conclusions of daily reasoning are usually defeasible: it is quite often that new knowledge contradicts a previous conclusion, and the subjects turn to trust the new knowledge more, hence, withdrawing the previous conclusion, see, e.g., Stenning and Van Lambalgen (2008).

Psychological Plausibility

Traditional logic is to a huge extent a purely normative enterprise. Logicians try to characterize valid inferences. On the other hand, the goals of the theory of reasoning are more descriptive. The theory tries to characterize human reasoning. One approach here would be to ask which of the logical inferences are also psychologically plausible. For instance, logic can yield infinitely many vapid conclusions. From p follows 'p and p', so also 'p and p and p', etc. This is clearly not what people do. More generally, a typical logical system is closed on deduction; given some statements already belong to the theory, all its conclusions are also included. Clearly, human rationality is not logically closed. Knowing postulates of natural number theory does not give us insights into all truths about natural numbers. This problem with logically based theories of human reasoning is often referred to as 'logical omniscience'. The goal of the theory of reasoning is to provide us with an efficient characterization that tells us which conclusions people are likely to draw. See Johnson-Laird et al. (2015).

2.1.2 The Responses

Many theories have been proposed to face the challenges. Proponents of Bayesian rationality argue that, as uncertainty plays a role, daily reasoning should have a probabilistic basis, and that probabilistic validity should replace conventional logical validity (see, e.g., Oaksford and Chater (2007)). They assume that degree of belief corresponds to subjective probability. Non-monotonic logics were proposed to achieve defeasible reasoning systems (see, e.g., Antoniou (1997)). The theory of mental models (Johnson-Laird (1986)) proposes that reasoning, as a mental procedure, is the generation and then verification of models. The theory is psychologically plausible and also admits defeasible conclusions.

2.1.3 The Probabilistic Basis of Reasoning

Chater and Oaksford (1999) have argued that theories of reasoning should generalize to everyday reasoning that is defeasible. However, traditional logic is not defeasible. Towards a solution some psychologists have proposed that probability should be the underlying basis of everyday reasoning (see Oaksford and Chater (1996) for an analysis of Wason's selection task based on the same idea). Probabilistic logic has been later proposed as the new paradigm (see, e.g., Oaksford and Chater (2007); Pfeifer (2013)). The proponents have assumed that reasoning is a probabilistic calculus: degrees of belief depends on subjective probability, and that probabilistic validity should replace logical validity, at least in daily

reasoning.

In the context of the syllogistic reasoning the sentences are assigned probabilistic interpretations. For example, *All A are B* is naturally understood as "the probability of B given A is 1". Let P be a probability assignment and consider the predicates as probabilistic events. "All A are B" is interpreted as $P(B|A) = 1$; "Some A are B" is interpreted as $P(B|A) > 0$; "No A are B" is interpreted as $P(B|A) = 0$; "Some A are not B" as $P(B|A) < 1$. An inference is **probabilistically valid** (p-valid) if whenever the premises hold in the probabilistic manner, the conclusion holds as well.

2.2 The Mental Logic Proposal

Rips (1994) has proposed theories of quantified reasoning based on formal inference rules (see also Braine and O'Brien (1998) for similar designs). The theory is based on the hypothesis that formulas are the underlying mental representation of reasoning and that inference rules are the basic reasoning operations of the mind. Rips has argued that, deductive reasoning, as a psychological procedure, is the generation of "a set of sentence linking the premises to the conclusion", and "each link is the embodiment of an inference rule that subjects consider intuitively sound". He has formulated a set of rules that includes both sentential connectives and quantifiers.

The input of the reasoning system (referred to as **PSYCOP**) are the representations of the logical assertions where quantifiers are replaced by names and variables. For example, *"Every paper has an author."* would be represented as

$$\text{IF Paper(x) then Author}(a_x, \text{x}).$$

Where x stands for a universally quantified variable. And a_x is a name whose value is determined by x through a Skölem function, which replaces an existential quantifier.

Rips has made necessary constraints on the application of the inference rules to avoid restrictions to its computational power and approximate 'psychologically complete' theory of reasoning. Rips has assumed that people reason from two directions: from the premises to the conclusion, and from the conclusion back to the premises. He has also made the distinction between forward rules and backward rules: the forward rules are applied on the premises towards the conclusions while the backward rules are applied to the conclusions to find out what has to be proved in advance in order to prove that conclusion.

The PSYCOP model uses relatively abstract rules and formal representations (roughly corresponding to the natural deduction system for first-order logic). It is possible that such design reduces the psychological plausibility of the system: humans are unlikely to use such abstract and sophisticated formal system (see also Johnson-Laird (1997)). Besides, the model explains only the logically valid inferences, and has no mechanism to explain why and how people make systematic mistakes (the model does predicts which steps are more likely to yield mistaken operations though).

2.3 The Natural Logic Program

2.3.1 Monotonicity

Some natural language quantifiers admits the property of **monotonicity** (see also Icard III and Moss (2014)). As an example, consider "Some". The inference

Some pines are green.
∴ Some plants are green.

is valid since all pines are plants. But note that the "pines" in "some pines are plants" can be replaced by any object that contains all pines. For example, since "trees" is a super set of "pines", it follows that Some trees are green.

The quantifier "Some" takes two arguments (in our example, they are "pines" and "green"). We say "Some" is **upward entailing** in its first argument, since the word standing as the first argument can be replaced by anything it entails, preserving the validity of the assertion. On the other hand, we say a quantifier is **downward entailing** in an argument if the word that is considered as this argument can be replaced with anything that entails it, preserving the validity of the assertion. As an example, "All" is downward entailing in its first argument. This is illustrated by the following inference:

All trees are green.
∴ All pines are green.

2.3.2 The Natural Logic Program

Monotonicity is a pervasive feature of natural language. People can reason based on monotonicity, even when the underlying meaning of terms is unclear for them. For example, from "Every Dachong has nine beautiful tails" people would infer "Every Dachong has nine tails", without knowing the meaning of "Dachong" (which simple means tiger in Chinese). It is hence fair to say that monotonicity operates on the surface of natural language. On the other hand, reasoning is often narrated in natural language; people often think with natural language when reasoning, sometimes even "think loudly". It is hence interesting to ask whether there is a natural logic that operates on the surface forms of natural language (see, e.g., van Benthem (1986,9, 2008)). Though it is not hard to see that some fragments of natural language admits "natural logic" (e.g., anaphora reasoning), yet it is an interesting question "how much can these fragments possibly cover?".

2.4 Syllogistic Reasoning

Aristotle defined humans to be the rational animals. Indeed, rationality is crucial to many aspect of human life such as law, social communication, decision making, etc. But, after all, what is rationality per se? One plausible definition of rationality is to reason according to the rules of logic, i.e., rational subjects make deductions only according to the reliable relations between the propositions (or in terms of modern logic, rational subjects only make the "valid inferences"). Aristotle proposed the syllogisms, as the first known attempt to formally characterize human reasoning. Although modern logic has gone far beyond the syllogisms in most aspects, since the syllogisms investigates the most fundamental fragment of human reasoning, it continuously receive attention of researchers (see Khemlani and

Johnson-Laird (2012) for a review of the theories of syllogisms).
The sentences of syllogisms are of four different sentence types (or "moods"), namely:

All A are B : universal affirmative (A)
Some A are B : particular affirmative (I)
No A are B : universal negative (E)
Some A are not B : particular negative(O)

Each syllogism has two sentences as the premises, and one as the conclusion. Traditionally, according to the arrangements of the terms in the premises, syllogisms are classified in to four categories, or "figures":

Figure 1	*Figure 2*	*Figure 3*	*Figure 4*
B C	C B	B C	C B
A B	A B	B A	B A
A C	A C	A C	A C

Syllogisms are customarily identified by their sentence types and figures. For example, "AI3E" refers to the syllogism whose premises are of sentence types A and I, and whose terms are arranged according to figure 3, and whose conclusion is of type E. Therefore, altogether, "AI3E" refers to the following syllogism:

All B are C
Some B are A

No A are C

As there are four different sentence types and four different figures, there are 256 equivalent syllogisms in total. These syllogisms are also referred to as the ones that follows the scholastic order. Some psychologists (e.g., Johnson-Laird (1986)), however, believe that the order that the two premises are presented to the reasoners also makes a difference. Namely, they believe that:

All B are C
Some B are A

No A are C

and

Some B are A
All B are C

No A are C

are different syllogisms. In this thesis we stick to the syllogisms following the scholastic order, since the order that the premises are presented does not make a difference in our model. Traditionally, the truth value of the syllogistic sentences are similar to their semantics in modern predicate logic. The difference lies in that, according to the former, both "All A are B" and "No A are B" implies that A is not empty, while according to the latter this is not necessary. There are 256 equivalent scholastic order syllogisms in total, of which 24 are valid according to the semantics of traditional syllogistic logic, and 15 of these 24 are valid according to the semantics of modern predicate logic.

2.5 Experimental Investigations Of Syllogistic Reasoning

Experimental psychologists have developed a battery of tests to study human reasoning. In one typical experimental design, the subjects are presented with the premises and asked "What follows necessarily from the premises ?" Chater and Oaksford (1999) compared five experimental studies and found that differences[1] in the designs of the experiments appear to have little effect on the results. They computed the weighted average, i.e., percentage that each conclusion is drawn. The data is shown in Table 2.1.

[1]These experiments differ in various ways. For example, to answer the question "What follows necessarily from the premises", some researchers (e.g.,Johnson-Laird (1986)) asked the participants to narrate in natural language the entailments while the rest of them made it explicit that the conclusions are among the four syllogistic conclusions and "nothing follows".

Syllogism	Conclusion					Syllogism	Conclusion				
	A	I	E	O	NVC		A	I	E	O	NVC
AA1	**90**	**5**	0	0	5	AO1	1	6	1	57	35
AA2	58	8	1	1	32	AO2	0	6	3	**67**	24
AA3	57	**29**	0	0	14	AO3	0	10	0	66	24
AA4	75	**16**	1	1	7	AO4	0	5	3	72	20
AI1	0	92	3	3	2	OA1	0	3	3	68	26
AI2	0	57	3	11	29	OA2	0	11	5	56	28
AI3	1	**89**	1	3	7	OA3	0	15	3	**69**	13
AI4	0	71	0	1	28	OA4	1	3	6	27	63
IA1	0	72	0	6	22	II1	0	41	3	4	52
IA2	13	49	3	12	23	II2	1	42	3	3	51
IA3	2	**85**	1	4	8	II3	0	24	3	1	72
IA4	0	**91**	1	1	7	II4	0	42	0	1	57
AE1	0	3	59	6	32	IE1	1	1	22	16	60
AE2	0	0	88	**1**	11	IE2	0	0	39	30	31
AE3	0	1	61	13	25	IE3	0	1	30	33	36
AE4	0	3	**87**	**2**	8	IE4	0	42	0	1	57
EA1	0	1	**87**	**3**	9	EI1	0	5	15	**66**	14
EA2	0	0	**89**	**3**	8	EI2	1	1	21	**52**	25
EA3	0	0	64	**22**	14	EI3	0	6	15	**48**	31
EA4	1	3	61	**8**	28	EI4	0	2	32	**27**	39
OE1	1	0	14	5	80	OO1	1	8	1	12	78
OE2	0	8	11	16	65	OO2	0	16	5	10	69
OE3	0	5	12	18	65	OO3	1	6	0	15	78
OE4	0	19	9	14	58	OO4	1	4	1	25	69
IO1	3	4	1	30	62	OI1	4	6	0	35	55
IO2	1	5	4	37	53	OI2	0	8	3	35	54
IO3	0	9	1	29	61	OI3	1	9	1	31	58
IO4	0	5	1	44	50	OI4	3	8	2	29	58
EE1	0	1	34	1	64	EO1	1	8	8	23	60
EE2	3	3	14	3	77	EO2	0	13	7	11	69
EE3	0	0	18	3	78	EO3	0	0	9	28	63
EE4	0	3	31	1	65	EO4	0	5	8	12	75

Table 2.1: Percentage of times each syllogistic conclusions was endorsed. The data is from a meta-analysis by Chater and Oaksford (1999). "NVC" stands for "No Valid Conclusion", all numbers have been rounded to the closest integer. A bold number indicates that the corresponding conclusion is valid.

One important observation made by Chater and Oaksford (1999) is that validity is a crucial factor in the performance of the participants. Firstly, the average percentage of subjects arriving at a valid conclusion is 51%, while that of arriving at an invalid conclusion is 11%: participants, indeed, made an effort along the path of validity. Secondly, subjects tends to mistakenly arrive at invalid syllogism that is different from a valid one just by its figure. For example, the *AO2O* syllogism is the only valid one among the four *AOO* syllogisms, how-

ever, subjects endorse the other three AOO syllogisms (namely $AO1O, AO3O$ and $AO4O$) with fairly high probability. This might be a sign that people are actually not that bad at syllogistic reasoning (Geurts (2003)): even if an error is made, the most probable wrongly endorsed syllogism is quite similar to a valid one, which differs only in the figure. Thirdly, the mean entropy of the syllogistic premises that yields at least one valid conclusion, according to the table above, is 0.729, however, that of the ones that yield no valid syllogisms is 0.921. The difference indicates that the psychological procedures triggered by the two groups of premises are likely to be different.

2.6 The Meta-Analysis of Theories of Syllogisms

Khemlani and Johnson-Laird (2012) published a meta-analysis of twelve existing theories of syllogisms. They classified the theories into three categories: heuristic theories that capture principles that could underlie intuitive responses; theories of deliberative reasoning based on formal rules of inference akin to those of logic; and theories of deliberative reasoning based on set-theoretic diagrams or models. They collected the experimental data from six empirical studies, and compared the predictions of seven theories out of the twelve[2]. A brief version of their comparison is shown in Table 2.2.

Theory	Correct Predictions (%)	Lower Limit	Upper Limit
Verbal Models Theory	84	80	89
Conversion	83	80	86
Mental Models Theory	78	75	81
Atmosphere	78	75	81
PSYCOP Model	77	73	80
Probability Heuristic Model	73	69	77
Matching	71	66	75

Table 2.2: Meta Analysis of Seven Theories of Syllogisms. The limits admit 95% confidence intevals.

The authors pointed out that all twelve theories of syllogisms lack a system to determine how an inferential task is carried out. They proposed that a unified theory of monadic reasoning should be able to explain the following:

- the interpretation and mental representation of monadic assertions, including syllogistic premises;
- what the brain computes and how it carries out all inferential tasks with such assertions;
- the differences in difficulty from one inference to another, and common errors;
- how contents affect performance;

[2]The empirical data they collected includes the syllogisms arranged according to the scholastic arrangements as well as those that are not. Hence, the theories whose predictions are not available on all syllogisms (possibly only available for those arranged according to the scholastic orders) are excluded from the meta-analysis.

- how the ability to reason with monadic assertions develops;
- differences in performance from one person to another, which are likely to reflect the processing capacity of working memory, experience in deductive reasoning tasks, and motivation (these subject differences, however, call for much more research).

2.7 Theories of the Syllogisms

In this section we introduce a few syllogistic theories.

2.7.1 The Atmosphere Hypothesis

The atmosphere hypothesis proposes that a conclusion should fit the premises' "atmosphere", namely, the sentence types of the premises (Sells (1936); Woodworth and Sells (1935); Begg and Denny (1969)). In particular, whenever at least one premise is negative, the most likely conclusion should be negative; whenever at least one premise contains "some", the most likely conclusion should contain "some" as well; otherwise the conclusion are likely to be affirmative and universal.

