INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

ORDER EFFECTS IN HUMAN BELIEF REVISION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Hongbin Wang, M.S., M.A., B.S.

* * * * *

The Ohio State University
1998

Dissertation Committee:

Professor Jiajie Zhang, Adviser

Professor Richard Jagacinski, Co-Adviser

Professor Neal Johnson

Professor Todd Johnson

Approved by

Co-Advisers
Department of Psychology

UMI Number: 9900929

UMI Microform 9900929

Copyright 1998, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI

**300 North Zeeb Road
Ann Arbor, MI 48103**

Copyright by
Hongbin Wang
1998

ABSTRACT

Human belief revision, a process in which one revises one's current belief in the light of new information, is ubiquitous in everyday life and scientific discovery. The order effect, a phenomenon in which the final belief is significantly affected by the temporal order of information presentation, is a robust empirical finding in human belief revision. Normative theories, such as Bayes' Theorem, have no room for the order effect. Consequently, the order effect is generally regarded as a manifest of human biases and an indication of human irrationality. It is the goal of this dissertation research to study how the order effect occurs and seek the rational foundations of the order effect.

Human belief revision has been extensively studied in the fields of Philosophy, Artificial Intelligence, Psychology, and Cognitive Science. A large body of previous research has shown that human belief revision has a coherence foundation, a multi-component structure, and an adaptive characteristic.

The experiment presented in Chapter 3 was designed to study how the different levels of experience (confidence) affect the belief revision patterns. The results show that 1) participants could gradually tune themselves to capture the fundamental base rate information; 2) though order effects exist at the early stage of the training process, they

tend to decrease and disappear as the training continues to the later stage; and 3) participants tend to become less sensitive to new evidence as confidence increases.

UECHO, a proposed computational model of human belief revision, is described in Chapter 4. UECHO takes the coherence foundation of human belief as one of its foundations. In addition, UECHO is equipped with mechanisms to dynamically revise beliefs, learn from experience, and distinguish between confidence and probability. The UECHO modeling results match the experimental results remarkably well.

In conclusion, both the experimental results and the UECHO model suggest that confidence play an important role in human belief revision. Order effects in human belief revision occur where confidence is low and disappear when confidence increases. UECHO provides a computational model of human belief revision and order effects.

Dedicated to my family

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my adviser, Professor Jiajie Zhang, for his intellectual support and encouragement which made this dissertation possible, and for his patience in correcting both my writing and scientific errors.

I wish to thank my co-adviser, Professor Richard Jagacinski, for his guidance in experimental design and carefully reading and polishing the dissertation drafts.

I wish to thank my committee member, Professor Neal Johnson, for his encouragement and insightful comments.

Special thanks go to Professor Todd Johnson. I have been closely working with him for more than 3 years. His intellectual support, helpful discussion, and detailed comments permeated every aspect of this dissertation.

This research was supported by a grant from the Office of Naval Research and a summer fellowship from the Center for Cognitive Science at The Ohio State University.

A preliminary version of ECHO in C, on which all the UECHO codes are based, was provided available by Paul Thagard.

VITA

December 14, 1968Born - Shaanxi, China

1991.....B.S., Psychology,
Peking University

1991 - 1993Research Assistant,
Department of Psychology,
Peking University

1995.....M.A., Psychology,
The Ohio State University

1998.....M.S., Computer and Information Science,
The Ohio State University

1993 - presentResearch & Teaching Associate,
Department of Psychology,
The Ohio State University

ACADEMIC HONORS

Cognitive Science Society
David Marr Award (1998)

PUBLICATIONS

Wang, H., & Zhang, J. (1996). Multiple-level analysis of memory dissociations. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

- Wang, H., Johnson, T. R., & Zhang, J. (1998). UEcho: A model of uncertainty management in human abductive reasoning. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Johnson, T. R., Zhang, J., & Wang, H. (1994). Bottom-Up recognition learning: A compilation-based model of limited-lookahead learning. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Johnson, T. R., Zhang, J., & Wang, H. (1997). A hybrid learning model of abductive reasoning. In R. Sun & F. Alexandre (Eds.), *Connectionist symbolic integration*. Hillsdale, NJ: Erlbaum.
- Johnson, T. R., Wang, H., & Zhang, J. (1998). Modeling speed-up and transfer of declarative and procedural knowledge. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.
- Zhang, J., Johnson, T. R., & Wang, H. (1996). Order effects and frequency learning in belief updating. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Zhang, J., Johnson, T.R., & Wang, H. (1997). The relation between order effects and frequency learning in tactical decision making. *Thinking and Reasoning*, 4(2), 123-145.
- Zhang, J., Johnson, T. R., & Wang, H. (1998). Isomorphic representations lead to the discovery of different forms of a common strategy with different degrees of generality. . In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.
- Zhang, J., & Wang, H. (1995). The integration of internal and external information in numerical tasks. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Zhu, Y., Wang, H., Fan, J., & Zhou, A. (1991). Levels of processing, declarative retrieval strategies and implicit memory. *Acta Psychologica Sinica*, 23(3), 264-271.

FIELDS OF STUDY

Major Field: Cognitive/Experimental Psychology

Minor Field: Quantitative Psychology

TABLE OF CONTENTS

	<u>Page</u>
Dedication.....	iv
Acknowledgments.....	v
Vita.....	vi
List of Tables	x
List of Figures	xii
 Chapters:	
1. Introduction.....	1
2. Literature Review.....	6
2.1. Philosophy of Human Belief.....	7
2.1.1. What is Belief?.....	8
2.1.2. Belief and Knowledge.....	12
2.1.3. Belief Justification	14
2.1.4. Belief Revision.....	16
2.1.5. Summary	20
2.2. Quantitative Belief Management	21
2.2.1. Uncertainty.....	23
2.2.1.1. Imprecision and Uncertainty.....	23
2.2.1.2. External and Internal Uncertainty	26
2.2.1.3. Formal Approaches to Uncertainty	27
2.2.2. Probability Theory	29
2.2.2.1. Frequentism vs Bayesianism.....	30
2.2.2.2. The Ignorance Problem.....	33
2.2.2.3. Bayes' Theorem for Belief Revision	36
2.2.3. The Theory Belief Functions	39
2.2.3.1. Simultaneous Belief and Ignorance	42
2.2.3.2. Evidence Combination.....	44
2.2.4. Graphic Belief Modeling	47

2.2.5. Summary	54
2.3. Order Effects in Human Belief Revision	55
2.3.1. Heuristics and Biases	57
2.3.2. Bounded Human Rationality	62
2.3.3. The Order Effect	67
2.3.3.1. Task Characteristics	71
2.3.3.2. Theories	75
2.3.4. Summary	81
2.4. General Summary	83
3. An Experimental Study	85
3.1. Overviews	87
3.1.1. The Multi-Component Structure	87
3.1.2. The Adaptive Characteristic	90
3.1.3. Summary	91
3.2. Experiment	91
3.2.1. Task	91
3.2.2. Design	95
3.2.3. Participants and Procedure	100
3.3. Experimental Results	101
3.3.1. Training Data	102
3.3.2. Evaluation Data	111
3.4. Discussion	122
4. UECHO: A Computational Model	130
4.1. Overviews	132
4.1.1. The Theory of Explanatory Coherence	133
4.1.2. ECHO	136
4.1.3. Toward UECHO	142
4.1.4. Summary	144
4.2. UECHO	145
4.2.1. Sequential Belief Revision	146
4.2.2. Learning	148
4.2.3. Confidence	155
4.2.4. Summary	163
4.3. Modeling Results	164
4.3.1. Training Data	166
4.3.2. Evaluation Data	172
4.4. Discussion	183

5. General Discussion and Conclusion	188
Appendices:	
A. Experiment Instructions	195
A.1. General Instructions	195
A.2. Instructions before the Initial Evaluation Sub-Block.....	198
A.3. Instructions before the Other Evaluation Sub-Block	198
B. Conditional Independence.....	200
C. Some Details of Those Invalid Participants	201
D. UECHO Parameters	204
D.1. Existing Parameters in ECHO.....	204
D.2. New Parameters in UECHO.....	205
E. A Sample of UECHO Simulation Trace	206
F. The Effect of Training Order	213
References.....	195