The idea of atmosphere is plausible since the types of the premises may encode a considerable amount of information about the type of the conclusions. Indeed, the atmosphere effect seems to capture one of the most significant phenomena of human behavior in the syllogistic reasoning, however, this one observation is clearly not sufficient to tell the complete story. One inevitable drawback is that the hypothesis has no mechanism to possibly explain why people can correctly conclude that nothing follows (but see Revlis (1975) for a model based on the atmosphere hypothesis where the subject can make errors).

Johnson-Laird and Byrne (1989) reported an experiment whose outcome might be contrary to the atmosphere hypothesis. The experiment included reasoning with the quantifier "only" or sentences like "only A are B". The meaning of "only" is, in essence, negative, since "only A are B" means "not A" entails "not B". Experimental results show that when both premises include "only", a mere 16% of the conclusion drawn by the subjects contains it; when one of the premises contains "only", it appears in just 2% of the conclusions.

2.7.2 The Illicit Conversion

It is observed that subjects often make invalid conversions on sentences of types "All" and "some not", i.e., they illicitly conclude "All B are A" from "All A are B" and "Some B are not A" from "Some A are not B" (Revlis (1975); Chapman and Chapman (1959)). Illicit conversions might have a probabilistic basis: quite often these conversions result in true conclusions in everyday life: no man is woman hence no woman is man; sugar is sweet hence sweet things contains sugar; no car can fly hence a flying thing is not a car. Additionally, subjects do make one-step illicit conversions when they are requested to infer from even one single premise sentence (e.g., they directly conclude "All B are A" from "All A are B"), and particularly when abstract predicates like A and B are used (Wilkins (1929); Sells (1936)).

Illicit conversions can explain why would subjects endorse some invalid inferences, for example

$$\frac{\begin{array}{l}\text{All C are B}\\ \text{All A are B}\end{array}}{\text{All A are C}} \quad \text{(Invalid)}$$

It could be that, by the illicit conversion, people infer "All B are A" from "All A are B", hence with "All C are B" they get "All C are A". The illicit conversion can also explain why subjects often endorse invalid syllogisms that are similar to valid ones (in the sense that they only differ in the figure and hence needs only few conversions, see also Section 2.5).

2.7.3 The Mental Model Theory

The Dual Processing Theory

Famously, recently it is often hypothesized that there are two cognitive systems for reasoning. System 1 makes rapid, unconscious, heuristic guesses while system 2 makes slower, conscious considerations based on systematic principles (see, e.g., Evans (2003); Sloman (1996); Kahneman and Frederick (2002)). Mental model theorists (Johnson-Laird (1983)) distinguish the systems from a computational perspective: system 1 has no access to working memory and is hence equivalent to finite state automata (at least restricted reasonably in size); system 2 has access to working memory and can carry out all recursive procedures, at least before the memory runs out.

As for how the two systems cooperate in the context of syllogistic reasoning, it could be that subjects use system 1 to generate plausible conclusions and then use system 2 to deliberate on them. It could also be that system 2 is used in the beginning, however later, subjects may turn to system 1 due to the increasing complexity or the exhaustion of the cognitive resources. It appears that system 1 is faster and costs less resources and may be less precise, hence, accounts more for the errors, yet it is proposed that subjects still makes errors when reasoning consciously and deliberately (Johnson-Laird and Byrne (1991)).

Mental Models

Johnson-Laird (1975) formulated the mental model theory. The theory proposes that subjects represent sets or models **iconically**. They build an iconic mental models for reasoning, each icon representing one object. For example, "all trees are plants" may be represented as

tree	plant
tree	plant
tree	plant

or alternatively, as

tree	plant
tree	plant
tree	plant
	plant
	plant

Each row in these models represents an object of the model. A set of assertions may yield multiple possible models and each model represents a distinct possibility that goes consistent with the assertions. The models explicitly represent only what is true in each possibility (e.g., in the second model, the fifth object is a plant), however, each model can be fleshed out to be a fully explicit representation (also, in the second model, the fifth object is not a tree).

In terms of algorithmic level theory of reasoning, the theory of mental models is a dual processing theory. System 1 has no access to working memory and rapidly generates mental models and conclusions by heuristics (Khemlani and Johnson-Laird (2013)). System 2, however, has access to working memory and can perform any recursive procedure before the cognitive capability becomes overloaded. In particular, system 2 verifies whether a conclusion is consistent with the generated mental models. If a conclusion is falsified by a model then it will be withdrawn, otherwise it will be output as a conclusion.

The theory of mental models has been further developed and implemented as the "mReasoner" (Khemlani and Johnson-Laird (2013)). In a meta analysis (Johnson-Laird et al. (2015)), mReasoner outperformed all other theories available for the analysis.

2.7.4 The Probability Heuristic Model

Marr's Levels

To take a better view of the model, we firstly introduce Marr's levels of cognitive models. When analyzing the visual system, David Marr proposed that, from the computational perspective, tasks performed by the cognitive system must be analyzed at three levels (Marr and Vision (1982)):

- The computational level: what is the function computed by the cognitive system?
- The algorithmic level: what is the algorithm used by the brain to obtain the solution?
- The implementation level: how is the algorithm implemented in the neural system?

These levels are not independent from each other, as it is well possible that considerations at each level constrain the answers at the other levels. In fact, the more fundamental levels can be seen as an implementation of the level above it.

Marr's levels have been a popular and useful guideline for analyzing the cognitive system (however some argued that the classification is not without shortcomings. See McClamrock (1991) for a proposed refinement). Many theories of reasoning are mainly devoted to the algorithmic level theory and has an underlying computational level theory (but see also Section 5.3 for different opinions).

Probability Heuristic Model

Chater and Oaksford (1999) proposed the Probability Heuristic Model (PHM). The model has a probabilistic basis. According to the proponents, an appropriate computational level theory of reasoning should be formalized as probability calculus. As for the algorithmic level theory, Chater and Oaksford (1999) did not suppose that subects are able to compute the p-validities in order to reason, but instead postulated that subjects are equipped with

a few heuristics that often result in p-valid inferences. They propose that, people use these heuristics to generate putative conclusions (system 1); some subject will later test (according to the proponents, the "test" part is not well developed among humans) the validities of these possible conclusions (system 2).

The generation heuristics are:

- *Min-heuristics*: the most preferred conclusion has the same sentence type as the least informative premise.

- *P-entailments*: the second most preferred conclusion is a p-entailment of the most preferred conclusion.

- *Attachment heuristic*: If just one possible subject noun phrase (e.g., Some R) matches the subject noun phrase of just one premise, then the conclusion has that subject noun phrase.

The terminology "**informativeness**" origins from Shannon's information theory: the smaller the probability of a signal, the greater the amount of information it carries. Hence, sentence that carry the least information admits the highest probability. Therefore, subjects are supposed to have a preference for sentences of the less informative sentence types over the rest. Taking the p-entailment relation into consideration (means, the entailed sentence should not contain more information than its premise), a rank order for the sentence types of syllogistic sentences is: $all > some > no > some\_not$.

The two test heuristics are:

- *Max-heuristics*: subjects' confidence in the conclusions generated is in positive proportion to the informativeness of the more informative premise. Lower confidence level means that subject is more likely to conclude that nothing follows.

- *Some_not heuristic*: avoid producing some_not sentences since they are highly uninformative.

Chater and Oaksford (2008) used the following example to illustrate the heuristics.

Premises:	All P are Q	
	Some R are not Q	

	Some Not	(Min-Heuristic)
	Some R are not P	(Attachment Heuristic)
A further conclusion:		
	Some R are P	(p-entailment)

According to Chater and Oaksford (1999) the PHM performs well in predicting human behavior in syllogistic reasoning. The introduction of informativeness, both in evaluating subjective confidence and in generating the conclusions, interestingly brings the probabilistic basis into the modelling. However, there are critics as well: the model sometimes arrives

at conclusions that are not p-valid, which is inconsistent with its own proposal; besides, the five heuristics appears to be quite "magical": it is hard to imagine how the brain could have developed such capabilities.

2.7.5 The Natural Logic Approach

Geurts (2003) designed a proof system for syllogistic reasoning that operates on the syllogistic sentences. He further enriched the proof system with a difficulty weights assigned to each inference rules to evaluate the difficulty of valid syllogisms. The most important component of the proof system is the monotonicity rule. Geurts' proof system for syllogistic reasoning, equivalently, consists of the following rules.

- *All-Some*: "All A are B" implies "Some A are B".
- *No-Some_not*: "No A are B" implies "Some A are not B".
- *Conversion*: "Some A are B" implies "Some B are A"; "No A are B" implies "No B are A".
- *Monotonicity*: If A entails B, then the A in any upward entailing position can be substituted by a B, and the B in any downward entailing position can be substituted by an A.
- *Additional Rule*: "No A are B" and "Some C are A" implies "Some C are not B".

Here the additional rule is a supplement of the monotonicity rule: we see "No A are B" is equivalent as "All A are not B", or "A" entails "not B"; by applying the monotonicity rule to "Some C are A" we have "Some C are not B".

Geurts assumed that different rules cost different amount of cognitive resources. He gives each subject an initial budget of 100 units; each use of the monotonicity rule costs 20 units (the extra rule costs 30 units); a proof containing a "Some Not" proposition costs an additional 10 units. Taking the remaining budget as an evaluation of the difficulty of each syllogism, the evaluation system fits the experimental data from Chater and Oaksford (1999) well. However, the system cannot make any evaluation on most invalid syllogisms, hence cannot explain why subjects can possibly arrive at invalid conclusions (according to Geurts, the system was never intended to give a "full-blown account of syllogistic reasoning" in the first place, see also Khemlani and Johnson-Laird (2012)).

2.7.6 The Mental Logic System

The mental logic theory (see Section 2.2), proposed as a unified theory of reasoning, also yields its syllogistic version. Rips investigated the model through a number of experiments. He has showed that the system is able to fit the data reasonably well. The model, however, is not able to make concrete predictions about the invalid syllogisms that are likely to be endorsed by people: after all, the system only makes logically sound inferences, hence, has no mechanism to explain the mistakes made by people. According to Rips, errors may occur for various reasons, such as failure to recognize the possibility of applying a rule, failure to retrieve a rule, trying to apply a complicated rule, etc. Although the model can predict which steps of the inference are more likely to lead to mistakes, the model is not able to specify which particular mistakes the subject is likely to make.

Chapter 3

Motivations

This thesis is devoted to the psychology of reasoning, or in particular, a probabilistically weighted natural logic for syllogistic reasoning. Mathematically the logic takes the form of a generative model in which the values of the parameters are derived from data using machine learning techniques. In this chapter we motivate the key factors that shape the design of the model.

3.1 Mental Representation

The intimate connection between reasoning and language generation indicates that it is worthwhile to ask whether reasoning has a linguistic mental representation, or to what extent can the mental state of reasoning be represented by natural language. We have mentioned the mental logic proposal by Rips, for which a natural deduction system is designed and implemented. Let us note that apart from logical quantifiers, Rips has also included sentential connectives in his system. However, the system is still formulated as sophisticated formal language, its rules and language being relatively abstract, and thus, it is hard to imagine that these formal systems are implemented in the human brain, especially for those who have no formal trainings (see also Johnson-Laird (1997)).

As an attempt to make the mental representation more intuitive and psychologically plausible we hope to design natural logic based theories for reasoning. That means, the mental representations will be given directly as natural language sentences, without an intermediate layer of an abstract formal language.

3.2 Probability on the Arena: Reasoning as a Stochastic Process

We assume that reasoning, at least as far as we can learn from psychological experiments, admits an underlying stochastic process. The randomness may come from two sources. Firstly, randomness may origin from the subject level, namely, each subject may adopt different possible inferential operations with different probabilities. Secondly, randomness may origin from the population level. It may be that each subject makes relatively constant choices when reasoning, however, different subjects may vary significantly in their 'reasoning style', hence, the reasoning on the population level may be best viewed as a random

process. Our model is designed assuming that the randomness comes from the subject level, however, when the population is measured as a whole, we cannot distinguish which of the two possibilities (or both) yields the source of the randomness.

3.3 Mistakes or "Realistic Validity"

The Wason selection task (see Chapter 2) shows that human reasoning does not follow the path predicted by traditional logic. We may say that people make mistakes. The use of the word "mistake" is justified here by the fact that what people do is not logically valid; or we may say that traditional logic fails to capture the pattern of human reasoning, and there may be a "realistic validity" beyond traditional logical validity, whose characterization would be the ultimate task of the cognitive computational level theories of human reasoning. In the context of this thesis we will prefer the latter solution. We would like to capture the human reasoning, in some sense, regardless of its level of agreement with traditional logical validity. We shall, in our model, enable the reasoners to diverge from the track of logical validity and make their "favorite" mistakes.

3.4 The Difficulty of Reasoning

Evaluating difficulty of reasoning is a very interesting topic of research. Geurts (2003) designed a natural logic for syllogistic reasoning (see Section 2.7). He estimated the cognitive difficulty of each inference rule and used the complexity of the minimal proof, which depends on the difficulty of the inference rules, to evaluate the difficulty of each syllogisms. Similarly, Gierasimczuk et al. (2013) designed a model of the deductive mastermind game based on analytic tableau method. By analyzing the data collected online the authors were able to find out a few factors that are crucial to the difficulty of the problem items.

We agree that each inference step is of different cognitive complexities and that the difficulty of each syllogism depends on the overall complexity of its proof. We will adopt machine learning techniques to evaluate these difficulties, and int turn, weight the logic with them. This should improve the descriptive performance of the reasoning model based on natural logic. Methodologically, machine learning techniques serve as a natural extension of the statistical tools experimental psychologists have been using to analyze the experimental data.

Chapter 4

A Generative Model For Syllogistic Reasoning

In this chapter we propose models for syllogistic reasoning and then train them from the data. The model has three versions of increasing complexity. In the following, we introduce the models, the training method, and the results of experimental evaluation.

4.1 Bayesian Generative Model

A Bayesian generative model is a probabilistic model that randomly generates observable data and possibly depends on a few parameters. It could be used to simulate a procedure modeled as a stochastic process. As an example, a random number generator that generates zeros with probability 0.4 and ones with probability 0.6 could simulate unfair coin flips.

In this thesis we will use a Bayesian generative models to simulate the procedure of reasoning. We assume that the mental events constituting reasoning have a random structure. The conclusions of reasoning are the observable data. The randomness depends on a few parameters, and the training will help to find the values of these parameters under which the model achieves optimal performance with respect to the data.

4.2 A Generative Model: Version 1

Version 1 of the model is based on a sound and complete probabilistic natural logic for the syllogistic fragment. We begin this section by discussing some theoretical assumptions; afterwards we formally define the model.

4.2.1 Theoretical Assumptions

Mental Representation

Similar to what Rips (1994) proposed, we take set of syllogistic sentences as the mental representation of reasoning. Namely, the subject maintains a set of sentences in the working memory to represent the state of reasoning, or more specifically, the subject keeps a record of the sentences that he considers true at the moment. As an example, given the "AE4" premises, the subject generates the mental representation that looks like

All C are B
No B are A

We will refer to each representation as a **state**. Reasoning operations change the mental states. When performing reasoning, the subject generates a sequence of states in the working memory, where the initial state is the set of premises, and the final state contains the conclusion. These states are linked by the **reasoning events**, which can be a specific adoption of an inference rule. For example, given the "AE4" premises, if the subject adopt the "All - Some" rule (i.e., "All A are B" implies "Some A are B") on the premise "All C are B", a "Some C are B" will be obtained, possibly as a conclusion. The subject may also terminate the reasoning and decide that nothing follows, see Figure 4.2.1.

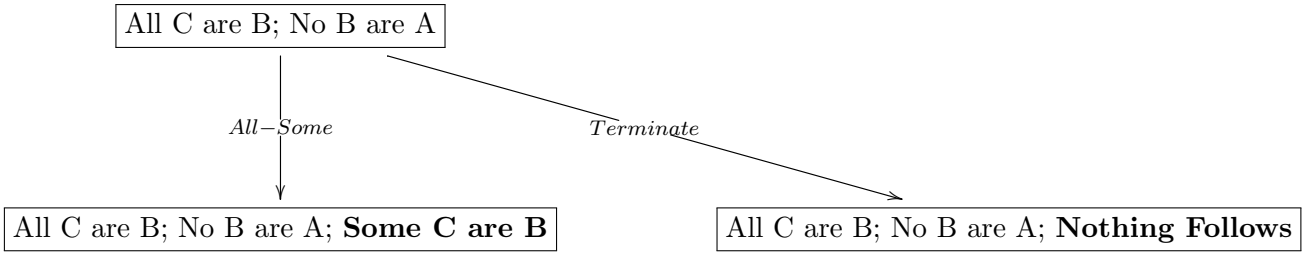

Figure 4.1: The Mental Representations

We would like to point out here that the sentences in each state may not be logically consistent. There are many reasons for this assumption. For example, people tends to adopt illicit conversions (see Section 2) which often leads to the inconsistency. After all, people do often make mistakes resulting in conclusions that are inconsistent with assumptions, even if reasoning in a conscious, deliberate way (see, e.g., Johnson-Laird and Byrne (1991)).

4.2.2 Model Definition

Language and Proof System

We restrict ourselves to the syllogisms following the scholastic order (see Section 2.4 for the details). We base our model on a slightly simplified version of the proof system provided in Geurts (2003), which is sound and complete with regard to the traditional validity of the syllogistic fragment. The proof system has four rules. The most important component is the **monotonicity rule**:

$$\frac{A \Rightarrow B \quad \ldots A^{+} \ldots}{\ldots B^{+} \ldots} \qquad \frac{B \Rightarrow A \quad \ldots A^{-} \ldots}{\ldots B^{-} \ldots} \qquad \textbf{Monotonicity}$$

where a "$+$"($-$) indicates that the sentence is upward (downward) entailing at the corresponding argument (see Section 2.7.5). In words, any predicate A in an upward entailing position can be replaced by its entailments; any predicate A in an downward entailing position can be replaced by the predicate that entails it. Specifically, in the context of syllogistic reasoning, we have

- ⋆ If A entails B (i.e., All A are B), then the A in any upward entailing position can be substituted by a B, and the B in any downward entailing position can be substituted by an A.
- ⋆ "No A are B" and "Some C are A" implies "Some C are not B".[1]

The second rule is the **conversion** rule:

$$\frac{\text{Some A are B}}{\text{Some B are A}} \qquad \frac{\text{No A are B}}{\text{No B are A}} \quad \textbf{Conversion}$$

One source of psychological plausibility of this rule is that people may have preference for symmetric relations (see, e.g., Dickstein (1981)). The remaining two rules are:

$$\frac{\text{All A are B}}{\text{Some A are B}} \quad \textbf{All-Some}$$

$$\frac{\text{No A are B}}{\text{Some A are not B}} \quad \textbf{No-Some not}$$

Note that these rules implicitly indicate that both "All A are B" and "No A are B" implies that A is not empty. We summarize the rules as follows.

- *All-Some*: "All A are B" implies "Some A are B".
- *No-Some_not*: "No A are B" implies "Some A are not B".
- *Conversion*: "Some A are B" implies "Some B are A"; "No A are B" implies "No B are A".
- *Monotonicity*:
  - If A entails B (i.e., All A are B), then the A in any upward entailing position can be substituted by a B, and the B in any downward entailing position can be substituted by an A.
  - "No A are B" and "Some C are A" implies "Some C are not B".

As an illustration, a proof for the EA2E syllogism is as follows.

$$\begin{array}{lll} \text{No C are B} & (1) & \\ \text{All A are B} & (2) & \\ \hline \text{No B are C} & (3) & \text{Conversion}(1) \\ \text{No A are C} & (4) & \text{Monotonicity}(2,3) \end{array}$$

[1] To make it explicit that this is also an adoption of the monotonicity rule, note that "No A are B" means "A entails not B". See also Section 2.7.5

The Tree Representation

Recall that each syllogism consists of two sentences as the premises. If we take the representations of the mental states (i.e., the formulas that have been "proved" to be true) as the nodes and the adoption of inference rules as the edges, the reasoning process can be represented by a tree.

Definition 4.2.1 (The inference tree)**.** *Given P as the set of premises, $R = \{Monotonicity, Conversion, Allsome, Nosomenot\}$ as the set of inference rules, the* ***inference tree*** *is determined as follows.*

- *Each* ***node*** *S is identified by $\langle P_S, ind_S \rangle$, where P_S is the set of sentences that have been proved at node S, ind_S is an index that is unique for each node.*
- *The root R represents the initial state. Its set of proved sentences $P_R = P$.*
- *An* ***inference event*** *is the specific adoption of a particular inference rule; the* ***type*** *of the event is the name of the inference rule. For each node S, the set $E(S)$ collects all events that can happen at node S, i.e., all possible proper adoptions of the inference rules ("proper" meaning the adoption* ***expands*** *the set of proved sentences).*
- *Each element of $E(S)$ yields a child node of S, with the set of "proved" sentences modified accordingly.*

As an example, consider the EI1 premises for which the set of premises is

$$\begin{array}{ll} \text{Some A are B} & (1) \\ \text{No B are C} & (2) \end{array}$$

Now consider the inference options: we can apply the conversion rule on any one of the sentences, and we can apply the monotonicity rule on both sentences. Hence the set of possible events at the root is

$$E(R) = \{Conversion(1), Conversion(2), Monotonicity(1,2)\}$$

Now consider the OE1 premises. The set of premises is

$$\begin{array}{ll} \text{No A are B} & (1) \\ \text{Some B are not C} & (2) \end{array}$$

The set of events at the root is

$$E(R) = \{No - Somenot(1), Conversion(1)\}$$

There are two elements in $E(R)$, hence the root has two children. For the first one, event $No - Somenot(1)$ would take place, expanding the set of premises with a "Some A are not B"; for the other one, event $Conversion(1)$ would occur, adding a "No B are A" to the set of proved sentences. See Figure 4.2.2.

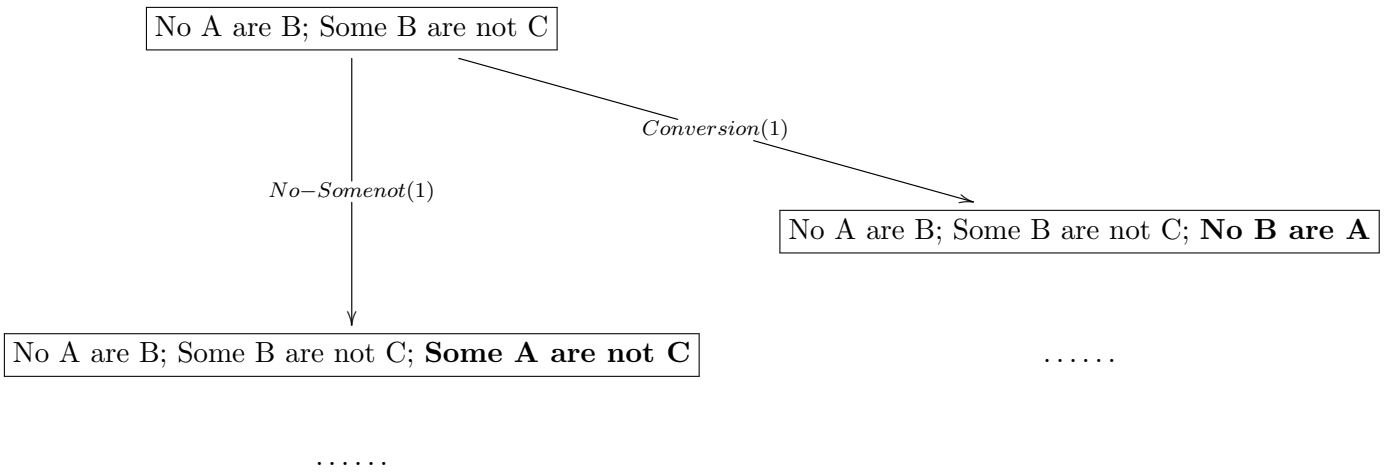

Figure 4.2: The first two levels of the inference tree of the OE1 premise

Note that each path in the tree, from the root to a leaf, yields a proof of the syllogism that consists of the premises and the latest member of the set of proved sentences of the leaf.

We made two technical adjustments to the inference tree, due to both computational and psychological considerations.

We excluded the adoption of the rules that generates no new members for the set of proved sentences. From the computational perspective, we want to make sure that the tree is finite (without the assumption this is not guaranteed: as an example, the conversion rule can be infinitely adopted whenever there is a sentence of type "Some" or "No"); from the psychological perspective, it is simply not plausible to consider the scenario in which a subject repeatedly use one rule.

We limit the height of the tree to four, i.e., each path from the root to a leaf has a maximum length of four, endpoints included (that means, each proof sequence has at most five sentences). From the computational perspective, limiting the height of the tree keeps the size of the tree reasonable, which makes possible the training procedure (this also makes the tree finite, with or without the first assumption). Besides, the restriction is safe, since it has been verified that all valid syllogisms can indeed be proved by our system within three steps, and three is minimal (as for the model, the subject does not necessarily follow the optimal paths). From the psychological perspective, it is plausible to assume that subjects do have restricted cognitive resources that may correspond to something like a limited steps of inference. One could naturally question our arbitrary choice of the cut-off point. However, the number has been actually determined from the data: the model performs best with the step restriction equal three. So it is the data that tells us what is the proper threshold.

Probabilistic Weights

We now bring probability into play, making the model a Bayesian generative model. We base the model on the inference tree we just defined. We assume that different inference

rules are of different cognitive complexities: the possibility to apply some of the rule can be quite straightforward while that of adopting some other may be harder. Or under alternative interpretation, the subject may prefer some rules over others. We represent the difference by the probabilities with which the subject adopts each rule: easier (more preffered) rules get higher probability, while more difficult (less preferred) ones receives lower probability. Formally, we assign each inference rule a **tendency** value, which is intended to be positively related to the probability that the rule is assigned. Consequently, each node of the inference tree receives a probability.

Definition 4.2.2 (The generative story. Version 1)**.** *The set of parameters are the tendencies:*

$$\theta_0 = \{T_{monotonicity}, T_{conversion}, T_{All-Some}, T_{No-Somenot}\}$$

The generating probabilities are determined as follows.

- *The root receives probability 1.*

- *For each node S, the probability N receives is inherited by its children. Specifically, a node S_r resulting from the adoption of inference rule r at S receives conditional probability*

$$p_0(S_r|S, \theta_0) = \frac{T_r}{\sum_{r \in R} c_r \cdot T_r}$$

where R is the set of inference rules; c_r is the number of ways that rule r can be adopted at S; T_r represents the tendency value of rule r.

As an illustration, consider the OE1 premises we have discussed above. Recall that the set of premises is

$$\begin{array}{cc} \text{No A are B} & (1) \\ \text{Some B are not C} & (2) \end{array}$$

The set of events at the root is

$$E(R) = \{No - Somenot(1), Conversion(1)\}$$

Hence by the definition, the probability to adopt the No-Somenot rule at the root, namely the conditional probability that the first node N_l receives is

$$p_0(S_l|R, \theta_0) = \frac{T_{No-Somenot}}{1 \cdot T_{No-Somenot} + 1 \cdot T_{Conversion}}$$

and the second node receives probability

$$p_0(S_r|R, \theta_0) = \frac{T_{Conversion}}{1 \cdot T_{No-Somenot} + 1 \cdot T_{Conversion}}$$

The Predictions of the Model

The model predicts probability with which every syllogistic inference should be endorsed. For any premise set (note that this the figure is also determined by the two particular premise sentences in the premise set), there are five possible conclusions in total:

$$Y = \{A, I, E, O, NVC\}$$

where A, I, E, O, NVC stands for All A are C, Some A are C, No A are C, Some A are not C and 'nothing follows' (or No Valid Conclusion), respectively. When the inference terminates (i.e., when the subject arrive at a leaf node), the subject draw conclusion A, I, E or O if the sentence is included in the set of proved sentences of the leaf node; the subject concludes 'nothing follows' if none of the A, I, E, O conclusion sentence has been proved. We say that a node N is **consistent** with a conclusion y if y can be concluded from N.

Note that each leaf node S uniquely determines a path from the root to itself. We denote the path by $S_0 \dots S_n$ where $S_0 = R$ (the root) and $S_n = S$ (the current node). Hence the node S receives probability

$$p_0(S|R, \theta_0) = \prod_{0 \le i < n} p_0(S_{i+1}|S_i, \theta_0)$$

where each conditional probability is defined as in Definition 4.2.2. The probability that a conclusion y is drawn is hence

$$p_0'(y|R, \theta_0) = \sum_{\text{S is a leaf consistent with y}} p_0(S|R, \theta_0)$$

in order to align the predictions with the experimental data, we normalize the probabilities, i.e., the predicted probability of conclusion y given premises X is

$$p_0(y|X, \theta_0) = \frac{p_0'(y|R, \theta_0)}{\sum_{y' \in Y} p_0'(y'|R, \theta_0)}$$

4.3 Data

We use the data from the meta-analysis by Chater and Oaksford (1999), as is shown in Table 2.1. Occasionally we denote the data set as $\{X_i, y_i\}_{i \le n}$, where X_i stands for the pair of premises and y_i stands for the conclusion. We randomly select 50% of the premises (i.e., half the member of X) and use the corresponding data as the training data, according to which we train the model; while the rest of the data are used as the testing data (which was kept invisible during the training).

4.4 Training

We train the model according to the maximum likelihood target function. Namely, we need to compute

$$\arg\max_{\theta_0} \; p_0(\{(X_i, y_i)\}_{i \le n}|\theta_0)$$

Intuitively, we want to find the values of the parameters that maximizes the probability of the observed data, which is one of the plausible definitions of "optimal performance". We use the Expectation-Maximization algorithm to train the model. The initial values of the parameters are sampled from the uniform distribution over $[0, 1]$. The sketch of the training goes as follows[2]

- $\boldsymbol{E - Step_t}$: For each X_i, y_i, we compute

$$p_0(y_i|X_i, \theta_0^t)$$

- $\boldsymbol{M - Step_t}$: The purpose is to estimate θ_o^{t+1} by

$$\theta_0^{t+1} = \arg\max_{\theta_0} L(\theta_0; \theta_0^t) = \arg\max_{\theta_0} \sum_{i \leq n} \sum_{S \in S(X_i, y_i)} \frac{p_0(S|X_i, \theta_0^t)}{p_0(y_i|X_i, \theta_0^t)} \cdot p_0(S|X_i, \theta_0)$$

where $S(X_i, y_i)$ collects all nodes in the tree of X_i that has a child who is consistent with y_i. In each iteration, to follow the gradient direction, we compute

$$\theta_0^{t+1} = \theta_0^t + \eta \cdot \frac{\partial L(\theta_0; \theta_0^t)}{\partial \theta_0}$$

where η is the step size.

4.5 Evaluation

We use a mixed means of evaluation. We mainly use the evaluation method proposed by Khemlani and Johnson-Laird (2012) that is based on the signal detection theory, and is applicable to all theories of the syllogisms. Where applicable, we also adopt a few more means of evaluation as an attempt to extract more information from the data.