LIST OF TABLES

<u>Table</u>	<u>Page</u>
3.1. The trial distribution used in the experiment	97
3.2. Theoretical base rates and conditional probabilities in the current design. $P(\text{Friendly})$ indicates the base rate. $P(\text{Friendly} +)$ is the conditional probability that the target is friendly given any one of positive evidence. $P(\text{Friendly} ++)$ is the conditional probability that the target is friendly given that both evidence items are positive. The base rates and conditional probabilities for hostile identity are not shown but can be easily calculated (i.e., the corresponding complementary of each cell).	98
3.3. The number of participants in each experiment condition. (The numbers in parentheses represent the number of valid participants, see Section 3.3 for details)	102
C.1 The number of participants excluded from the data analysis due to various reasons	202

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1.1. A schematic diagram of order effects. E1-E2 and E2-E1 are sequences of the same evidence items (E1 and E2) with different presentation orders. Note that the final judgments are different. While the top panel represents the recency effect (i.e., the final judgments are significantly affected by the final piece of evidence), the lower panel represents the primacy effect (i.e., the final judgments are significantly affected by the early piece of evidence).....	4
2.1. A simple Bayesian belief network to solve the Monty Hall puzzle.....	52
2.2. Evidence for the independence of verbal memory and belief revision in impression formation. Recall curve shows strong recency over the last six serial positions. Judgment curve shows primacy. (Reproduced from N. H. Anderson, 1991, page 10).....	73
3.1. The version of CIC used as the task domain.....	93
3.2. The experimental design.	99
3.3. Initial belief judgments (belief0) across the 144 training trials. The corresponding locally-weighted regression (loess) curves are also plotted. Based on the experimental data in all three base rate conditions. The numbers in parentheses indicate the corresponding theoretical base rates of friendly targets in terms of percentage.....	104
3.4. Initial confidence judgments (confidence0) across the 144 training trials. The corresponding locally-weighted regression (loess) curves are also plotted. Based on the experimental data in all three base rate conditions.....	106
3.5. Proportion of friendly decisions across the four training blocks. Based on the experimental data in all three base rate conditions.	107

3.6.	Proportion of friendly decisions across the four training blocks, separated for different evidence sequences. The dashed lines indicate the theoretical probabilities (see Table 3.2). Based on the experimental data in all three base rate conditions.	110
3.7.	The belief revision pattern based on the experimental data from the initial evaluation block (i.e., block0). “init”, “e1”, and “e2” indicate the time points before any evidence, after the first evidence, and after the second evidence, respectively. “+” denotes a positive evidence, and “-“ denotes a negative evidence.	112
3.8.	The belief revision patterns in the 1:1 base rate condition. Based on the experimental data from the four evaluation blocks (block1 to block4).	115
3.9.	The area estimates across the four training blocks. Based on the experimental data from the four evaluation blocks (block1 to block4).	116
3.10.	The belief revision patterns in the 3:1 base rate condition. Based on the experimental data from the four evaluation blocks (block1 to block4).	117
3.11.	The belief revision patterns in the 1:3 base rate condition. Based on the experimental data from the four evaluation blocks (block1 to block4).	119
3.12.	The belief revision patterns in all three base rate conditions. Based on the experimental data from all five evaluation blocks (block0 to block4). Only data from inconsistent evidence sequences (i.e., +- and +-) are showed.	120
3.13.	From block1 to block4, belief judgments approach the built-in theoretical values. Based on the experimental data from the first and the fourth evaluation blocks (block1 and block4) in all three base rate conditions.	123
3.14.	The confidence revision patterns in all three base rate conditions. Based on the experimental data from the four evaluation blocks (block1 to block4).	124
4.1.	A graphical representation of learning in UECHO. Panel A represents the original network, in which dashed lines indicate inhibitory relations and the strengths are determined by the	

parameter <i>inhibitory weight</i> , and solid lines indicate excitatory relations and the strengths are determined by the parameter <i>excitatory weight</i> . Panel B represents the network after learning. Note that the connection strengths have been tuned to reflect the corresponding associations based on experiences.	149
4.2. The UECHO network for the CIC task. Inhibitory connections are represented by dashed lines and excitatory connections are represented by solid lines. In this specific example, it assumes that R+ and ID+ are observed and thus they are connected to the SEU. The numbers in parentheses are the corresponding activation values after the network settles. It is clear that the hypothesis of friendliness (with an activation of 0.64) beats the hypothesis of hostility (with an activation of -0.62).	152
4.3. The logistic mapping rule between activation values and probabilities, with τ being -3.2.	154
4.4. The flowchart of associative learning in UECHO. Panel A represents the initial UECHO network with R+ and ID+ being true. Note that the SEU is not shown. Panel B represents the settled network. Note that the friendly node has a higher activation than the hostile node, represented by the degrees of grayness. Panel C indicates that a feedback of the true identity is provided (in this case it is friendly). Learning via the Rescorla-Wagner rule is shown in Panel D.	156
4.5. The parameter of skepticism, θ , as a function of the <i>experience</i> number. As θ decreases, confidence increases. Based on equation 4.11, and $\theta_0=0.1$, $a=1.5$, $b=-0.032$, and $c=0$	159
4.6. Initial belief judgments (belief0) across the 144 training trials. The corresponding locally-weighted regression (loess) curves are also plotted. Based on the simulation data in all three base rate conditions. The numbers in parentheses indicate the corresponding theoretical base rates of friendly targets in terms of percentage.	168
4.7. Proportion of friendly decisions across the four training blocks. Based on the simulation data in all three base rate conditions.	170
4.8. Proportion of friendly decisions across the four training blocks, separated for different evidence sequences. The dashed lines indicate the theoretical probabilities (see Table 3.2). Based on the simulation data in all three base rate conditions.	173

4.9.	The belief revision pattern based on the simulation data from the initial evaluation block (i.e., block0). “init”, “e1”, and “e2” indicate the time points before any evidence, after the first evidence, and after the second evidence, respectively. “+” denotes a positive evidence, and “-“ denotes a negative evidence.	174
4.10.	The belief revision patterns in the 1:1 base rate condition. Based on the simulation data from the four evaluation blocks (block1 to block4).	176
4.11.	The area estimates across the four training blocks. Based on the simulation data from the four evaluation blocks (block1 to block4).	177
4.12.	The belief revision patterns in the 3:1 base rate condition. Based on the simulation data from the four evaluation blocks (block1 to block4).	179
4.13.	The belief revision patterns in the 1:3 base rate condition. Based on the simulation data from the four evaluation blocks (block1 to block4).	181
4.14.	The belief revision patterns in all three base rate conditions. Based on the simulation data from all five evaluation blocks (block0 to block4). Only data from inconsistent evidence sequences (i.e., -+ and +-) are showed.	182
C.1.	Examples of invalid participants who used tricks. The initial belief judgments are plotted against the 144 training trials.	203
F.1.	The belief revision patterns in all three base rate conditions. Based on the experimental data when the training order is consistent with the evaluation order.	214
F.2.	The belief revision patterns in all three base rate conditions. Based on the experimental data when the training order is inconsistent with the evaluation order.	215
F.3.	The belief revision patterns in all three base rate conditions. Based on the simulation data when the training order is consistent with the evaluation order.	216
F.4.	The belief revision patterns in all three base rate conditions. Based on the simulation data when the training order is inconsistent with the evaluation order.	217

CHAPTER 1

INTRODUCTION

Uncertainty is a specter that haunts human beings. On the one hand, the environment around us is full of uncertainties, ranging from simple situations like tossing a coin and playing a gamble game, to much more complex ones like natural weather patterns, global stock market, and world peace. On the other hand, uncertainty inevitably arises in one's mental processes as well. At the lowest level, biological processes are rarely clear-cut and without noise. At the perceptual level, one may have difficulty recognizing an object given noisy sensory input. At the cognitive level, uncertainty results from inadequate information sources, limited information processing capacities, or ambiguities in natural language. People, as John Locke put it, are most of the time condemned to live in the twilight of uncertainty rather than the midday sun of certainty, for "our ignorance [is] infinitely larger than our knowledge" (see Wolterstorff, 1996, page 60).