4.5.1 The Khemlani and Johnson-Laird (2012) Method

Khemlani and Johnson-Laird (2012) assume that the conclusions of the participants are noisy, in the sense that unsystematic errors occur frequently. Hence, according to Khemlani and Johnson-Laird (2012), the experimental data are classified into two categories: those conclusions that appear reliably more often than chance level, which a theory of the syllogisms should predict to occur; and those that do not occur reliably more than chance level, which a theory should predict will not occur.

In our context, there are five possible conclusions that can be drawn by subject. The chance level is thus 20%. In the following we count a conclusion as reliable if it is drawn significantly often, i.e., in at least 30%[3] of the trials.[4]

[2] The details of the training algorithm is not important for our thesis. In case of need, one could cf. e.g., Dempster et al. (1977); Baum et al. (1970).

[3] To align with Khemlani and Johnson-Laird (2012), we set the significant level as 1.5 times the probability of the chance level as they did.

[4] This is slightly different from what used by Khemlani and Johnson-Laird (2012) since they also included the non-scholastic order syllogisms, hence there are nine possible conclusions in their experiments, while we have five.

The method is generally applicable to all theories of the syllogisms. As far as a theory predicts what will be concluded from each pair of premises, the method can be applied to evaluate the theory. According to the type of fitting, the predictions of a model are classified into four categories:

Predictions \ Exp. Data	$< 30\%$	$\geq 30\%$
$< 30\%$	Correct Rejection	Miss
$\geq 30\%$	False Alarm	Hit

Table 4.1: Break-down of Predictions

4.5.2 The Entropy Based Measurements

The **entropy** of a discrete random variable X with possible values $\mathcal{X}$ is defined as

$$H(p) = \sum_{x \in \mathcal{X}} -p(x) \cdot log(p(x))$$

where $p(x)$ is the probability of $\{X = x\}$. For each item $x \in \mathcal{X}$, $-log(p(x))$ is also called the **surprisal** of x. The idea is that, the less the probability, the higher the amount of information that is contained in the occurrence of the item. Entropy measures the average amount of information contained in each item of the distribution, weighted by the probabilities. It is a measurement of the amount of information (or uncertainty) contained in a probabilistic distribution. To be more elaborative, the entropy of constants is zero; the entropy of a binary distribution where the items takes probabilities $0.01, 0.99$ is close to zero (since the amount of uncertainty is low: for 99% probability the second item shall occur); the entropy of a binary distribution where the items take probabilities $0.5, 0.5$ is 1, which is highest among binary distributions: the distribution is completely random. Entropy is a crucial parameter of a distribution. In the context of cognitive science, it is also a lower bound of the amount of information that is processed during a cognitive task. Besides, differences in the entropy are indicators that the corresponding cognitive processes are different.

The formulation of the predictions of our model and that of the experimental data coincides with each other, in the sense that they both take the form of probabilistic distributions: if we fix the premises, the experimental data we use provides the probability (or frequency) that participants draw each conclusion; on the other hand, our model is a Bayesian generative model, which, given the set of premises, predicts the probability that the subject draws each conclusion (this is not the case for many other syllogistic theories that only predict which conclusions should be drawn). Therefore, we are enabled to use a few entropy based measurements as applied to probabilistic distributions: the mean entropy, the mean entropy error and the mean KL divergence.

The mean entropy computes the average entropy of the probabilistic distributions of the conclusion following from each pair of premises. Namely, we compute

$$MEnt = \sum_{i \leq n} (\frac{1}{n} \cdot \sum_{y \in Y} -p(y|X_i) log(p(y|X_i)))$$

where X_is are the pairs of premises and Y collects all possible conclusions. For each pair of premises, it measures the average amount of information of the distributions of the possible

conclusions. It is also a lower bound of the amount of information that is processed by the cognitive system, or in some senses, the workload. For example, if the experimental data admits a $99\% - 1\%$ distribution, of which the entropy is low, then the cognitive system should have done that with ease; if the distribution is instead $50\% - 50\%$, then it is well possible that the participants struggled between the two options, and that the cognitive system made a lot of difficult decisions to arrive at the conclusion. Khemlani and Johnson-Laird (2012) also observed that the entropy of the distribution appears positively related to the difficulties of the syllogisms: the higher the entropy, the more difficult the choice.

The mean entropy error is the mean absolute value of the differences between the entropies of the predicted distributions and the distributions provided by experimental data, and is computed as

$$MEE = \sum_{i \leq n} (\frac{1}{n} |H(p(\cdot|X_i)) - H(p'(\cdot|X_i))|)$$

where $H(\cdot)$ computes the entropy of a random variable. The mean entropy error is used to aid the mean entropy measurement in evaluating the difference in the entropy of the distributions.

The mean KL divergence from the data distribution to the predictions of a model is computed as

$$MKLD = \sum_{i \leq n} (\frac{1}{n} \cdot D(p(\cdot|X_i)||p'(\cdot||X_i))) = \sum_{i \leq n} (\frac{1}{n} \cdot \sum_{y \in Y} p(y|X_i) log(p(y|X_i)/p'(y|X_i)))$$

where $p'(y|X_i)$ is the predicted probability of conclusion y given pair of premises X_i. The KL divergence measures the information loss caused by replacing the data distributions with the predictions of the model, or in some senses, the "distance" from the data distribution to the predictions (the KL divergence is not symmetric, though). The KL divergence $D(p||q)$ is only defined if for all x, $q(x) = 0$ implies $p(x) = 0$. In our context, that means whenever a conclusion is drawn in at least one experimental trial, the model should assign positive probability to the corresponding prediction. This requirement is only fulfilled by the complete version of our model, which is also the only case when we compute the mean KL divergence. As a summary, the $MKLD$ is positively related to the difference between the two distributions, every item counted.

4.5.3 The Approval Rate

We compute the proportion of premises for which the conclusion that receives the highest probability agrees with the experimental data (i.e., is also the one that is most likely to be endorsed by the experimental participants. We refer to this item as the *chief conclusion* in the following, and the proportion is referred to as the *approval rate*). Compared with the KL divergence which measures the distance between the distributions, the approval rate evaluates the predictions of the model on the item that is endorsed by most experimental participants.

The approval rate is interesting from the following perspectives: the chief conclusion holds the highest probability (and quite often, a dominating majority), and carries a lot of information of the distribution; it is also the one that is most likely to be endorsed, and a model

should be able to predict the most likely response well.

The approval rate is applicable only when the predictions of the model takes the form of probabilistic distributions. All versions of our model fulfils this condition.

4.6 Version 1: Training Result and Discussion

4.6.1 Training Result

The training results are shown in the following tables.

Data Set	Correct Prediction		Size	Mean Entropy	
	Count	Percentage		Predictions	Data
Test Set	133	83.1%	160	0.140	0.875
Training Set	128	80.0%	160	0.210	0.852
Complete Set	261	81.6%	320	0.175	0.864
NVC Premises*	187	83.1%	225	0	0.921
Valid Syl. Premises*	74	77.9%	95	0.589	0.729
Valid Syllogisms	23	95.8%	24	N/A	N/A

Table 4.2: Version 1: Training results evaluated according to the Khemlani and Johnson-Laird (2012) method.
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows; the valid syllogisms are the subject syllogisms that is valid.

We see that the model is good at prediction the behavior of participants on valid syllogisms, making 95.8% correct predictions; whereas the number is 83.1% on the test set.

Data set	Approval Rate	Mean Entropy Error
Testing Set	0.66	0.783
Training Set	0.56	0.784
Complete Set	0.61	0.783
NVC Premises*	0.42	0.921
Valid Syl. Premises*	0.69	0.457

Table 4.3: Evaluation: Version 1
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

The approval rate is around 0.6. The mean entropy error is very high, which indicates that what simulated by the model is probability far from the realistic psychological process. The following table shows a break down of the predictions.

Data set	Hit	Correct Rejection	Miss	False Alarm	Size
Testing Set	26	107	15	12	160
Training Set	25	103	18	14	160
Complete Set	51	210	33	26	320
NVC Premises	35	152	28	10	225
Valid Syl. Premises	16	58	5	16	95
Valid Syllogisms	14	9	0	1	24
Experimental Data	85	235	0	0	320

Table 4.4: Version 1: Break Down of Predictions
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

The table shows the break down of predictions according to the Khemlani and Johnson-Laird (2012) method. We see that the performance of the model admits no noticeable difference between the test set and the training set. Also we see that the experimental data admits some bias between the positive samples and the negative samples: 85 out of 320 syllogisms are endorsed by the participants, while 235 are rejected.

The logarithms of values of the parameters are shown in the following table.

Tendencies	$T_{monotonicity}$	$T_{conversion}$	T_{all_some}	$T_{no\_somenot}$
Logarithm Values	0.78	2.92	2.06	0.96

Table 4.5: Version 1: Logarithms of the Values of the Parameters.

The model appears good at predicting the behavior of subjects on valid syllogisms, making 95.8% correct predictions. The only valid syllogism that the model is not able to predict correctly is the EI4O syllogism: experiments admits a percentage of 27% that subjects endorse it, while the model predicted a 32%, causing a close false alarm.

The shortcomings of the model is also obvious. In most cases, the model has no mechanism to explain the errors subjects make. When no valid conclusion follows from the premises, the model can only conclude that nothing follows (this is also why the mean entropy of the predictions are so low), however, the experimental data clearly indicates that subjects do make systematic mistakes. The later versions will be devoted to the modelling of these mistakes.

4.6.2 Summary

We now summarize the basic version of the model.

Based on a probabilistic natural logic of syllogistic reasoning, we implemented a Bayesian generative model by assuming that different inference rules are adopted with different probabilities by subjects. The theoretical assumptions are as follows.

- Sentences are the mental representation of the state of reasoning.

- Inference rules are among the basic mental operations of reasoning.
- Different inference rules are of different cognitive difficulties. The differences are reflected by the differences in the probabilities that they are adopted.
- Adoptions of inference rules are probabilistically independent of the previous rules.
- Subjects have limited cognitive resources for reasoning. This difference is reflected by a maximum step they can make in a proof.
- Subjects make no redundant application of inference rules, at least in our context when the lengths of the proofs are restricted.

The performance of the model is summarized as follows.

- The model appears good at predicting the behavior of subjects on valid syllogisms.
- One vital shortcoming is that the model has little mechanism to predict the patterns of mistakes. In fact, the model can only conclude nothing follows when no valid conclusion follows from the set of premises, which is clearly different from the behavior of the subjects.

4.7 The Generative Model: Version 2

4.7.1 Model Definition

Previous discussion shows that version 1 has the principle shortcoming when it comes to predicting the errors. The second version attempts to partially solve the problem by including the **illicit conversions**.

One systematic mistake people make is adopting the **Illicit Conversions**, which is observed in a number of experiments and is psychologically plausible (see Section 2.7.2). We expect the inclusion of these operations to simulate some systematic mistakes. We decide not to distinguish between the illicit conversions and their licit counterparts: it is simply psychologically implausible that people distinguish between them while adopting the illicit conversions without doubt as if they were licit. Hence what we do is to extend the existing conversion rule to include the illicit conversions. After integrating the illicit conversions into the model, the rules of our underlying proof system are now

- *All-Some*: "All A are B" implies "Some A are B".
- *No-Some_not*: "No A are B" implies "Some A are not B".
- *Conversion*: For all sentence types Q, Q(B,A) follows from Q(A,B).
- *Monotonicity*:
  - If A entails B (i.e., All A are B), then the A in any upward entailing position can be substituted by a B, and the B in any downward entailing position can be substituted by an A.
  - "No A are B" and "Some C are A" implies "Some C are not B".

The definition of the inference tree and the training methods remains unchanged.

4.7.2 Training Results

The following tables show the training result.

Data Set	Correct Prediction		Size	Mean Entropy	
	Count	Percentage		Predictions	Data
Test Set	149	93.1%	160	0.228	0.875
Training Set	145	90.6%	160	0.354	0.852
Complete Set	294	91.9%	320	0.291	0.864
NVC Premises*	205	91.1%	225	0.534	0.921
Valid Syl. Premises*	89	93.7%	95	0.188	0.729
Valid Syllogisms	23	95.8%	24	N/ A	N/ A

Table 4.6: Training results evaluated according to the Khemlani and Johnson-Laird (2012) method.
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

Data set	Approval Rate	Mean Entropy Error
Testing Set	0.97	0.665
Training Set	0.88	0.538
Complete Set	0.92	0.601
NVC Premises*	0.95	0.733
Valid Syl. Premises*	0.91	0.290

Table 4.7: Evaluation: Version 2
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

We see that the performance of the model has been significantly improved compared with the first version. The model now makes 93.1% correct predictions on the test set (and 91.9% on the complete set), which is around 10% higher than the previous version, or makes less than half the mistakes. Also, the approval rate is significantly better, reaching a 0.97 on the test set. However, the mean entropy error remains quite large, which again indicates that the model is not telling the whole story. Also note that the mean entropy error of the premises that yields at least one valid conclusion (the item "Valid Syl. Premises", the mean entropy error is 0.290) is dramatically lower than that of the rest (e.g., the test set, where the error is a 0.665). The difference indicates that what our model fails to capture about human reasoning is most likely related to what people do when there is no valid conclusion: do they guess, and how do they make mistakes.
The following table shows a break-down of the predictions.

Data set	Hit	Correct Rejection	Miss	False Alarm	Size
Testing Set	33	116	8	3	160
Training Set	34	111	9	6	160
Complete Set	67	227	17	9	320
Valid Syllogisms	14	9	0	1	24
NVC Premises*	47	158	16	4	225
Valid Syl. Premises*	20	69	1	5	95
Experimental Data	85	235	0	0	320

Table 4.8: Break-down of Predictions
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

The values of the parameters are shown in the following table.

Tendencies	$T_{monotonicity}$	$T_{conversion}$	T_{all_some}	$T_{no\_somenot}$
Logarithms of Values	1.84	4.50	-0.63	-0.24

Table 4.9: Values of the Parameters.

4.7.3 Discussion

The performance of the model saw a significant improvement. The inclusion of illicit conversions improved the model's capability for predicting the errors in: the vanilla version suffered 33 misses, while the number is halved to 17 for the extended version.

The model reached an overall percentage of correctness of 91.9% which appears promising. However, the training results clearly indicate that the model is not telling the complete story. Even though the inclusion of illicit conversions allowed the model to predict some mistakes, there are still a lot that the model cannot predict. For example. in the case of the II, IO, EE, OI, OE, OO syllogisms, the model can conclude nothing but 'nothing follows'. According to the Khemlani and Johnson-Laird (2012) method, the failure to predict the behavior of the participants on these syllogisms is systematic. Therefore, the model needs further improvements to predict these mistakes. We can also see the shape difference between the mean entropies: the mean entropy of the experimental data is 0.864, while that of the predictions is 0.291. This is a clear signal that what the model simulates is a different procedure from the psychological process.

4.7.4 Summary

We extended the conversion rule in version 1 of the model to include the illicit conversions. The modification is made based on the following assumptions.

- Subjects can deliberate and still err when reasoning.
- Subjects frequently adopt the illicit conversions, without realizing that the conversions are different from their licit counterparts.

The training results of the extended version shows that the inclusion of illicit conversions significantly improved the performance of the model. However, the predictions still suffer a lot of systematic problem; meanwhile, the mean entropy of the predictions differs a lot from that of the data, which indicates that what the model does is still quite different from the behavior of the subjects. The mean entropy error is significantly lower for the premises that yield at least one valid conclusion, which indicates that what the model fails to capture about human reasoning may have mostly to do with handling the situation when no valid conclusions follow.

4.8 The Generative Model: the Complete Version

4.8.1 Theoretical Assumptions

As an attempt to improve the performance of version 2 of the model, and in particular, to allow the model to further predict the errors subjects make, we extend the model further to include the **guessing events**, which, in parallel with the **inference events** (i.e., the adoption of inference rules), are yet another kind of mental events that lead to the expansion of the set of proved sentences. The underlying assumption is that in the reasoning procedure, apart from the formal reasoning, subjects have a probability to "fall off" and make a heuristic guess. The guessing procedure is a less deliberate process, intended to explain what happens when subjects partially give up on the inference and try to obtain a most likely conclusion within one inference step. The reason to turn to the guessing scenario may have to do with an increasing complexity of reasoning or the subject doubting the conclusion that was already obtained. Conceptually, the complete version of the model has two parts: the inference part, as is already defined, and the guessing part, which is to be introduced.

Our model of the guessing scenario consists of two components. One of them determines the probability that a subject guesses "nothing follows", the other determines the probability that she guesses the most likely conclusion from the remaining set of conclusions.

For the first part, we involve the notion of **informativeness**. Recall that informativeness is the measure of the amount of information that each type of sentence carries (see Section 2.7.4). Chater and Oaksford (1999) proposed that the level of **confidence** subjects have in conclusions is related to the level of informativeness of the premises. In particular, the more informative the premise sentences, the more confident are the subjects in the existence of their conclusions, namely, the less likely they will conclude "nothing follows". We adopt this assumption by making the probability that a subject guesses "nothing follows" negatively related to the informativeness of the premise sentences. We further assume that, when the subject concludes 'nothing follows' from the inference, there is a probability, which is positively related to the informativeness of the premises, that she doubts the conclusion and makes a guess to try to get another conclusion.

For the second part, we integrate the **atmosphere hypothesis** into our model. Recall that the atmosphere hypothesis claims that when there is a negation in the premises, subjects are likely to draw a negative conclusion; when there is a "some" in the premises, subjects tend to include a "some" in the conclusion; when neither is the case, the conclusion is often affirmative. The premises are creating a kind of "atmosphere" that affects the conclusion.

In the following we will refer to the type of sentence that is predicted by the atmosphere hypothesis as the **dominant type**. To be precise, the dominant type in the premises is the one that is greater by the order

$$A < I < E < O$$

[5] We assume that when making a guess, the subject is likely to arrive at the conclusion with the dominant type.

One interesting question concerns he potential relationship between atmosphere and informativeness. The proponents of informativeness assumed the following order:

$$A > I > E > O$$

which is exactly the reverse of our order of dominant type. It could be that people prefer the conclusions indicated by atmosphere because they are less informative, hence, have a higher probability to hold. However, we would like to clarify that in our model, atmosphere and informativeness are two independent components. We implemented the atmosphere hypothesis and the order of dominant types, yet made no assumptions of the amount of informativeness of each sentence type, but rather leave the model to learn it from the data.

To summarize, based on the atmosphere hypothesis and the research on informativeness, we added guessing events to enrich the mechanism of our model. When the subject makes an inference, or when she concluded 'nothing follows' but doubt the conclusion since the premises might be highly informative, a **guessing event** occurs with a certain probability.

4.8.2 Model Definition

The underlying proof system remains unchanged. Firstly we modify the inference tree to include the leaf nodes generated by guessing events.

The Tree Representation

In addition to the **inference events**, we now include in our model the **guessing events** which represents the scenario when the subject makes guesses. After each guessing event, the subjects arrive at a conclusion and the proof terminates. Therefore, guessing events, if considered as edges, leads to the leaf nodes in the tree.

Definition 4.8.1 (The inference tree, complete version)**.** *Given P as the set of premises, $R = \{Monotonicity, Conversion, Allsome, Nosomenot\}$ as the set of inference rules, the **inference tree** is determined as follows.*

- *Each **node** S takes the form $\langle P_S, ind_S \rangle$, where P_S is the set of sentences that have been proved at node S, ind_S is an index that is identical for each node.*
- *The root R represents the initial state. Its set of proved sentences is $P_R = P$.*
- *An **inference event** is a specific adoption of a particular inference rule; the **type** of the event is the name of the inference rule.*

[5] Not all formulations of atmosphere include the order $I < E$. We included this pair since from the data E is indeed dominant over I.

- *A **guessing event** represents the incident when a subject makes a guess and gets to a conclusion, instantly terminating the reasoning. A guessing event is likely to happen if there is at least one inference event that is possible (in other words, whenever the subject makes inference, there is a probability to drop off and make a guess, probably leading to a mistake), or the subject has concluded 'nothing follows' through inference events.*

- *For each node S, the set $E(S)$ collects all the events that can possibly happen at node S, including the inference events and the guessing events. There are five possible guessing events when a guess takes place, corresponding to five possible conclusions: guessing A, I, E, O or nothing follows.*
- *Each element of $E(S)$ yields a child node of S, with the set of "proved" sentences modified accordingly.*
- *As before, inference events that generates no new formulas for the set of proved sentences are not considered; maximum number of occurrences of inference events is restricted to three in any path of the tree (see section 4.2.2).*

As an example, consider the EI1 premise, i.e., the set of premises is

Some A are B (1)
No B are C (2)

Now, consider the inference options: we can apply the conversion rule to any one of the sentences, and we can apply the monotonicity rule to both sentences. Hence, the set of inference events at the root is

$$E_1(R) = \{Conversion(1), Conversion(2), Monotonicity(1,2)\}$$

The set of the guessing events at the root is

$$E_2(R) = \{A, I, E, O, NVC\}$$

Hence, the set of the events at the root is

$$E = E_1 \cup E_2 = \{Conversion(1), Conversion(2), Monotonicity(1,2), A, I, E, O, NVC\}$$

Now, consider a node S_e with the following set of proved sentences

Some A are B (1)
Some B are A (2)
Some B are C (3)
Some C are B (4)

We see that no sentences of the form "Q(A,C)" has been proved. Further, no inference rules set can be adopted. In this situation, the subject has to conclude 'nothing follows'. However, according to our new model, the reasoner may consider the premises (in our example, the premises can be, for example, II1) so informative that he doubts the conclusion he just derived. "Really? Nothing follows?", he says, "the premises are so informative and I should have inferred a lot!" He may then, with a probability, decides that the inference

he just made is not reliable and he should guess what is the correct conclusion based on his life experiences. There is a probability that he concludes 'nothing follows' immediately, or his doubt may trigger a guessing event. The set of events at the node is hence

$$E(S_e) = \{A, I, E, O, NVC, NVC'\}$$

Note that NVC stands for the event that the subject turned to a guessing event but still guessed 'nothing follows', while NVC' stands for the event that the subject continue to trust the outcome of the inference and conclude that 'nothing follows'.

The model inherits the constraints on the tree we made for version 1, namely, inference events that does not generate a new sentence are not applied; in each path, a maximum of three adoptions of inference rules can be made. These does not apply to the guessing events, though.

Probabilistic Weights

We now define the probability that each node in the inference tree will receive.

Definition 4.8.2 (The generative story. Complete Version)**.** *There are three types of parameters. The tendency parameters*

$$T = \{T_{monotonicity}, T_{conversion}, T_{All-Some}, T_{No-Somenot}\}$$

represents the tendencies with which the subject adopts each inference rule. The ***atmosphere strength***

$$A = \{AS\}$$

which represents the ***strength*** *of the atmosphere effect, which is positively related to the probability that subjects guess the conclusion predicted by the atmosphere hypothesis. The* ***informativeness parameters***

$$IF = \{IF_A, IF_I, IF_E, IF_O\}$$

represent the informativenesses of each type of sentence. We denote the full set of parameters as

$$\theta = T \cup A \cup IF$$

The probabilities are determined as follows.

- *The root receives probability 1.*
- *For each node S, the probability N receives spreads over its children. Specifically, if the node S is not consistent with 'nothing follows', i.e., if a sentence of the form $Q(A, C)$ (which is a potential conclusion) is already proved at the node or some inference rules can still be applied to the set of proved sentences properly, the subject could either make an inference step or make a guess.*
  - *A node S_r resulting from the adoption of inference rule r at S receives conditional probability*

    $$p(S_r|S, \theta) = \frac{T_r}{\sum_{r \in R} c_r \cdot T_r + 1}$$

    where R is the set of inference rules; c_r is the number of ways that rule r can be adopted at S.

– *Guessing events take place with the rest of the probability, namely*

$$p(Guess|S,\theta) = \frac{1}{\sum_{r\in R} c_r \cdot T_r + 1}$$

There are five candidates for the guess, namely, the conclusion predicted by the atmosphere hypothesis, 'nothing follows' and one of the remaining three.

– *We denote the* ***confidence level*** *of the premise set by*

$$CL = IF_{t_1} + IF_{t_2}$$

where t_1, t_2 represents the sentence types of the premises. The probability of guessing 'nothing follows' is positively related to its inverse, the ***doubt level***

$$DL = \frac{1}{CL}$$

the probability of guessing 'nothing follows' is

$$p(NVC|Guess, S, \theta) = \frac{DL}{3 + AS + DL}$$

– *the probability of guessing the dominant type is*

$$p(t_d|Guess, S, \theta) = \frac{A}{3 + AS + DL}$$

where t_d stands for the dominant type.

– *the probability of guessing any one of the remaining three options is*

$$p(Non\text{-}Dominant\text{-}Type|Guess, S, \theta) = \frac{1}{3 + AS + DL}$$

- *if at node S some conclusion has already been proved and no further inference rule could be applied, the inference terminates and the subject concludes the conclusion;*

- *if the node S is consistent with 'nothing follows' (i.e. no potential conclusion has been proved) and no further inference rule could be applied, there are two possibilities: firstly, the subject feels sufficiently confident in the inference and concludes 'nothing follows'; secondly, the subject doubts the conclusion since the premises are so uninformative, and decides to give up the inference conclusion and make a guess.*

  – *The probability that the subject continues to trust the formal inferences and concludes 'nothing follows' is*

$$p(NVC'|S, \theta) = \frac{1}{1 + CL}$$

Namely, the more confident the subject is in the premises, the less likely that the subject trusts 'nothing follows' conclusion.

  – *The probability the subject doubts the conclusion so much (since the confidence level is high) that he decides it to be unreliable, and takes a guess is*

$$p(Guess|S, \theta) = \frac{CL}{1 + CL}$$

– *If the subject chooses to make a guess, the conditional probability that each option receives is the same as what we already defined, i.e.,*

  * *the probability of guessing 'nothing follows' is*

$$p(NVC|Guess, S, \theta) = \frac{DL}{3 + AS + DL}$$

  * *the probability of guessing the dominant type is*

$$p(t_d|Guess, S, \theta) = \frac{AS}{3 + AS + DL}$$

    where t_d stands for the dominant type.

  * *the probability of guessing any rest of the options is*

$$p(Non\text{-}Dominant\text{-}Type|Guess, S, \theta) = \frac{1}{3 + AS + DL}$$

4.8.3 Training Results and Discussions

The training methods remains the same. The following tables show the training result.

Data Set	Correct Prediction		Size	Mean Entropy	
	Count	Percentage		Predictions	Data
Test Set	153	95.6%	160	0.901	0.875
Training Set	151	94.4%	160	0.830	0.852
Complete Set	304	95.0%	320	0.870	0.864
NVC Premises*	212	94.2%	225	0.939	0.921
Valid Syl. Premises*	92	96.8%	95	0.706	0.729
Valid Syllogisms	23	95.8%	24	N/ A	N/ A

Table 4.10: Training results evaluated according to the Khemlani and Johnson-Laird (2012) method.
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

Data set	Approval Rate	Mean KL Divergence	Mean Entropy Error
Testing Set	0.94	0.161	0.270
Training Set	0.78	0.246	0.276
Complete Set	0.86	0.203	0.273
NVC Premises*	0.82	0.212	0.297
Valid Syl. Premises*	0.95	0.184	0.214

Table 4.11: Evaluation: The Complete Version
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

We see that the proportions of correct predictions have been further improved. Besides, the predictions and the data admits similar mean entropy. The mean entropy error also decreased to a reasonable level (which is around one third of that of the previous version). The following table shows a break-down of the predictions.

Data set	Hit	Correct Rejection	Miss	False Alarm	Size
Testing Set	37	116	4	3	160
Training Set	37	114	6	3	160
Complete Set	74	230	10	6	320
Valid Syllogisms	14	9	0	1	24
NVC Premises*	55	157	8	5	225
Valid Syl. Premises*	19	73	2	1	95
Experimental Data	85	235	0	0	320

Table 4.12: Break-down of Predictions
* The NVC Premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

The values of the parameters are shown in the following tables.

Tendencies	$T_{monotonicity}$	$T_{conversion}$	T_{all_some}	$T_{no\_somenot}$
Logarithms of Values	1.82	4.45	-0.66	-0.28

Table 4.13: Values of the Tendency Parameters.

Sentence Types	A	I	E	O
Logarithms of Informativenesses	1.11	0.33	0.19	-0.78

Table 4.14: Values of the Informativeness Parameters.

Parameter Name	Atmosphere Strength
Logarithms of Value	1.32

Table 4.15: Values of the Informativeness Parameters.

Overall Performance

The addition of the guessing events, to some degree, reshaped the model. Compared with the previous version, the performance of the model has been, in general, noticeably improved. The proportion of correct predictions reached around 95% (that of the test set/ training set/complete data set being 95.6%, 94.4% and 95.0% respectively); also the mean entropy of the predictions is quite close to that of the data (error is less that 3% on all data sets). The mean KL divergence from the data to the predictions is around 0.2. The mean entropy error dropped significantly to around 0.27 (it was, in version 2, around 0.60). The approval rate is around 0.8, which is lower than that of version 2. This drop is partially

predictable: in version 2, only the NVC conclusion and what can be proved through the more-than-complete proof system can possibly receive positive probability, which covers most items that are endorsed by most participants; however, the complete version is designed to explain why participants endorse the syllogisms that received zero probability in version 2 (i.e., those not NVC and not provable in the proof system of version 2), which consequently caused the most favored conclusion to receive lower probability.

The Informativeness Parameters

The values of the informativeness parameters, as shown in Table 4.14, allow to make an interesting observation. Recall that we assumed that informativeness determines the confidence the subject has in the premises and, hence, the probability with which he concludes 'nothing follows'. We made no assumptions on "which type of sentences" are more informative. The training results show that the amounts of informativeness follow the order:

$$A(1.11) > E(0.33) > I(0.19) > O(-0.78)$$

which completely coincides with the order proposed by Chater and Oaksford (1999). Besides, we see that sentence type "O" is exceptionally un-informative, which also agrees with the authors' proposal. The values of the informativenesses were learnt by the model. The result supports the proposal that the probabilistic validity plays an important role in human reasoning.

The Tendency Parameters

The values of the tendency parameters are sharply different from each other. Recall that these parameters are positively related to the probability that the corresponding rule is adopted. And note that the values in Table 4.13 are the logarithms. The values yield the order

$$T_{conversion}(4.45) > T_{monotonicity}(1.82) > T_{all\_some}(-0.28) > T_{no\_somenot}(-0.66)$$

The most immediate observation is that according to our model, the conversion rule is more likely to be adopted than the other rules, which is plausible since the conversion rule is simple and natural. To follow is the monotonicity rule, which is slightly surprising, since it involves two sentences and has a good reason to be cognitively more difficult than the all-some rule and the no-somenot rule. One explanation of this phenomenon is that both the all-some rule and the no-somenot rule contains the procedure to introduce new type of sentence that does not appear in the premises. In that case, the adoption of those two rules might have to employ the cognitive resources to process the new objects.