While pursuing *absolute knowledge* is the ultimate goal of human life, uncertainty is a pervasive fact with which life must be prepared to contend. It has long been believed, through epistemological inquiries, that forming and maintaining various *beliefs* (e.g.,

scientific theories and religions) is one way, if not the only way, that human beings try to overcome uncertainty and embrace the ultimate goal of human life.

Encyclopaedia Britannica gives the following definition of belief. Belief is “a mental attitude of acceptance or assent toward a proposition without the full intellectual knowledge [and evidence] required to guarantee its truth. ... Beliefs have been distinguished according to their degree of certainty: a surmise or suspicion, an opinion, or a conviction.” (Britannica Online, 1998).

Then how do various beliefs help people? First of all, belief is often associated with knowledge in the sense that knowledge is nothing more than belief plus complete conviction and justification. Second, when knowledge is not available, belief provides a direction for further investigation and a basis for immediate decision making (Schmitt, 1992). As long as the underlying belief closely conforms to the reality, one hopes that the direction it points to is not misleading and the decision making it supports is not fallacious.

It is generally agreed that one constantly conducts *belief revision* – a process in which one revises one’s beliefs in the light of new information, with a goal to maintain a reasonably consistent and up-to-date belief system. The ultimate reason for belief revision is that belief, unlike knowledge, can be wrong. Since the environment changes dynamically, as time passes new pieces of relevant information constantly become available. These new pieces of information may suggest that one’s current beliefs are incomplete, inappropriate, incorrect, inconsistent, or out-of-date. How to sufficiently and efficiently incorporate the new information into the current belief system to reflect the changed reality is a big challenge one faces. It is of great philosophical and

psychological interest to investigate whether one is able to achieve such a goal and what the underlying regularities are.

Psychological investigations of human belief revision have revealed an important finding -- *order effects* (e.g., Asch, 1946; Hogarth & Einhorn, 1992; Schlottmann & Anderson, 1995; Zhang, Johnson, & Wang, 1997). Generally speaking, the order effect refers to the phenomenon that the temporal order in which information is presented affects the final judgment of an event (see Figure 1.1). Undoubtedly, the temporal order of incoming evidence often carries some important information about the true meaning of an event. However, robust order effects have been found even in situations where the temporal order of incoming evidence seems not meaningful. It is these cases that make the order effect a very interesting phenomenon. Given that the temporal order is not informative, why does one still show order effects? What are the underlying mechanisms? What do these effects suggest about human belief revision in general?

These questions are what this dissertation aims to answer. The primary contribution of this dissertation is the development of a computational model of human belief revision that naturally accounts for order effects. The key conclusion is that confidence plays an important role in human belief revision, and when little is known about the uncertainty of the environment, order effects result from one's coherently and dynamically adapted expectations of the statistical properties of the environment.

Figure 1.1. A schematic diagram of order effects. E1-E2 and E2-E1 are sequences of the same evidence items (E1 and E2) with different presentation orders. Note that the final judgments are different. While the top panel represents the recency effect (i.e., the final judgments are significantly affected by the final piece of evidence), the lower panel represents the primacy effect (i.e., the final judgments are significantly affected by the early piece of evidence).

This dissertation is organized as follows. In chapter 2, I present a literature review of belief, belief revision, and order effects. Some formal approaches to uncertainty management and belief revision are described. Important previous empirical studies of order effects are reviewed. Order effects in a large context of human judgment under uncertainty are discussed.

In chapter 3 I present an empirical study that investigates how the patterns of order effects change with more and more experience about the environment being obtained. The hypothesis tested here is that as one gains more and more experience about the environment, one's reaction to new information would be finely tuned thus more likely to reflect the true statistical structure of the environment. That is, one would show more realistic adaptation rather than over-reaction or under-reaction, which usually results in order effects. It is shown that belief has a multi-dimensional structure. Besides the ordinary degree of belief dimension, the confidence dimension measures the amount of previous experience about the environment. Different levels of confidence result in different belief revision patterns.

In chapter 4 I describe how a computational model called UECHO is developed to model the experimental results. First, ECHO, a connectionist model of human abductive reasoning, is reviewed. Then I show why UECHO is a sophisticated extension of ECHO, with human belief revision and order effects in mind. UECHO clearly reveals the different patterns of belief revision at different confidence levels. The model performance is compared with the experimental results.

The final chapter summarizes the major results of this dissertation research and outlines some directions for future work.

CHAPTER 2

LITERATURE REVIEW

Belief and belief revision are large and complicated subjects. Philosophers are interested in them. Due to its close relationship with knowledge, belief has become a central topic in epistemology – the study of the nature and origin of knowledge. Researchers in Artificial Intelligence (AI) are also interested in them. In order for the artificial systems that AI researchers build to be really “intelligent”, it is essential that these systems be able to formally reason and adapt with uncertainty and belief. Psychologists and cognitive scientists, of course, are interested in them as well. Belief and belief revision are ubiquitous in both scientific discovery and everyday life. It is important to understand how people manage them.

It should be noted that these different perspectives are closely related. Philosophical inquiries help to clarify fundamental questions like what belief is and how belief is justified. Formal AI approaches to belief management identify *normative* and mathematical methods to represent and revise beliefs. Psychological studies empirically illustrate important underlying principles and constraints in *human* belief revision. It has

been of great interest to compare observed human behaviors with various normative predictions. Clearly, ideas and methods from different perspectives need to be combined to reveal the full spectrum of the subject.

In this chapter, I review some related topics from the above perspectives. In section 2.1, I present some philosophical investigations about belief and belief revision. In section 2.2, I describe some formal probabilistic approaches to uncertainty and belief management. Section 2.3 reviews the literature of human judgment under uncertainty, with a focus on order effects in human belief revision. The final section in this chapter summarizes the reviews.

2.1. Philosophy of Human Belief

Belief is a central topic in epistemology. A philosophical analysis of belief focuses on the necessary and sufficient conditions for something being a belief and typically consists of three sub-queries. First of all, the exact meaning of belief is inquired. That is, when one says that “I believe that ...”, what does he/she really mean? Second, the relationship between belief and knowledge is examined. In other words, what’s the difference between “I believe that ...” and “I know that ...”? Finally, it is of importance to inquire how belief is justified and revised. That is, why does one believe? And why and how does one change one’s belief?

2.1.1. What is Belief?

One's repertoire of beliefs changes in nearly every waking moment. The sound of the wind or the sight of a maple leaf, adds a belief to one's store. These are trivial beliefs, quickly acquired and quickly dropped. Other beliefs endure: the belief that the Sun rises in the morning, and the belief that everybody will eventually die. Belief, in fact, is one of the most commonplace and familiar things in the world. And it is just these commonplace and familiar things that prompt the most interesting philosophical questions. "What is belief?" is one such question.

There exist two different ways of analyzing beliefs (see Price, 1969; Quine & Ullian, 1978; Schiller, 1979). In the traditional approach, it is assumed that belief is a sort of mental occurrence or mental act. This mental occurrence is introspectible and can be described in a clarified language. Therefore, the main task of a philosophy of belief is to examine this introspectible mental occurrence, and to distinguish it from other mental occurrences, such as the one that occurs when one entertains a proposition neutrally without accepting or assenting it. Clearly, the definition of belief cited previously from *Encyclopaedia Britannica* is consistent with this approach.

The modern approach to belief is quite different. Belief is not regarded as an introspectible mental occurrence, but as a disposition (Quine & Ullian, 1978). In this view, although a belief may manifest itself in various sorts of mental occurrences, these occurrences themselves are not the belief. A belief reveals itself in consequential behaviors. Therefore, a belief is not something that happens at a particular moment, but something that one has or possesses throughout a period. It makes perfect sense to

continue to believe a proposition for a long time even when one does not think (or talk) about it at all or when one is asleep.