We see that $T_{conversion} = 4.45$, and that $T_{monotonicity} = 1.82$, that means, when both rules can be adopted, the probability to adopt the conversion rule is $e^{4.45-1.82} \approx 14$ times that of the probability of adopting the monotonicity rule; the probability of adopting the monotonicity rule is $e^{1.82--(0.28)} \approx 8$ times that of adopting the no-somenot rule. Only the tendencies of the all-some rule and the no-somenot rules appear comparable to one another. The first two differences are dramatic. This suggest that people simply may have a deterministic preference order among the inference rules. The result can be also consistent with a mixture of 'preference order', 'complexity', and the availability of rules in this particular reasoning task.

The Remaining Errors

Most sharp errors occur in the IE syllogisms. Our model incorrectly favors the O conclusions. The reason is, given the IE premises, quite often only one application of the additional monotonicity rule (may be accompanied by necessary adoptions of the conversion rule) is enough to results in the O conclusion. For example the IE4 premises

Some C are B
No B are A

The solution should be an adjustment of the proof system. One possible approach would be assigning higher costs to the additional monotonicity rule.

4.8.4 Summary

The complete version is based on the following theoretical assumptions.

- Sentences are the mental representation of the state of reasoning.
- Inference rules are among the basic mental operations of reasoning.
- Different inference rules are of different cognitive difficulties. The differences are reflected by the differences in the probabilities that they are adopted with.
- Adoptions of inference rules are probabilistically independent of the previous rules.
- Subjects have limited resources for reasoning (the difference is reflected by a maximum step they can make in a proof).
- Subjects make no redundant application of inference rules, at least in our context when the lengths of the proofs are sufficiently short.
- Subjects can deliberate and still err when reasoning.
- Subjects frequently adopt the illicit conversions, without realizing that the conversions are different from their licit counterparts.
- In the reasoning procedure, subjects sometimes depart from the formal approach and make a heuristic guess.
- The informative level of the premises determines the confidence level subjects have for their conclusions. That means, if they arrive at "nothing follows" in the inference part but are not confident at the premises, they are likely to abandon the conclusion from the inferences and make a guess instead.
- The atmosphere effect exist: when making a guess, subjects tend to guess the type of sentences that is supported by the atmosphere hypothesis, i.e., that is greater by the order $A < I < E < O$.

The model, making around 95% correct predictions, performs well under our means of evaluation.

Chapter 5

Further Discussion

5.1 Psychological Plausibility

In this section we discuss various psychological aspects of the model.

- **Different inference rules are of different cognitive difficulties. The differences are reflected by the differences in the probabilities that they are adopted with.**
  Assuming the mental logic framework, it is plausible that different mental operations have different cognitive complexities, as they are supposed to corresponds to different mental processes. However, whether this difference is reflected by probability remains questionable. From the subject level, it is hard to imagine that the brain is equipped with a random number generator and people adopt the operations randomly. It is even harder to imagine that if a participant is asked to infer from one pair of premises repeatedly, his conclusions approximates some probabilistic distribution. However, even if reasoning does not admit underlying probabilistic basis, it may still exhibit probabilistic behavior when we investigate the reasoning behavior of the population through psychological experiments.

  We assumed the tendency parameters to reflect some sort of inherent cognitive complexity. But is it possible what reflected is rather a kind of preference order (e.g., preferences gained through past experience)? As a matter of fact, the training result of our model indicates that the latter could be the case: the tendency parameters admits sharp differences, sometimes to the point that it could be considered as a non-probabilistic model (as the preferred inference rule could make a dominating majority of the probability). The mechanism could be a non-probabilistic one: the subject may have a strict preference order for the inference rules, and always adopt the one that is most preferred.

- **In the reasoning procedure, subjects sometimes depart from the formal approach and make a heuristic guess.**
  The plausibility of the guessing scenario is hard to evaluate. Perhaps, one possible explanation is the dual processing theory (see Evans (2012)). It might be that guessing, which uses no working memory and is fast, happens in parallel with the formal inference procedure which is deliberate and slower.

- **The assumptions on informativeness.**
  We assumed that each type of sentence has a informativeness level, and that subject make use of the value as if they know the informativeness. The question is, how do people come equipped with this knowledge? One explanation is that it becomes accumulated in daily reasoning. Based on life experience, people may become sophisticated in unconsciously evaluating the amount of information to expect from each type of sentence.

5.2 A Parallel Comparison to Other Theories of the Syllogisms

We examined the predictions of a number of existing theories of the syllogisms. We were able to obtain the predictions of the PSYCOP model from its proponent. The rest of the predictions were obtained from table 7 of Khemlani and Johnson-Laird (2012)[1]. The summary is shown in Table 5.1.

Theory	Hit	Miss	False Alarm	Correct Rejection	Correct Predictions
Atmosphere	44	41	20	215	259 /80.9%
Matching	41	44	55	180	221 /69.1%
Conversion	52	33	12	223	275 /85.9%
PHM*	40	45	63	172	212 /66.3%
PSYCOP	45	40	26	209	254 /79.4%
Verbal Models*	54	31	29	206	260 /81.2%
Mental Models*	85	0	55	180	265 /82.8%
Ver. 1 Test Data	26	15	12	107	133/83.1%
Ver. 2 Test Data	33	8	3	116	149/93.1%
Ver. 3 Complete Data**	70	14	5	231	301/94.1%
Ver. 3 Test Data	37	4	3	116	153/95.6%

Table 5.1: Predictions of the Theories of Syllogisms: A Summary.
*: Due to the limitations of the data we were able to obtain, the corresponding theory is likely to perform better than what is shown in the table.
**: The data in this line result from a cross-test: we take the predictions on the test data, then switched the test data and the training data and train the model again to get the predictions on the other half of the data.

As far as we can see from the data we had access to, we see that based on a similar version of the means of evaluation proposed by Khemlani and Johnson-Laird (2012), our model outperforms other models whose predictions are available. Yet we would like to make a comment here:our model, although intended as a first step in designing a uniform theory of reasoning, now covers only the syllogistic fragment and is highly dependent on the structure

[1]The table provided the predictions of the syllogistic theories on both the syllogisms that follow the scholastic order and the ones that do not. We obtained the data we use by restricting the conclusions to the ones that follow the scholastic order. The restriction has no influence on the predictions of the atmosphere, matching and conversion theories. However, for the PHM, the verbal model theory, and the mental model theory, we are unsure about the consequences.

of the latter; whereas some other theories in the table (e.g., the PSYCOP, the mental models theory) are intended as a uniform theory of reasoning.

5.3 Towards a Uniform Theory of Reasoning

The syllogistic fragment is a good yet small arena for theories of reasoning. It is good since it investigates a few most fundamental quantifiers; however it is far from covering all aspects of reasoning. Our model is designed just for the syllogisms. We would like to discuss the possibility of extending it to a uniform theory of reasoning.

Conceptually, our model consists of two parts: the inference part, which consists of the probabilistic natural logic and the guessing part, which takes care of the guessing events. The first part is based on the fragment of natural language for which there is a natural logic operating on its surface (see, e.g., MacCartney and Manning (2009) for another example). The second part, however, yields no immediate means of extension. Recall that the second part of our model is based on the assumptions of informativeness and the atmosphere hypothesis. We so designed the model that the subject simply guesses the conclusion. Compared with the first part, where we made precise assumptions on the underlying psychological procedure of reasoning (which should belong to the *algorithmic level* according to David Marr, see Section 2.7.4), the second part simply tells what happens during a guessing event, instead of how it happens. And yes, the second part of our model belongs to the computational level. We do not know the details of guessing as a psychological procedure.

The situation may appear messy at first glance: we are proposing a Bayesian generative model that is represented by a tree. The first part of the model is on the algorithmic level, however, closer to the leaves the tree is decorated with a lot of black-boxes, namely, the guessing events, which encapsulates the guessing procedure as computational level theories. We expect a second glance to make it a little bit more intuitive: after all, we could only give sophisticated theories (in our context, an algorithmic level theory) for the procedure we have more knowledge about; what remains has to be described in less detail. As research goes deeper these black-boxes could be open and modelled more precisely.

Now let us return to the original topic: how to extend our model toward a uniform theory of reasoning? As discussed, the guessing events are used to encapsulate what is not captured by the inference part of the model. Therefore, one way to extend our model would be, firstly, design a natural logic for a larger fragment of natural language (that does not necessarily follow traditional logical validity); secondly, extend the logic into a Bayesian generative model by determining probability distributions among the inference operations that could be applied, train the model with experimental data; thirdly, describe what could not be captured by the first part as a psychologically plausible computational level theory; thirdly, when further knowledge is acquired, investigate the computational level theories and model the cognitive task.

One crucial remark is that our model has a high requirement on the amount of experimental data. It is the relatively abundant experimental data on syllogistic reasoning that makes possible the training of a Bayesian generative model: the probability that people arrive at every possible conclusion from every possible premise set has been approximated by

experiments. For larger fragments of natural language, we definitely do not have the luxury to expect such abundance, as the amount of required data would increase dramatically with the size of the fragment. One possible way out might be to automatically extract the instances of reasoning that is encoded in natural language corpora, which is another challenging task.

Chapter 6

Conclusion

6.1 Summary and Future Work

We designed a probabilistic logical model for syllogistic reasoning. The model combines the ideas from the mental logic, the natural logic approach, the probabilistic basis of reasoning and the atmosphere hypothesis. Our theory infers not only the valid conclusions but also some invalid ones.
We trained the model and evaluated it in multiple ways. The performance of the model is promising, making around 95% correct predictions. As far as we can see, our model outperformed all other theories of the syllogistic reasoning whose predictions were available to us. The results indicate that natural logic based probabilistic model may be a reasonable candidate for the theory of reasoning.

One interesting future direction is to try to extend the theory to richer fragments of natural language (see Section 5.3). Besides, more benchmarks for evaluating are also needed.

6.2 Coda: Theory of Reasoning, a Mosaic?

Elegant theories has been popular in the scientific world. Simple and universal theories are favored by many researchers. Every pair of particles obey the law of gravitation. All recursive functions are computed by a Turing machine. Every member of the market is rational. It hence appears natural to seek an elegant theory of reasoning. For now, the mental model theorists propose that the mental representation of reasoning is iconic; the mental logic proposal assumes that the state of reasoning is represented by formulas; the proponents of the probabilistic paradigm propose that the computational level theory should be a kind of probabilistic calculus. Unfortunately, these proposals, although all being psychologically plausible and capturing some aspects of human reasoning, are not completely consistent with each other. Besides, reasoning is no simple procedure. Subjects differ in their reasoning strategies; even one particular subject may keep switching his strategies. It makes perfect sense to use natural language to reason when trying to convince a friend that not everyone is rational and switch to graphs or icons when imagining the scene described in an exercise about classical mechanics or models of set theory.

The point we want to make here is that various theories might be able to coexist harmoniously with each other The mechanism of reasoning may be a mosaic of different paradigms,

switching from one to another when the reasoning task changes. As van Benthem (2008) wrote,

> ... it is a telling fact that mathematicians have never abandoned natural language in favor of logical formalisms... mathematicians use mixtures of both, with the logical notation coming in dynamically when natural language needs to be made more precise. This mixture suggests that 'natural logic' and 'modern logic' can coexist harmoniously, because both have their place...

The way each subject reason may, to some degree, be shaped by the previous experiences, which depend on a random yet particular life trajectory; these life experiences further depend on the living environments of the subject that are shaped by a particular history of the society. It may well be that people trigger a mental proof to reason about history of logic and adopt an iconic representation when puzzling about models of set theory. To summarize, in terms of a uniform theory of reasoning, what we expect is not an elegant and concise theory, but rather a mosaic of plausible theories, one that form some shape at first glance but highly depends on the particular task of reasoning. That is why in this thesis we rather focused on investigating how much natural logic can help in building the model of syllogistic reasoning than developing another grand theory of reasoning.

Acknowledgements

I hereby record my gratitude to those who I have been long in debt to.
First of all come my family, especially my parents and my wife. Their love , company and and support made me feel like home in the small, unfamiliar foreign city of Amsterdam.

I thank Feng Ye for showing me the elegance of formal logical theories for the first time which made me determined to become a logician. I thank Yanjing Wang for letting me know about the opportunity to join the ILLC and helping with the application.

I thank the Evert Willem Beth Foundation for generously providing me with a scholarship.

The ILLC is definitely an outstanding place for interdisciplinary study and research. I owe every member of the ILLC for creating such an excellent environment. Among them I thank Benedikt Löwe and Yde Venema for their excellent lectures.

I owe my thesis supervisor a lot. To begin with, it is always a pleasant to work with Jakub Szymanik, who supervised my master thesis and also a research project. He is super friendly, helpful and humorous, willing to sacrifice his weekend to take the trouble to review my messy documents written in weird English, and comment with his lovely accent and hand-writing. He brought me a lot of interesting questions of research and without him I would not have finished my study.

Ivan Titov co-supervised my thesis project. As an expert in NLP he took the great trouble to communicate with me who knew little about what algorithm we need to use and knew little machine learning terminology in English. Going through the interdisciplinary barrier, he was able to enable me to implement the EM algorithm and provided enlightening ideas for my thesis project.

I thank Maria Aloni for chairing my defense, and all other committee members, i.e., Jelle Zuidema, Johan van Benthem, Robert van Rooij, for reviewing my messy thesis. I thank Henk Zeevat for joining the defense.

Thank everyone mentioned above again. Your help was really indispensable in my life. It is a deep pleasure to know everyone of you.

References

Antoniou, G. (1997). *Nonmonotonic reasoning*. MIT Press.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, pages 164–171.

Begg, I. and Denny, J. P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology*, 81(2):351.

van Benthem, J. (1986). *Essays in logical semantics*. Number 29. Springer.

van Benthem, J. (1987). Meaning: interpretation and inference. *Synthese*, 73(3):451–470.

van Benthem, J. (2008). A brief history of natural logic. *Logic, Navya-Nyaya and Applications*.

Van Benthem, J. (2008). Logic and reasoning: do the facts matter? *Studia Logica*, 88(1):67–84.

Braine, M. D. and O'Brien, D. P. (1998). *Mental logic*. Psychology Press.

Chapman, L. J. and Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, 58(3):220.

Chater, N. and Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, 38(2):191–258.

Chater, N. and Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38.

Dickstein, L. S. (1981). Conversion and possibility in syllogistic reasoning. *Bulletin of the Psychonomic Society*, 18(5):229–232.

Dotlačil, J., Szymanik, J., and Zajenkowski, M. (2014). Probabilistic semantic automata in the verification of quantified statements. In P. Bello, M.McShane, M. G. B. S., editor, *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 1778–1783.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459.

Evans, J. S. B. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, 18(1):5–31.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, 86(3):223–251.

Gierasimczuk, N., Van der Maas, H. L., and Raijmakers, M. E. (2013). An analytic tableaux model for deductive mastermind empirically tested with a massively used online learning system. *Journal of Logic, Language and Information*, 22(3):297–314.

Griggs, R. A. and Cox, J. R. (1982). The elusive thematic-materials effect in wason's selection task. *British Journal of Psychology*, 73(3):407–420.

Icard III, T. and Moss, L. (2014). Recent progress in monotonicity. *LiLT (Linguistic Issues in Language Technology)*, 9.

Isaac, A. M., Szymanik, J., and Verbrugge, R. (2014). Logic and complexity in cognitive science. In *Johan van Benthem on Logic and Information Dynamics*, pages 787–824. Springer.

Johnson-Laird, P., Khemlani, S. S., and Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, 19(4):201–214.

Johnson-Laird, P. N. (1975). Models of deduction. *Reasoning: Representation and process in children and adults*, pages 7–54.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

Johnson-Laird, P. N. (1986). *Mental models*. Number 6. Harvard University Press.

Johnson-Laird, P. N. (1997). An end to the controversy? a reply to rips. *Minds and Machines*, 7(3):425–432.

Johnson-Laird, P. N. and Byrne, R. M. (1989). Only reasoning. *Journal of Memory and Language*, 28(3):313–330.

Johnson-Laird, P. N. and Byrne, R. M. (1991). *Deduction*. Lawrence Erlbaum Associates, Inc.

Kahneman, D. and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49.

Khemlani, S. and Johnson-Laird, P. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 138(3):427.

Khemlani, S. and Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, 4(1):4–20.

MacCartney, B. and Manning, C. D. (2009). An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics*, pages 140–156. Association for Computational Linguistics.

Marr, D. and Vision, A. (1982). A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company.*

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, 1(2):185–196.

Oaksford, M. and Chater, N. (1996). Rational explanation of the selection task.

Oaksford, M. and Chater, N. (2007). *Bayesian rationality the probabilistic approach to human reasoning.* Oxford University Press.

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning*, 19(3-4):329–345.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior*, 14(2):180–195.

Rips, L. J. (1994). The psychology of proof.

Sells, S. B. (1936). The atmosphere effect: an experimental study of reasoning. *Archives of Psychology (Columbia University).*

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3.

Stenning, K. and Van Lambalgen, M. (2008). *Human reasoning and cognitive science.* MIT Press.

Verbrugge, R. (2009). Logic and social cognition. *Journal of Philosophical Logic*, 38(6):649–680.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, 20(3):273–281.

Wason, P. C. and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology*, 23(1):63–71.

Wilkins, M. C. (1929). The effect of changed material on ability to do formal syllogistic reasoning. *Archives of Psychology.*

Woodworth, R. S. and Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, 18(4):451.

Appendix A

The Predictions Of the Model With the Experimental Data

A.1 Predictions: Version 1

Premises	A	I	E	O	NVC
AA1	37.0*	27.0*	0.0	0.0	36.0
AA1	90.0	5.0	0.0	0.0	5.0
AA2	0.0	0.0	0.0	0.0	100.0
AA2	58.0	8.0	1.0	1.0	32.0
AA3	0.0	24.0*	0.0	0.0	76.0
AA3	57.0	29.0	0.0	0.0	14.0
AA4	0.0	14.0*	0.0	0.0	86.0
AA4	75.0	16.0	1.0	1.0	7.0
AI1	0.0	68.0*	0.0	0.0	32.0
AI1	0.0	92.0	3.0	3.0	2.0
AI2	0.0	0.0	0.0	0.0	100.0
AI2	0.0	57.0	3.0	11.0	29.0
AI3	0.0	66.0*	0.0	0.0	34.0
AI3	1.0	89.0	1.0	3.0	6.0
AI4	0.0	0.0	0.0	0.0	100.0
AI4	0.0	71.0	0.0	1.0	28.0
AE1	0.0	0.0	0.0	0.0	100.0
AE1	0.0	3.0	59.0	6.0	32.0
AE2	0.0	0.0	32.0*	6.0*	62.0
AE2	0.0	0.0	88.0	1.0	11.0
AE3	0.0	0.0	0.0	0.0	100.0
AE3	0.0	1.0	61.0	13.0	25.0
AE4	0.0	0.0	30.0*	4.0*	66.0
AE4	0.0	3.0	87.0	2.0	8.0
AO1	0.0	0.0	0.0	0.0	100.0
AO1	1.0	6.0	1.0	57.0	35.0
AO2	0.0	0.0	0.0	100.0*	0.0
AO2	0.0	6.0	3.0	67.0	24.0
AO3	0.0	0.0	0.0	0.0	100.0

AO3	0.0	10.0	0.0	66.0	24.0
AO4	0.0	0.0	0.0	0.0	100.0
AO4	0.0	5.0	3.0	72.0	20.0
IA1	0.0	0.0	0.0	0.0	100.0
IA1	0.0	72.0	0.0	6.0	22.0
IA2	0.0	0.0	0.0	0.0	100.0
IA2	13.0	49.0	3.0	12.0	23.0
IA3	0.0	68.0*	0.0	0.0	32.0
IA3	2.0	85.0	1.0	4.0	8.0
IA4	0.0	66.0*	0.0	0.0	34.0
IA4	0.0	91.0	1.0	1.0	7.0
II1	0.0	0.0	0.0	0.0	100.0
II1	0.0	41.0	3.0	4.0	52.0
II2	0.0	0.0	0.0	0.0	100.0
II2	1.0	42.0	3.0	3.0	51.0
II3	0.0	0.0	0.0	0.0	100.0
II3	0.0	24.0	3.0	1.0	72.0
II4	0.0	0.0	0.0	0.0	100.0
II4	0.0	42.0	0.0	1.0	57.0
IE1	0.0	0.0	0.0	0.0	100.0
IE1	1.0	1.0	22.0	16.0	60.0
IE2	0.0	0.0	0.0	0.0	100.0
IE2	0.0	0.0	39.0	30.0	31.0
IE3	0.0	0.0	0.0	0.0	100.0
IE3	0.0	1.0	30.0	33.0	36.0
IE4	0.0	0.0	0.0	0.0	100.0
IE4	0.0	1.0	28.0	44.0	27.0
IO1	0.0	0.0	0.0	0.0	100.0
IO1	3.0	4.0	1.0	30.0	62.0
IO2	0.0	0.0	0.0	0.0	100.0
IO2	1.0	5.0	4.0	37.0	53.0
IO3	0.0	0.0	0.0	0.0	100.0
IO3	0.0	9.0	1.0	29.0	61.0
IO4	0.0	0.0	0.0	0.0	100.0
IO4	0.0	5.0	1.0	44.0	50.0
EA1	0.0	0.0	32.0*	11.0*	57.0
EA1	0.0	1.0	87.0	3.0	9.0
EA2	0.0	0.0	30.0*	8.0*	62.0
EA2	0.0	0.0	89.0	3.0	8.0
EA3	0.0	0.0	0.0	11.0*	89.0
EA3	0.0	0.0	64.0	22.0	14.0
EA4	0.0	0.0	0.0	8.0*	92.0
EA4	1.0	3.0	61.0	8.0	27.0
EI1	0.0	0.0	0.0	43.0*	57.0
EI1	0.0	5.0	15.0	66.0	14.0
EI2	0.0	0.0	0.0	35.0*	65.0
EI2	1.0	1.0	21.0	52.0	25.0

EI3	0.0	0.0	0.0	35.0*	65.0
EI3	0.0	6.0	15.0	48.0	31.0
EI4	0.0	0.0	0.0	32.0*	68.0
EI4	0.0	2.0	32.0	27.0	39.0
EE1	0.0	0.0	0.0	0.0	100.0
EE1	0.0	1.0	34.0	1.0	64.0
EE2	0.0	0.0	0.0	0.0	100.0
EE2	3.0	3.0	14.0	3.0	77.0
EE3	0.0	0.0	0.0	0.0	100.0
EE3	0.0	0.0	18.0	3.0	79.0
EE4	0.0	0.0	0.0	0.0	100.0
EE4	0.0	3.0	31.0	1.0	65.0
EO1	0.0	0.0	0.0	0.0	100.0
EO1	1.0	8.0	8.0	23.0	60.0
EO2	0.0	0.0	0.0	0.0	100.0
EO2	0.0	13.0	7.0	11.0	69.0
EO3	0.0	0.0	0.0	0.0	100.0
EO3	0.0	0.0	9.0	28.0	63.0
EO4	0.0	0.0	0.0	0.0	100.0
EO4	0.0	5.0	8.0	12.0	75.0
OA1	0.0	0.0	0.0	0.0	100.0
OA1	0.0	3.0	3.0	68.0	26.0
OA2	0.0	0.0	0.0	0.0	100.0
OA2	0.0	11.0	5.0	56.0	28.0
OA3	0.0	0.0	0.0	100.0*	0.0
OA3	0.0	15.0	3.0	69.0	13.0
OA4	0.0	0.0	0.0	0.0	100.0
OA4	1.0	3.0	6.0	27.0	63.0
OI1	0.0	0.0	0.0	0.0	100.0
OI1	4.0	6.0	0.0	35.0	55.0
OI2	0.0	0.0	0.0	0.0	100.0
OI2	0.0	8.0	3.0	35.0	54.0
OI3	0.0	0.0	0.0	0.0	100.0
OI3	1.0	9.0	1.0	31.0	58.0
OI4	0.0	0.0	0.0	0.0	100.0
OI4	3.0	8.0	2.0	29.0	58.0
OE1	0.0	0.0	0.0	0.0	100.0
OE1	1.0	0.0	14.0	5.0	80.0
OE2	0.0	0.0	0.0	0.0	100.0
OE2	0.0	8.0	11.0	16.0	65.0
OE3	0.0	0.0	0.0	0.0	100.0
OE3	0.0	5.0	12.0	18.0	65.0
OE4	0.0	0.0	0.0	0.0	100.0
OE4	0.0	19.0	9.0	14.0	58.0
OO1	0.0	0.0	0.0	0.0	100.0
OO1	1.0	8.0	1.0	22.0	68.0
OO2	0.0	0.0	0.0	0.0	100.0

OO2	0.0	16.0	5.0	10.0	69.0
OO3	0.0	0.0	0.0	0.0	100.0
OO3	1.0	6.0	0.0	15.0	78.0
OO4	0.0	0.0	0.0	0.0	100.0
OO4	1.0	4.0	1.0	25.0	69.0

Table A.1: The Predictions of Version 1 of the Model. First line is the prediction, followed by another line that gives the experimental data. A * indicates that the conclusion is valid.

A.2 Predictions: Version 2

Premises	A	I	E	O	NVC
AA1	86.0*	1.0*	0.0	0.0	14.0
AA1	90.0	5.0	0.0	0.0	5.0
AA2	85.0	0.0	0.0	0.0	15.0
AA2	58.0	8.0	1.0	1.0	32.0
AA3	85.0	0.0*	0.0	0.0	15.0
AA3	57.0	29.0	0.0	0.0	14.0
AA4	85.0	0.0*	0.0	0.0	14.0
AA4	75.0	16.0	1.0	1.0	7.0
AI1	0.0	86.0*	0.0	0.0	14.0
AI1	0.0	92.0	3.0	3.0	2.0
AI2	0.0	85.0	0.0	0.0	15.0
AI2	0.0	57.0	3.0	11.0	29.0
AI3	0.0	86.0*	0.0	0.0	14.0
AI3	1.0	89.0	1.0	3.0	6.0
AI4	0.0	85.0	0.0	0.0	15.0
AI4	0.0	71.0	0.0	1.0	28.0
AE1	0.0	0.0	69.0	1.0	31.0
AE1	0.0	3.0	59.0	6.0	32.0
AE2	0.0	0.0	72.0*	1.0*	28.0
AE2	0.0	0.0	88.0	1.0	11.0
AE3	0.0	0.0	69.0	0.0	31.0
AE3	0.0	1.0	61.0	13.0	25.0
AE4	0.0	0.0	71.0*	1.0*	28.0
AE4	0.0	3.0	87.0	2.0	8.0
AO1	0.0	0.0	0.0	85.0	15.0
AO1	1.0	6.0	1.0	57.0	35.0
AO2	0.0	0.0	0.0	86.0*	14.0
AO2	0.0	6.0	3.0	67.0	24.0
AO3	0.0	0.0	0.0	86.0	14.0
AO3	0.0	10.0	0.0	66.0	24.0
AO4	0.0	0.0	0.0	85.0	15.0
AO4	0.0	5.0	3.0	72.0	20.0

IA1	0.0	85.0	0.0	0.0	15.0
IA1	0.0	72.0	0.0	6.0	22.0
IA2	0.0	85.0	0.0	0.0	15.0
IA2	13.0	49.0	3.0	12.0	23.0
IA3	0.0	86.0*	0.0	0.0	14.0
IA3	2.0	85.0	1.0	4.0	8.0
IA4	0.0	86.0*	0.0	0.0	14.0
IA4	0.0	91.0	1.0	1.0	7.0
II1	0.0	0.0	0.0	0.0	100.0
II1	0.0	41.0	3.0	4.0	52.0
II2	0.0	0.0	0.0	0.0	100.0
II2	1.0	42.0	3.0	3.0	51.0
II3	0.0	0.0	0.0	0.0	100.0
II3	0.0	24.0	3.0	1.0	72.0
II4	0.0	0.0	0.0	0.0	100.0
II4	0.0	42.0	0.0	1.0	57.0
IE1	0.0	0.0	0.0	64.0	36.0
IE1	1.0	1.0	22.0	16.0	60.0
IE2	0.0	0.0	0.0	65.0	35.0
IE2	0.0	0.0	39.0	30.0	31.0
IE3	0.0	0.0	0.0	65.0	35.0
IE3	0.0	1.0	30.0	33.0	36.0
IE4	0.0	0.0	0.0	68.0	32.0
IE4	0.0	1.0	28.0	44.0	27.0
IO1	0.0	0.0	0.0	0.0	100.0
IO1	3.0	4.0	1.0	30.0	62.0
IO2	0.0	0.0	0.0	0.0	100.0
IO2	1.0	5.0	4.0	37.0	53.0
IO3	0.0	0.0	0.0	0.0	100.0
IO3	0.0	9.0	1.0	29.0	61.0
IO4	0.0	0.0	0.0	0.0	100.0
IO4	0.0	5.0	1.0	44.0	50.0
EA1	0.0	0.0	72.0*	1.0*	28.0
EA1	0.0	1.0	87.0	3.0	9.0
EA2	0.0	0.0	71.0*	1.0*	28.0
EA2	0.0	0.0	89.0	3.0	8.0
EA3	0.0	0.0	69.0	1.0*	31.0
EA3	0.0	0.0	64.0	22.0	14.0
EA4	0.0	0.0	69.0	1.0*	31.0
EA4	1.0	3.0	61.0	8.0	27.0
EI1	0.0	0.0	0.0	70.0*	30.0
EI1	0.0	5.0	15.0	66.0	14.0
EI2	0.0	0.0	0.0	67.0*	33.0
EI2	1.0	1.0	21.0	52.0	25.0
EI3	0.0	0.0	0.0	67.0*	33.0
EI3	0.0	6.0	15.0	48.0	31.0
EI4	0.0	0.0	0.0	66.0*	34.0

EI4	0.0	2.0	32.0	27.0	39.0
EE1	0.0	0.0	0.0	0.0	100.0
EE1	0.0	1.0	34.0	1.0	64.0
EE2	0.0	0.0	0.0	0.0	100.0
EE2	3.0	3.0	14.0	3.0	77.0
EE3	0.0	0.0	0.0	0.0	100.0
EE3	0.0	0.0	18.0	3.0	79.0
EE4	0.0	0.0	0.0	0.0	100.0
EE4	0.0	3.0	31.0	1.0	65.0
EO1	0.0	0.0	0.0	0.0	100.0
EO1	1.0	8.0	8.0	23.0	60.0
EO2	0.0	0.0	0.0	0.0	100.0
EO2	0.0	13.0	7.0	11.0	69.0
EO3	0.0	0.0	0.0	0.0	100.0
EO3	0.0	0.0	9.0	28.0	63.0
EO4	0.0	0.0	0.0	0.0	100.0
EO4	0.0	5.0	8.0	12.0	75.0
OA1	0.0	0.0	0.0	85.0	15.0
OA1	0.0	3.0	3.0	68.0	26.0
OA2	0.0	0.0	0.0	86.0	14.0
OA2	0.0	11.0	5.0	56.0	28.0
OA3	0.0	0.0	0.0	86.0*	14.0
OA3	0.0	15.0	3.0	69.0	13.0
OA4	0.0	0.0	0.0	85.0	15.0
OA4	1.0	3.0	6.0	27.0	63.0
OI1	0.0	0.0	0.0	0.0	100.0
OI1	4.0	6.0	0.0	35.0	55.0
OI2	0.0	0.0	0.0	0.0	100.0
OI2	0.0	8.0	3.0	35.0	54.0
OI3	0.0	0.0	0.0	0.0	100.0
OI3	1.0	9.0	1.0	31.0	58.0
OI4	0.0	0.0	0.0	0.0	100.0
OI4	3.0	8.0	2.0	29.0	58.0
OE1	0.0	0.0	0.0	0.0	100.0
OE1	1.0	0.0	14.0	5.0	80.0
OE2	0.0	0.0	0.0	0.0	100.0
OE2	0.0	8.0	11.0	16.0	65.0
OE3	0.0	0.0	0.0	0.0	100.0
OE3	0.0	5.0	12.0	18.0	65.0
OE4	0.0	0.0	0.0	0.0	100.0
OE4	0.0	19.0	9.0	14.0	58.0
OO1	0.0	0.0	0.0	0.0	100.0
OO1	1.0	8.0	1.0	22.0	68.0
OO2	0.0	0.0	0.0	0.0	100.0
OO2	0.0	16.0	5.0	10.0	69.0
OO3	0.0	0.0	0.0	0.0	100.0
OO3	1.0	6.0	0.0	15.0	78.0

OO4	0.0	0.0	0.0	0.0	100.0
OO4	1.0	4.0	1.0	25.0	69.0

Table A.2: The Predictions of Version 2 of the Model. First line is the prediction, followed by another line that gives the experimental data. A * indicates that the conclusion is valid.