It is critical to note that this dispositional approach does not deny that there are mental occurrences. However, it does deny these mental occurrences are beliefs. A disposition manifests itself from time to time by actual mental occurrences, but a disposition is something that one still has or possesses at times when it is not actually mentally manifested at all. In other words, a belief is something dispositional and behavioristic, something which one possesses for a period, no matter whether it occurs mentally or not.

Standing between the two approaches is an interesting question: whether animals have beliefs. Philosophers have generally agreed that an answer to this question involves a conceptual understanding about the nature of belief (Ackermann, 1972). From the traditional point of view, which often claims that belief is conceptually related to linguistic competence, it is best to deny that animals have beliefs. Although animals are often goal-oriented and may behave as if they had beliefs, the mental occurrences, if they exist, cannot be examined by formulating sentences. Therefore, animals are best conceived as sophisticated organisms that behave without beliefs. The dispositional and behavioristic approach, on the contrary, argues that animals do have beliefs. Conscious awareness with linguistic competence is a reflection of sophistication in design, but is neither a necessary condition for belief nor a boundary between human beings and animals. Beliefs are revealed in actions and may not be accompanied by any conscious formulation of the belief.

It is important to note that conscious awareness has been a critical factor that distinguishes the two approaches. While the traditional approach emphasizes conscious beliefs, the dispositional approach emphasizes behavioral beliefs. Behavioral beliefs are often called non-conscious beliefs to indicate that they may not be accompanied by conscious awareness. In fact, some philosophers argue that there exist a third of kind of belief – unconscious belief (see Ackermann, 1972). Unconscious beliefs are long-standing beliefs that can influence behaviors over a long period of time and resist conscious recognition upon query or examination. Different from non-conscious behavioral beliefs, which are assumed to be easily recognizable at the conscious level in response to a query or scrutiny, unconscious beliefs exist at the sub-symbolic level or in the psychoanalytic sense, and they do not identify themselves easily to a query or scrutiny.

The notion of unconscious belief has obtained psychological support. For example, Kahneman and Tversky (1982) distinguished conscious active anticipations and unconscious passive expectations in human perception. They noticed that unconscious expectations and conscious anticipations could conflict. More importantly, they showed evidence that an observer could be unconsciously “primed” for one event while consciously expecting another, and could show physiological evidence of surprise at the occurrence of an event that was consciously expected. Therefore, it seems that an individual can have conflicting beliefs for the same event at the same time, though one is conscious and the other is unconscious. These observations support the image of the

mind as a multi-level structure in which different levels have access to different data and hold different views or beliefs of the situation.

The authors further provided an explanation for the above phenomenon. They argue that these unconscious expectancies or beliefs have been developed over a life-time of visual experience and thus cannot be easily revised on demand for a particular occasion. These unconscious perceptual beliefs have a profound effect on perception and determine what one “sees” in an ambiguous stimulus.

The notion of unconscious beliefs clouds the possibility of analyzing beliefs through introspection and self-scrutiny. It also precludes any attempts to enumerate one’s actual beliefs. As a result, most of the philosophical work on belief focuses on consciously recognizable beliefs.

In addition, it is important to note that these consciously recognizable beliefs are usually expressed as propositions in the form of “I believe that ...”. Although degrees of belief have been generally admitted among philosophers, similar beliefs with different degrees are often expressed as separate propositions, thus counted as different beliefs. For example, *disbelief* (believing a proposition is false) and *nonbelief* (not believing it true) have been seriously distinguished (e.g., Quine & Ullian, 1978). Nonbelief can be further divided into *half-belief*, *quarter-belief*, and many other beliefs (e.g., Schiller, 1979). These beliefs are regarded as qualitatively different ones since it is very unlikely that a single person can hold all of them simultaneously. In this sense, a belief is either held or not held, and there is no uncertainty about it. As I will show, this treatment of

belief (which may be called propositional belief) is significantly different from the treatment given from the probabilistic perspective.

2.1.2. Belief and Knowledge

It has been generally agreed among philosophers that there are two types of knowledge. Depending upon the different types of things that are known, one can distinguish between *knowledge of facts* and *knowledge by acquaintance*. Knowledge of facts is related to facts or truths and is usually expressed in the form of the verb “to know” followed by a dependent clause. Examples of knowledge of facts include:

I know that $7 \times 7 = 49$.

We all know that Beijing is the capital of China.

Knowledge of acquaintance often describes personal experience about an entity or a group of entities. In knowledge of acquaintance the verb “to know” is not followed by any dependent clause. Instead, it is followed by a noun or a noun-phrase. Some examples are:

He knows Mr. Smith.

I know Buckeye Village.

Belief is often contrasted with knowledge in the sense that belief is an inferior substitute for knowledge (see Ackermann, 1972; Price, 1969). However, it is very important to note that although it is appropriate to contrast belief with knowledge of facts, there is no contrast between belief and knowledge by acquaintance. Through

sentences like “I believe him” and “I know him”, one is talking about different things. Obviously, believing a person is not an inferior substitute for knowing a person.

The contrast between belief and knowledge is most direct when one compares belief and knowledge of facts (i.e., “believe that” and “know that”). Only in this case, belief serves as a second-best substitute for knowledge, as shown in the following example:

I do not know if the Yellow River is in China, but I believe it is.

The contrast can be illustrated more specifically in the following manner. If one knows a proposition, then the proposition is true. For example, if one knows that it is raining, then it is raining. There is not doubt about it. It would be a contradiction to say something like

I know that it is raining, but perhaps it isn't.

Unfortunately, believing a proposition does not entail that the proposition is true. A proposition can be either true or false, no matter whether it is believed by someone or not. Therefore, there would be no contradiction to say that

I believe that it is raining, but perhaps it isn't.

The contrast between belief and knowledge, however, does not mask the close connection between the two. As suggested in the definition from *Encyclopaedia Britannica*, belief admits of degree, ranging from barely accepting at the one end, to complete conviction at the other. Knowledge is simply the name given to the belief with the highest degree of believability. That is, when one believes a proposition with complete conviction and based on conclusive evidence, one knows it. In other words,

knowledge is nothing more than belief plus complete conviction and conclusive justification.

2.1.3. Belief Justification

As stated previously, while knowledge is the goal of inquiries, belief is usually what one can get. A reasonable belief is based on good but not conclusive evidence. Given the fact that the unconvinced belief does not guarantee the truth of what is believed, how is the unconvinced belief justified? That is, on what ground is such a belief held without being dismissed?

There are two main philosophical views regarding how belief is justified: the *foundations* approach and the *coherence* approach (see Harman, 1973; Gardenfors, 1990; Doyle, 1992). The foundations approach argues that a rational individual derives beliefs from reasons for these beliefs. In other words, an individual holds a belief if and only if it possesses a satisfactory reason for that belief. More specifically, according to the foundations approach, a belief is justified if and only if 1) the belief is self-evident, or, 2) the belief can be derived from a set of other justified beliefs. Beliefs fulfilling the first condition are called *foundational* beliefs. These beliefs do not rest on other beliefs. It is assumed that the acceptance of these beliefs is indisputable, since they usually correspond to "hard facts", such as observations about the physical world. Beliefs satisfying the second condition are those which are logical consequences of the foundational beliefs. Every belief must have one or more justifications, and the justification chain must end in the foundational beliefs. No *circularity* is allowed.

The coherence approach, in contrast, maintains that pedigrees do not matter for belief justification. Instead, an individual holds a belief as long as it logically coheres with the individual's other beliefs. In comparison with the foundations approach, in the coherence approach beliefs can be held independently of how they may have been inferred. Moreover, since beliefs cohere and mutually justify each other, no belief is more fundamental than another, and cycles are allowed in the justification chains.