A.3 Predictions: Version 3

Premises	A	I	E	O	NVC
AA1	85.0*	1.0*	1.0	1.0	13.0
AA1	90.0	5.0	0.0	0.0	5.0
AA2	83.0	1.0	1.0	1.0	14.0
AA2	58.0	8.0	1.0	1.0	32.0
AA3	83.0	1.0*	1.0	1.0	14.0
AA3	57.0	29.0	0.0	0.0	14.0
AA4	84.0	1.0*	1.0	1.0	13.0
AA4	75.0	16.0	1.0	1.0	7.0
AI1	1.0	84.0*	1.0	1.0	13.0
AI1	0.0	92.0	3.0	3.0	2.0
AI2	1.0	82.0	1.0	1.0	14.0
AI2	0.0	57.0	3.0	11.0	29.0
AI3	1.0	84.0*	1.0	1.0	13.0
AI3	1.0	89.0	1.0	3.0	6.0
AI4	1.0	82.0	1.0	1.0	14.0
AI4	0.0	71.0	0.0	1.0	28.0
AE1	1.0	1.0	68.0	1.0	29.0
AE1	0.0	3.0	59.0	6.0	32.0
AE2	1.0	1.0	71.0*	2.0*	26.0
AE2	0.0	0.0	88.0	1.0	11.0
AE3	1.0	1.0	68.0	1.0	29.0
AE3	0.0	1.0	61.0	13.0	25.0
AE4	1.0	1.0	71.0*	1.0*	26.0
AE4	0.0	3.0	87.0	2.0	8.0
AO1	1.0	1.0	1.0	83.0	14.0
AO1	1.0	6.0	1.0	57.0	35.0
AO2	1.0	1.0	1.0	84.0*	13.0
AO2	0.0	6.0	3.0	67.0	24.0
AO3	1.0	1.0	1.0	83.0	13.0
AO3	0.0	10.0	0.0	66.0	24.0
AO4	1.0	1.0	1.0	83.0	14.0
AO4	0.0	5.0	3.0	72.0	20.0
IA1	1.0	82.0	1.0	1.0	14.0
IA1	0.0	72.0	0.0	6.0	22.0
IA2	1.0	82.0	1.0	1.0	14.0

IA2	13.0	49.0	3.0	12.0	23.0
IA3	1.0	84.0*	1.0	1.0	13.0
IA3	2.0	85.0	1.0	4.0	8.0
IA4	1.0	84.0*	1.0	1.0	13.0
IA4	0.0	91.0	1.0	1.0	7.0
II1	10.0	39.0	10.0	10.0	30.0
II1	0.0	41.0	3.0	4.0	52.0
II2	10.0	39.0	10.0	10.0	30.0
II2	1.0	42.0	3.0	3.0	51.0
II3	10.0	39.0	10.0	10.0	30.0
II3	0.0	24.0	3.0	1.0	72.0
II4	10.0	39.0	10.0	10.0	30.0
II4	0.0	42.0	0.0	1.0	57.0
IE1	2.0	2.0	6.0	59.0	32.0
IE1	1.0	1.0	22.0	16.0	60.0
IE2	2.0	2.0	6.0	60.0	31.0
IE2	0.0	0.0	39.0	30.0	31.0
IE3	2.0	2.0	6.0	60.0	31.0
IE3	0.0	1.0	30.0	33.0	36.0
IE4	1.0	1.0	6.0	63.0	29.0
IE4	0.0	1.0	28.0	44.0	27.0
IO1	9.0	9.0	9.0	34.0	39.0
IO1	3.0	4.0	1.0	30.0	62.0
IO2	9.0	9.0	9.0	34.0	39.0
IO2	1.0	5.0	4.0	37.0	53.0
IO3	9.0	9.0	9.0	34.0	39.0
IO3	0.0	9.0	1.0	29.0	61.0
IO4	9.0	9.0	9.0	34.0	39.0
IO4	0.0	5.0	1.0	44.0	50.0
EA1	1.0	1.0	71.0*	2.0*	26.0
EA1	0.0	1.0	87.0	3.0	9.0
EA2	1.0	1.0	71.0*	2.0*	26.0
EA2	0.0	0.0	89.0	3.0	8.0
EA3	1.0	1.0	68.0	2.0*	28.0
EA3	0.0	0.0	64.0	22.0	14.0
EA4	1.0	1.0	68.0	2.0*	29.0
EA4	1.0	3.0	61.0	8.0	27.0
EI1	1.0	1.0	6.0	64.0*	27.0
EI1	0.0	5.0	15.0	66.0	14.0
EI2	2.0	2.0	6.0	62.0*	29.0
EI2	1.0	1.0	21.0	52.0	25.0
EI3	2.0	2.0	6.0	62.0*	29.0
EI3	0.0	6.0	15.0	48.0	31.0
EI4	2.0	2.0	6.0	61.0*	30.0
EI4	0.0	2.0	32.0	27.0	39.0
EE1	2.0	2.0	8.0	2.0	85.0
EE1	0.0	1.0	34.0	1.0	64.0

EE2	2.0	2.0	8.0	2.0	85.0
EE2	3.0	3.0	14.0	3.0	77.0
EE3	2.0	2.0	8.0	2.0	85.0
EE3	0.0	0.0	18.0	3.0	79.0
EE4	2.0	2.0	8.0	2.0	85.0
EE4	0.0	3.0	31.0	1.0	65.0
EO1	4.0	4.0	4.0	13.0	76.0
EO1	1.0	8.0	8.0	23.0	60.0
EO2	4.0	4.0	4.0	13.0	76.0
EO2	0.0	13.0	7.0	11.0	69.0
EO3	4.0	4.0	4.0	13.0	76.0
EO3	0.0	0.0	9.0	28.0	63.0
EO4	4.0	4.0	4.0	13.0	76.0
EO4	0.0	5.0	8.0	12.0	75.0
OA1	1.0	1.0	1.0	83.0	14.0
OA1	0.0	3.0	3.0	68.0	26.0
OA2	1.0	1.0	1.0	83.0	13.0
OA2	0.0	11.0	5.0	56.0	28.0
OA3	1.0	1.0	1.0	84.0*	13.0
OA3	0.0	15.0	3.0	69.0	13.0
OA4	1.0	1.0	1.0	83.0	14.0
OA4	1.0	3.0	6.0	27.0	63.0
OI1	9.0	9.0	9.0	34.0	39.0
OI1	4.0	6.0	0.0	35.0	55.0
OI2	9.0	9.0	9.0	34.0	39.0
OI2	0.0	8.0	3.0	35.0	54.0
OI3	9.0	9.0	9.0	34.0	39.0
OI3	1.0	9.0	1.0	31.0	58.0
OI4	9.0	9.0	9.0	34.0	39.0
OI4	3.0	8.0	2.0	29.0	58.0
OE1	4.0	4.0	4.0	13.0	76.0
OE1	1.0	0.0	14.0	5.0	80.0
OE2	4.0	4.0	4.0	13.0	76.0
OE2	0.0	8.0	11.0	16.0	65.0
OE3	4.0	4.0	4.0	13.0	76.0
OE3	0.0	5.0	12.0	18.0	65.0
OE4	4.0	4.0	4.0	13.0	76.0
OE4	0.0	19.0	9.0	14.0	58.0
OO1	6.0	6.0	6.0	23.0	58.0
OO1	1.0	8.0	1.0	22.0	68.0
OO2	6.0	6.0	6.0	23.0	58.0
OO2	0.0	16.0	5.0	10.0	69.0
OO3	6.0	6.0	6.0	23.0	58.0
OO3	1.0	6.0	0.0	15.0	78.0
OO4	6.0	6.0	6.0	23.0	58.0
OO4	1.0	4.0	1.0	25.0	69.0

Table A.3: The Predictions of the complete version of the model. First line is the prediction, followed by another line that gives the experimental data. A * indicates that the conclusion is valid.

A.4 Predictions of Other Theories of the Syllogisms

Syllogisms	Atm.	Mat.	Conv.	PHM	PSYCOP	V.Models	M.Models	Data
AA1	A	A	A	A,I	A,I	A	A	A
AA2	A	A	A	A,I	NVC	I,NVC	A,I,NVC	A,NVC
AA3	A	A	A	A,I	I	NVC	A,I	A
AA4	A	A	A	A,I	I	NVC	A,I	A
AI1	I	I,O	NVC	I,O	I,O	I	I	I
AI2	I	I,O	NVC	I,O	NVC	I,NVC	I,NVC	I
AI3	I	I,O	NVC	NVC	I,O	I,NVC	I	I
AI4	I	I,O	NVC	NVC	NVC	NVC	I,NVC	I
AE1	E	E	NVC	E,O	I	NVC	E,O,NVC	E,NVC
AE2	E	E	NVC	E,O	E,O	NVC	E	E
AE3	E	E	NVC	NVC	I	NVC	E,O,NVC	E
AE4	E	E	NVC	NVC	E,O	NVC	E	E
AO1	O	I,O	NVC	I.O	NVC	NVC	O,NVC	O,NVC
AO2	O	I,O	NVC	I,O	I,O	NVC	O,NVC	O
AO3	O	I,O	NVC	NVC	I	NVC	O,NVC	O
AO4	O	I,O	NVC	NVC	NVC	NVC	O,NVC	O
IA1	I	I,O	I	I,O	NVC	I	I,NVC	I
IA2	I	I,O	I	NVC	NVC	I,NVC	I,NVC	I
IA3	I	I,O	I	I,O	I,O	NVC	I	I
IA4	I	I,O	I	NVC	I,O	NVC	I	I
II1	I	I,O	NVC	I,O	NVC	I	I,NVC	I,NVC
II2	I	I,O	NVC	I,O	NVC	NVC	I,NVC	I,NVC
II3	I	I,O	NVC	I,O	NVC	NVC	I,NVC	NVC
II4	I	I,O	NVC	I,O	NVC	NVC	I,NVC	I,NVC
IE1	O	E	NVC	E,O	I	NVC	E,O,NVC	NVC
IE2	O	E	NVC	E,O	I	NVC	E,O,NVC	E,O,NVC
IE3	O	E	NVC	NVC	I	NVC	E,O,NVC	E,O,NVC
IE4	O	E	NVC	NVC	I	NVC	E,O,NVC	O,NVC
IO1	O	I,O	NVC	I,O	NVC	NVC	O,NVC	O,NVC
IO2	O	I,O	NVC	I,O	NVC	NVC	O,NVC	O,NVC
IO3	O	I,O	NVC	NVC	NVC	NVC	O,NVC	NVC
IO4	O	I,O	NVC	NVC	NVC	NVC	O,NVC	O,NVC
EA1	E	E	E	E,O	E,I,O	E	E	E
EA2	E	E	E	NVC	E	E,NVC	E	E
EA3	E	E	E	E,O	I,O	E,NVC	E,O,NVC	E

EA4	E	E	E	NVC	I,O	NVC	E,O,NVC	E
EI1	O	E	O	E,O	O,I	O	O,E,NVC	O
EI2	O	E	O	E	O,I	O,I,NVC	O,E,NVC	O
EI3	O	E	O	E	O,I	O,NVC	O,E,NVC	O,NVC
EI4	O	E	O	E	O,I	O,NVC	O,E,NVC	E,NVC
EE1	E	E	NVC	E,O	NVC	E,NVC	E,NVC	E,NVC
EE2	E	E	NVC	E,O	NVC	I,NVC	E,NVC	NVC
EE3	E	E	NVC	E,O	NVC	NVC	E,NVC	NVC
EE4	E	E	NVC	E,O	NVC	NVC	E,NVC	E,NVC
EO1	O	E	NVC	O,I	NVC	O,NVC	E,O,NVC	NVC
EO2	O	E	NVC	I,O	NVC	NVC	E,O,NVC	NVC
EO3	O	E	NVC	NVC	NVC	NVC	E,O,NVC	NVC
EO4	O	E	NVC	NVC	NVC	NVC	E,O,NVC	NVC
OA1	O	I,O	O	I,O	NVC	O	O,NVC	O
OA2	O	I,O	O	NVC	I	NVC	O,NVC	O
OA3	O	I,O	O	I,O	O	O,NVC	O,NVC	O
OA4	O	I,O	O	NVC	NVC	NVC	O,NVC	NVC
OI1	O	I,O	NVC	I,O	NVC	I,O	O,NVC	O,NVC
OI2	O	I,O	NVC	NVC	NVC	I,NVC	O,NVC	O,NVC
OI3	O	I,O	NVC	I,O	NVC	O,NVC	O,NVC	O,NVC
OI4	O	I,O	NVC	NVC	NVC	NVC	O,NVC	NVC
OE1	O	E	NVC	I,O	NVC	O,NVC	E,O,NVC	NVC
OE2	O	E	NVC	NVC	NVC	I,NVC	E,O,NVC	NVC
OE3	O	E	NVC	I,O	NVC	NVC	E,O,NVC	NVC
OE4	O	E	NVC	NVC	NVC	NVC	E,O,NVC	NVC
OO1	O	I,O	NVC	I,O	NVC	NVC	O,NVC	NVC
OO2	O	I,O	NVC	I,O	NVC	NVC	O,NVC	NVC
OO3	O	I,O	NVC	I,O	NVC	NVC	O,NVC	NVC
OO4	O	I,O	NVC	I,O	NVC	NVC	O,NVC	NVC

Table A.4: Predicted Responses for Each Syllogisms From Eight Theories of Syllogistic Reasoning.
Atm. = atmosphere; Mat. = maching; Conv. = conversion; V.Models = verbal models; M.Models = mental models; G.Model = generative model. Data from the first seven theories are from Khemlani and Johnson-Laird (2012)