Sosa (1980, page 23) describes the differences between the two approaches in the following way:

For the foundationalist every piece of knowledge stands at the apex of a pyramid that rests on stable and secure foundations whose stability and security does not derive from the upper stories or sections. For the coherentist a body of knowledge is a free-floating raft every plank of which helps directly or indirectly to keep all the others in place, and no plank of which would retain its status with no help from the others.

It has been argued that the coherence approach is the only rational non-foundational approach of belief justification (Cornman, 1978). Several different types of coherence have been suggested by making the notion of "coherence" precise (see Thagard, 1989). For example, one typical type of coherence is called *deductive coherence*, which requires that the belief set be logically consistent. Other types of coherence are based on additional relations between beliefs and thus require information beyond pure logic to measure the coherence of a belief set. For instance, *probabilistic coherence* depends on the beliefs having probability assignments. *Semantic coherence* is based on beliefs that have similar meanings. And *explanatory coherence* occurs when there are consistent explanatory relationships between the propositions in the belief set. More details about explanatory coherence will be discussed in chapter 4.

The coherence approach is generally preferred to the foundational approach (see Gardenfors, 1990; Thagard, 1989). It has been argued that the foundational approach involves excessive computational cost. It is intellectually very costly to keep track of the reasons of beliefs, and the benefits are usually out-weighed by the costs. It would be more intellectually parsimonious to neglect the pedigree of one's beliefs. Moreover, it has been shown that the foundational approach conflicts with observed human behavior. On the one hand, it is very common that one sometimes drops a belief for the purpose of maintaining consistency among beliefs, even if that belief has well-established evidential support. On the other hand, the *belief preservation effect* (Ross & Lepper, 1980; Ross & Anderson, 1982) suggests that people are reluctant to give up some beliefs even when the original evidential bases of these beliefs are completely destroyed.

In addition, it has also been suggested that the difference between the two approaches have been overemphasized. For example, it has been argued that the foundations approach represents the most direct way of mechanizing the coherence approach (Doyle, 1992).

2.1.4. Belief Revision

Belief revision is the process of incorporating new information into the current belief system. As stated previously, the ultimate reason for belief revision is that the constant changes in the environment always make one's belief system out of date or inappropriate. It is generally agreed that for belief revision to be rational or reasonable, some minimal requirements must be met. First of all, consistency should be maintained in the resulting

belief system. The requirement of consistency is fundamental since inconsistent beliefs describe no world and thus are useless. Second, since purely logical considerations are not sufficient to guide belief revision, some extra-logical means are needed. This is easily illustrated by two extreme examples. One can completely give up old beliefs to adopt new ones, or one can maintain old beliefs and completely ignore new ones. In both cases consistency is preserved, and there is no pure logical reason for preferring one to the other. One of the well-received guidelines for belief revision is that reasonable belief revision should be as conservative as possible and represent the minimal change of the current belief system to accommodate the new information consistently (see Gardenfors, 1992).

It should be obvious that the approaches to belief justification have very direct implications for how beliefs should be revised in the light of new information. According to the foundations approach, one should give up a belief whenever it is no longer adequately justified. Therefore, belief revision should consist of two steps. The first step is to give up all beliefs that no longer have a satisfactory reason. And the second step is to add new beliefs that have become well justified. In contrast, the coherence approach emphasizes consistency and conservatism and claims that a belief revision should keep the resulting belief system consistent and should only make a minimal change to the current belief system. Whether a belief is dropped or maintained during belief revision is independent of how it is inferred. As long as the coherence of the resulting state is preserved or improved, no belief needs to be dropped. In other words, one should retain as many of one's beliefs as possible while accommodating any

new evidence. The rationale behind coherent belief revision is that old beliefs contain valuable information, and losses of information should be minimal.

Examples of adopting the foundations approach to belief revision include the *Truth Maintenance System* developed by Doyle (1979) and the *Theory of Endorsement* proposed by Cohen (1985). Both systems keep track of the reasons (or endorsements) for beliefs and maintain that these reasons (or endorsements) be the sole criteria for belief revision.

Recently there has been an increasing emphasis on the coherence approach to belief revision. The so-called AGM theory of belief revision (Alchourron, Gardenfors and Makinson, 1985; Gardenfors, 1990) is one of the most well known examples of the coherence approach.

The AGM theory deals with propositional beliefs that are connected by propositional logic. That is, a belief is either accepted or rejected (i.e., no degrees of belief are considered), and a belief set contains a set of beliefs and all their derived logical consequences. The AGM theory distinguishes three kinds of operations for belief revision:

- (1) *Expansion*. In expansion, a new belief (and its logical consequences) is added to the current belief set (e.g., through learning). None of the current beliefs are dropped even if the resulting belief set is inconsistent.
- (2) *Revision*. In revision, a new belief (and its logical consequences) is added to the current belief set in the light of new information. In order to keep consistency in the resulting belief set, some old beliefs are dropped.

(3) *Contraction*. In contraction, some old beliefs (and some of their necessary logical consequences) are retracted without new beliefs being added.

While belief expansion is trivial to do, belief revision and contraction represent real challenges. The reason is that there are always multiple ways of doing them, but there is no logical way to make a choice. The AGM theory meets the challenges by proposing various *rationality postulates* that these operations must follow. By following these rationality postulates, the solution for belief revision and contraction would be unique. As a result, one can have a *revision function* and a *contraction function*.

The guiding principle underlying these rationality postulates is the maintenance of coherence with minimal changes. However, it turns out that it is difficult to define the “minimal change”. The notion of *epistemic entrenchment* is proposed to specify the measure. The idea behind the notion is that even if all beliefs in the belief set are equally accepted, their relative importance to problem solving and decision making is not equal. Some are more important than others. Epistemic entrenchment is a measure of such importance. Those beliefs with higher degrees of importance have larger entrenchment, and those beliefs with lower degrees of importance have smaller entrenchment. It is this epistemic entrenchment measure that determines which old belief to give up when belief revision and contraction are carried out. More specifically, when giving-up is needed, those beliefs with smallest entrenchment are dropped.

Unfortunately, quantitatively specifying epistemic entrenchment in general is difficult. Therefore, in AGM theory, epistemic entrenchment is only qualitatively specified. Recently, several models of the AGM theory have been developed, in which

application-specific epistemic entrenchment measurements have been quantitatively implemented (see Dixon, 1994). By choosing different epistemic entrenchment measures, it has also been shown that the AGM theory is very flexible in the sense that it is able to reason based on either the foundational or the coherence view of belief justification (Gärdenfors, 1990).

2.1.5. Summary

Belief, as a second-best substitute of knowledge, represents a sort of disposition, which in most cases guides human behaviors. Although beliefs often manifest themselves in the form of mental occurrences, not all beliefs are consciously introspectible. The notion of unconscious belief makes the enumeration of one's actual beliefs impossible.

Intuitively, beliefs held by an individual should be based on solid foundational evidence. However, the coherence approach to belief justification suggests that this may not be necessary. As long as one's beliefs are consistent and coherent with each other, they are justified.

Belief revision based on the foundations approach is straightforward. Those new beliefs with solid reasons are added, and those old beliefs that lose support are dropped. However, it is shown that this approach does not meet the requirement of conservatism and the principle of informational economy. Belief revision based on the coherence approach is to make minimal change of the current belief system while preserving

coherence. The AGM theory of belief revision adopts the concept of epistemic entrenchment to qualitatively define “minimal change”.

Although these philosophical analyses have offered valuable insights on belief and belief revision, they also raise additional questions.

First, the above philosophical treatments of belief and belief revision are only concerned with propositional beliefs. That is, a belief is either accepted or rejected, and there is no compromise between the two. These are different from the treatments of belief from the probabilistic perspective. In this view, a belief is subject to a continuous uncertainty judgment. A belief can be held completely, or it can be held only barely. The question is, are they different beliefs, or the same belief but with different levels of uncertainty? It seems that the latter conforms to intuition more naturally. In section 2.2, I will review studies from that perspective.

In addition, rational belief is assumed in these inquiries. However, it is always questionable whether the proposed rationality postulates in belief revision are supported by observed human behaviors. Hot debates exist on whether human behavior is rational. This topic will be discussed in section 2.3.

2.2. Quantitative Belief Management

The issues of belief and belief revision have also been extensively addressed in other disciplines. One of them is the research in UAI (Uncertainty in Artificial Intelligence, or Reasoning under Uncertainty). Since uncertainty is rarely avoidable when people interact with the environment, those AI researchers who are interested in building machines to

deal with real-world problems find themselves in a position where adequate and efficient belief management is essential. Instead of analyzing belief and belief revision qualitatively and phenomenologically as done by philosophers, AI researchers use various sophisticated mathematical and statistical methods to study the subject at a quantitative level. The basic weapon in their arsenal is various probability-related techniques.

The concept of uncertainty has become fundamental, as it should. Uncertainty, as the ultimate source of subjective belief, has been defined and measured quantitatively. Belief, generally expressed as the degree of believability of a proposition, is functionally dependent on the amount of uncertainty that the proposition involves.

Some researchers, especially those non-Bayesians, in this area prefer the term *uncertainty management* to the term *belief management*, due to the subjective semantics of the word “belief”. It should be clear that uncertainty and belief are indeed two complementary aspects of a single entity in the sense that belief is human’s *spear* to handle the *shield* of uncertainty. They are used interchangeably in this dissertation.

In this section, I review several influential formal techniques for belief management. I first present more detailed descriptions and analyses of uncertainty. Then I describe the two most important approaches to formal belief management: the theory of probability and the theory of belief functions. Finally, I describe some work on Bayesian belief networks and graphical belief modeling.

2.2.1. Uncertainty

When dealing with real-world problems, uncertainty is usually inevitable. For some people uncertainty is a plague that should be avoided by all means, but for some others uncertainty is a blessing. As soon as the mysterious veil is uncovered, the beauty of regularity and predictability will come out.

2.2.1.1. Imprecision and Uncertainty

It is well agreed that there are two general types of uncertainty (see Walley, 1991). On the one hand, due to various information deficiencies, one may not completely determine the truth of a proposition. However, the average proportion of that proposition being true in a long run can be precisely specified. The indeterminacy involved in this case is called *uncertainty*¹. Consider the tossing of an ordinary (fair) coin. One is unable to predict which side the coin will land facing upwards. Thus one is uncertain about the outcome. However, if the coin is tossed many times, one knows that approximately 50% of the times it will land with head (or tail) facing upwards.

On the other hand, in some cases, neither can one completely determine the truth of a proposition nor can one precisely specify the average proportion of that proposition being true in a long run. This type of uncertainty – the indeterminacy of the average behavior – is usually called *imprecision*. Consider a guessing game, in which one is to guess which one of the two boxes contains a ball. Each time the ball is put into one box by a magician, and the outcome is completely random and without any predictability even in a

long run. It is obvious that for any single trial, one cannot be certain which box contains the ball. Moreover, one cannot even have a precise evaluation of the long-term average behavior. In this case, if one wants to make a judgment that the ball is in a specific box, one is in a situation that is both imprecise and uncertain. However, if one only wants to say that the ball is either in box 1 or box 2 but not to specify which one, one is in a situation that is imprecise but certain.

The distinction between uncertainty and imprecision has also been illustrated from a different perspective by Dubois and Prade (1988). They note that a proposition can be defined as a quadruple: *<attribute, object, value, confidence>*. *Attribute* is a function that attaches a *value* to an *object*, and *confidence* indicates the degree to which the proposition conforms to the reality. In this context, they argue that imprecision relates to the content of the proposition (i.e., the *value* component of the quadruple), and uncertainty relates to the truth of the proposition (i.e., the *confidence* component of the quadruple). More specifically, a proposition is precise when the information about value is not vague or fuzzy, and a proposition is certain when its conformity to the reality is completely determined. Let us consider an example:

It is likely that John is a young fellow.

This proposition can be easily expressed as the following quadruple:

<age, John, young, likely>

¹ It is a little confusing to call both the general concept and one of its subtypes as uncertainty. However, it will be pretty clear from the context which I am talking about.

It is clear that this proposition is not only imprecise but also uncertain. It is imprecise because the word “young” is vague and fuzzy (i.e., no precise age is specified); and it is uncertain because its truth is not completely known (i.e., it is only “likely” to be true).

Based on this distinction, Dubois and Prade also notice that there is a close connection between imprecision and uncertainty. By making the content (i.e., the value) of the proposition more precise, the uncertainty tends to increase. Similarly, the uncertainty tends to decrease when the content becomes more imprecise. For example, while it is likely that John is young, the belief that John is 20 years old involves more uncertainty.

In fact, as it will become increasingly clear, the distinction between imprecision and uncertainty made by Dubois and Prade (1988) is not as general as it should be. Imprecision is not only applied to the *value* component of a proposition, it is also applied to the *confidence* component. When the uncertainty involved in a proposition cannot be precisely evaluated, imprecision appears. As readers may have already noticed, in the above example (i.e., “*it is likely that John is a young fellow*”) the word “likely” indicates not only uncertain but also imprecise. That is, no precise specification is given for “likely”.

As I will show later, the distinction between imprecision and uncertainty is such a fundamental issue that it causes the “holy war” in uncertainty management. On the one hand, one of the most serious criticisms of the theory of probability is that it is difficult to handle imprecision. On the other hand, fuzzy sets and the possibility theory (see Zadeh, 1965, 1978; Dubois & Prade, 1988; Klir, 1994) often deal with imprecision but not

uncertainty (e.g., Almond, 1995). The theory of belief functions, which I will discuss later, deals with both imprecision and uncertainty and thus captures a more thorough picture of uncertainty management.

2.2.1.2. External and Internal Uncertainty

Another interesting and insightful dichotomy of uncertainty has been given by Kahneman and Tversky (1982). They distinguish external uncertainty and internal uncertainty. External uncertainty comes from the external world, which one cannot control. Examples of external uncertainty include tossing a coin, or predicting the championship of World Cup 98. There are two modes to assess external uncertainty. One mode, the *distributional mode*, is to take an outside view and evaluate the relative frequencies in a reference class. The other mode, the *singular mode*, is to take an inside view and assess the propensities of the particular case at hand. Take the championship of World Cup 98 as an example. If the uncertainty is assessed in a distributional mode, then one would predict that Brazil is more likely to win since it won the most championships in the World Cup history. However, if it is assessed in a singular mode, one may prefer France since it is the host country and it has many superstars.

In contrast, internal uncertainty comes from one's ignorance, as in statements like "I believe Beijing is the capital of China". Here the uncertainty results from one's own mind, but not from the propensity of Beijing to be the Capital of China. As in the case of external uncertainty, the internal uncertainty can also be assessed by two modes. One's guess that Beijing is the capital of China may be based upon serious reasoning (e.g., "I

believe the capital of China is either Beijing or Shanghai, and I think Shanghai is unlikely, therefore I guess it is Beijing”), which is called the *reasoned mode*. Or, the guess may be based upon pure introspection and without any reasoning involved, which is called the *direct mode*.

It should be noted that the distinction between external and internal uncertainty is not static. The flexibility depends on many factors, one of which is timing. The uncertainty about the championship of World Cup 98 is external if it is assessed before July 12, 1998. After that, the uncertainty, if any, comes from the internal ignorance.

One of the most important conclusions those authors make is that even though there are many different types of uncertainty, the well-established classical theory of probability is only suitable to describe the external uncertainty assessed by a distributional mode. Clearly this conclusion is not surprising.

2.2.1.3. Formal Approaches to Uncertainty

Many real-world problems with inherent uncertainty can only be solved approximately. In addition, sometimes even when a complete and exact solution to a problem is available, one may still settle on an approximate or non-optimal solution in order to reduce the computational or cognitive cost. This is true for both human beings and artificial systems.

Various formal methods have been developed to formally manage uncertainty. They can be generally divided into two categories (see Bhatnager & Kanal, 1986). The first category adopts a numeric approach. According to this approach, uncertainty can be

represented by one or more numbers. Examples of this category include the theory of probability, which uses a single dimension of probability (i.e., a number in $[0,1]$) to summarize the uncertainty, and the theory of belief functions (Shafer, 1976), which uses two numbers (i.e., credibility and plausibility) as the summary.

The second category adopts a non-numerical approach. This approach tries to avoid the use of any numbers in the representation of uncertainty. According to this approach, the uncertainty involved in a problem usually results from several factors. The numerical approaches somehow summarize all these factors into one or more numbers. But an intelligent agent normally discriminates among these factors. Thus, the numeric representation is inadequate. The main strategy of this approach is therefore centered on the foundational bases of beliefs and to try to deal with reasons behind a belief. Examples of this approach include the Truth Maintenance System (Doherty, 1979) and the Theory of Endorsement (Cohen, 1985).

The numeric approach is generally preferred to the non-numeric approach. First of all, since a strict rank ordering is hard to obtain, the non-numerical approach encounters difficulties when competing beliefs need to be selected. Second, it has been shown that people do not keep track of all the underlying reasons for beliefs (e.g., Ross & Anderson, 1982). Finally, people do often feel natural to summarize their beliefs in numbers. In Laplace's famous phrase, probability theory is nothing more than good sense reduced to a calculus. This is also evident from Bernoulli's writing that the law of large numbers is a rule that "even the stupidest man knows by some instinct of nature per se and by no previous instruction".

It is worth pointing out that any theory of formal belief management need to address two fundamental issues. First, how is belief represented? Second, how is belief revised in the light of new information? In the following two sections, I present the treatments from the theory of probability and the theory of belief functions, with an emphasis on their differences.

2.2.2. Probability Theory

The theory of probability is the oldest and best-established formal method for uncertainty management. For a long time, probability theory has been the only theoretical approach for representation and inference with uncertainty. This approach has a sound theoretical foundation, is easy to understand, and has been successfully applied to a wide range of domains. In this section, I briefly review the basics of probability theory, with an emphasis on the Bayesian perspective.

An intuitive definition of probability is that it is a measure of how likely a proposition is to be true. More formally, probability measures the uncertainty of a proposition by defining a function that maps a finite set of propositions S to real numbers in $[0,1]$. This function by definition must satisfy the following axioms:

- i) *Positivity.* $P(A) \geq 0$, where A is any proposition in S .
- ii) *Normalization.* $P(T) = 1$, where T is the ever true proposition in S .
- iii) *Additivity.* If $A \cap B = F$ Then $P(A) + P(B) = P(A \cup B)$, where A and B are propositions in S , and F is the ever false proposition in S .

The concept of conditional probability is fundamental for probability updating in the light of new information. Let A and B be any two propositions in S . Further suppose it turns out later that B is true. Then the conditional probability of A given B (when $P(B) \neq 0$) is defined as follows:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (2.1)$$

More importantly, Bayes' Theorem provides a mechanism for inverting conditional probabilities:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad (2.2)$$

2.2.2.1. Frequentism vs Bayesianism

It is interesting to note that although the above axioms and definitions form the fundamental core of probability theory, they may also be the only commonality among different approaches under the single umbrella of probability. Controversies exist from the dawn of probability theory. The most significant one is concerned with the interpretation of probability.

While the classical theory of probability assumes that probability is the relative frequency of an event in a stochastic experiment in the long run and thus is a physical property of the world (e.g., *the* probability that a coin will land heads), the Bayesian theory of probability maintains that probability is a measure of subjective belief thus is a property of the person who assigns the probability (e.g., *your* or *my* probability that the coin will land heads). One important difference between physical probability and

Bayesian probability is that no repeated trials are necessary in order to measure the latter. For example, consider the probability that Brazil will win the championship of World Cup 98. For frequentists, this is an unsolvable question since there is no associated reference class and the event is not repeatable. For Bayesians, however, this question makes perfect sense. A subjective single-event probability can be assigned based on any appropriate personal knowledge.

From the Bayesian perspective, the concept of objective probability is an illusion. A probability number reflects an observer's knowledge about the world and nothing more. A probability assessment is always in the context of one's current knowledge. Outside of an individual "probability does not exist" (De Finetti, 1974, page x). Therefore, $P(A)$ is in fact $P_K(A)$ where K is the knowledge one has when the probability is assigned. Such probabilities are subjective in the sense that individuals with different (or even the same) knowledge may legitimately assess $P(A)$ differently. It has been argued that confusion arises when efforts are made to provide a frequency interpretation for all uncertainty-related problems. For the purpose of performing inference and managing beliefs it is the concept of probability in the Bayesian approach that is relevant.

Gigerenzer (1994) gives one of the best historical reviews about the relationships between the frequencism and the Bayesianism. He claims that in the enlightenment age, objective frequencies and subjective beliefs were unified. Observed frequencies are transformed into degrees of belief through then well-established mechanisms like "traces", "vibrations", "interior images", and "impressions". All these mechanisms assume the automatic and accurate mapping of experienced frequencies into subjective

probabilities. In the early nineteenth century, when research about the illusionary and emotional nature of human reasoning was popular, the notion of subjective belief was eliminated from the realm of rigorous mathematical probability theory. It was claimed that mathematical probability was not a measure of a subjective belief. Instead, mathematical probability was about the relative frequencies in the long run, which are the objective properties of the physical world. Subjective probability has regained acceptance in this century, due to the cognitive revolution in the 1950s and the pioneering work of Bruno De Finetti and Leonard Savage in the 1930s and 1950s. It has become the dominant approach in the areas of theoretical economics, decision making, and artificial intelligence.

One common question of Bayesian probability is that probabilities seem arbitrary. Why should degrees of belief obey the rules of probability? Many researchers have suggested different sets of reasonable requirements that should be satisfied by any calculus of subjective beliefs (e.g., Savage, 1972; De Finetti, 1974). It has been shown that each set eventually leads to the same rules: the rules of probability – a conclusion summarized in the Ramsey-DeFinetti Theorem (see also Howson & Urbach, 1993). The fact that different sets of requirements all lead to the rules of probability provides a strong argument for using probability to measure beliefs (e.g., Heckerman, 1997). In other words, “probability is inevitable” (Lindley, 1994, page 6).

2.2.2.2. The Ignorance Problem

Unfortunately, in Bayesian probability theory, how to assign prior probabilities in the absence of any relevant background knowledge poses a problem. The situation brings the notion of *ignorance*, which results from the lack of necessary information to assign probability values among alternatives. It should be clear that the lack of power to handle ignorance is closely related to the lack of power to handle imprecision. Being unable to easily handle ignorance (and imprecision) is a well-known criticism of the Bayesian probability theory. The need to assign prior probabilities when no sufficient knowledge is available also presents a major barrier to the widespread acceptance of the Bayesian approach.

Let us consider a simple coin-tossing example. How would one assign the probability that the coin will land heads when no information about the fairness of the coin is available? It seems intuitive to assume that $P(\text{the coin will land heads}) = P(\text{the coin will land tails}) = \frac{1}{2}$. In fact, this intuition is a direct reflection of the well-known *principle of indifference*, which says that if there are n mutually exclusive alternatives and there is no reason to believe one more than any other, then each of these alternatives has an equal probability (i.e., $1/n$) to be true.

However, this solution is questionable (see Howson & Urbach, 1993). It has been shown that the principle of indifference often leads to paradoxes in practice. Consider an example. Suppose that you are to draw a ball from an urn, which you are told contains white and colored (either red or blue or yellow) balls in some unknown proportions. What sort of ball will you draw? On the one hand, according to the principle of

indifference, your probability of drawing a white ball will be $\frac{1}{2}$, when the balls are classified as white or colored. On the other hand, according to the same principle, your probability of drawing a white ball will be $\frac{1}{4}$, when the balls are classified as white, red, blue, or yellow. This is a paradox in the sense that a uniform distribution over one classification leads to a skewed distribution over the other.

In addition, it has been argued by many that the principle of indifference is *too* powerful to be plausible. “for it (the principle of indifference) has established the existence of God from total ignorance, and it has measured with numerical precision the probability that the sun will rise tomorrow” (Keynes, 1921, page 89; see also Howson & Urbach, 1993, page 59).

One rationale behind the principle of indifference is that it leads to a solution with the least presumptions. In the same spirit, the principle has been extended as follows: if there is constraint which implies that the probabilities of all the alternatives cannot be all equal (i.e., not uniformly distributed), then the least presumptuous assignment of probabilities is the one which maximizes the entropy. This rule is called the *principle of maximum entropy*. More formally, the entropy T of a distribution is defined as:

$$T = -\sum_{i=1}^n P_i \ln P_i \quad (2.3)$$

where n is the total number of alternatives, and P_i is the probability of the i^{th} alternative. It can be shown that the maximal entropy is achieved when one comes as close as possible to distributing the probabilities equally among all alternatives while satisfying the constraint (i.e., all alternatives cannot be equally likely). Consider an example of

rolling a six-sided die. If the die is fair, in repeated rolls, the average number of spots would be 3.5. Now suppose evidence shows that in repeated rolls the average number of spots is 4.5 but not 3.5, then, according to the principle of maximum entropy, one can calculate the probability of getting each side on the next roll, which results in: 0.05435, 0.07877, 0.11416, 0.16545, 0.23977, 0.34749, from side 1 to side 6, respectively (see Neapolitan, 1989).

It is clear that the principle of indifference is a special case of the principle of maximum entropy. As a result, the principle of maximum entropy suffers the same criticisms mentioned above as that of the principle of indifference. Moreover, it has been shown that while the notion of entropy is useful for discrete distributions, it is hard to be extended to deal with continuous distributions (see Howson & Urbach, 1993).

Other solutions to the ignorance problem have also been suggested. For example, Cheeseman (1985) argues that specifying two numbers, that is, probability and its standard deviation, solves the problem. Ignorance can then be represented by a large standard deviation. This solution is in line with a more formal solution based on higher-order probability (see Almond, 1995). That is, the uncertainty about probability assignment can be represented by a second-order probability distribution. However, it turns out that how to specify a second-order probability distribution poses another problem. One proposal is to use non-informative priors. For example, it is common to assume a $\text{beta}(0,0)$ or $\text{beta}(\frac{1}{2}, \frac{1}{2})$ distribution for the above coin-tossing problem (see Mood & Graybill, 1963, for a description of the beta distribution). The principal argument in defense of non-informative priors is that they will eventually be

overwhelmed by new evidence. An alternative proposal, from the robust Bayesianism (e.g., Berger, 1985), suggests that it is useful to perform a *sensitivity analysis* to the choice of prior. That is, examining the posterior for several competing choices of prior measures the impact of the prior on the conclusions. A good decision should be robust across reasonable choices of prior information. Robust Bayesians have typically found that the impact of priors is overestimated. For example, they argue that the sensitivity to the choice of likelihood models (i.e., the forms of probability distribution given priors) was much higher than the sensitivity to the choice of priors (see Almond, 1995).

In section 2.2.3, I will describe how the theory of belief functions offers a more extensive treatment of the problem from a different perspective.

2.2.2.3. Bayes' Theorem for Belief Revision

It is generally agreed that Bayes' Theorem regulates the way in which beliefs are revised in the light of new evidence. To illustrate, let H and E be two propositions in a finite set S on which a probability function P_K is defined (where K is the background knowledge). If new evidence suggests that E is true, then the belief of H can be revised according to Bayes' Theorem:

$$P_K(H | E) = \frac{P_K(E | H)P_K(H)}{P_K(E)} \quad (2.4)$$

It states that the probability of H given E (or the posterior probability of H) is equal to the product of the probability of E given H (or the likelihood of H) and the prior probability of H divided by the probability of E .

The belief revision property of Bayes' Theorem is more explicit if Bayes' Theorem is expressed in another form:

$$\frac{P_K(H|E)}{P_K(\bar{H}|E)} = \frac{P_K(E|H)}{P_K(E|\bar{H})} \times \frac{P_K(H)}{P_K(\bar{H})} \quad (2.5)$$

where $\bar{H}$ is the negation of H. It says that the posterior odds of H are determined by the product of two factors – the likelihood ratio and the prior odds. For given prior odds, the posterior odds increase with the likelihood ratio.

Unfortunately, Bayes' Theorem as a general method for belief revision has also been greatly challenged (see Howson & Urbach, 1993). It has been shown that Bayes' Theorem is applicable only in very restricted conditions (see also Wang, 1993).

One common criticism is that in the process of belief revision what one really needs is an unconditioned probability $P_{K'}(H)$, where K' is the new background knowledge (i.e., $K \cup \text{"E is true"}$). However, Bayes' Theorem merely provides a conditioned posterior probability $P_K(H|E)$. While $P_{K'}(H) = P_K(H|E)$ is intuitively obvious, it has been questioned recently (see Diaconis & Zabell, 1982; Howson & Urbach, 1993). It has been shown that the equation is valid only if the following condition is met: $P_K(H|E) = P_{K'}(H|E)$. That is, after learning that E is true, one sees no reason to change the conditional probability of H. Fortunately, this condition is usually met.

In addition, it has been noted that Bayes' Theorem is applicable only when the belief of E is changed from a predefined nonzero value to a definite one (see Diaconis & Zabell, 1982). In other words, Bayes' Theorem requires that the evidence is conclusive enough to show the change from $P_K(E)=p$ (and p must not be 0) to $P_{K'}(E)=1$. The fact that $P_K(E)$

cannot be zero suggests that when new evidence shows that a previously impossible proposition E (i.e., $P_K(E)=0$) is actually true, Bayes' Theorem cannot be used to update beliefs. Similarly, when new evidence is not convincing enough to show that E is true (e.g., the evidence only suggests that the probability of E should be changed, say, from 0.3 to 0.8), Bayes' Theorem is helpless as well. In this latter case, Jeffrey's rule (see Jeffrey, 1968; Diaconis & Zabell, 1982; Howson & Urbach, 1993) is needed:

$$P_{K'}(H) = P_K(H | E) \times m + P_K(H | \bar{E}) \times (1 - m) \quad (2.6)$$

where m is the new probability of E . This rule is also called the "probability kinematics", and its motivation is that "it is rarely or never that there is a proposition for which the direct effect of an observation is to change the observer's degree of belief in that proposition to 1" (see Jeffrey, 1968, page 171). It is obvious that Bayes' Theorem is a special case of Jeffrey's rule when m is equal to 1 (i.e., E is true).

Finally, Wang (1993) has argued that Bayes' Theorem should be called a *belief updating* rule rather than a belief revision rule. What he means by belief updating is concerned with the new evidence. More specifically, when the new evidence suggests that E is true (or more generally, the probability of E should be m), the system's opinion about E is completely dominated by the new evidence (i.e., $P_{K'}(E)=1$, or $P_{K'}(E)=m$), regardless of the previous opinion about E (i.e., $P_K(E)$). He claims that such a complete abandonment of old beliefs seldom occurs in human reasoning. As implied by the coherence approach to human belief revision, this criticism of Bayes' Theorem is certainly valid.

In summary, the theory of probability models belief in a precise and uncertain manner. It is difficult to handle ignorance or imprecision. In addition, Bayes' Theorem as a general method for belief revision has been greatly challenged.

2.2.3. The Theory of Belief Functions

The theory of belief functions (or the theory of evidence) was developed by Shafer (1976) as an extension to probability theory. It is also referred to as the Dempster-Shafer theory due to its extensive use of Dempster's rule for evidence combination. Different from probability theory, which only deals with uncertainty, the theory of belief functions is capable of handle both uncertainty and imprecision. In addition, it offers a different approach to belief revision. In this section, I briefly review the basics of this theory.

It is easiest to introduce the theory of belief functions from the set-theoretical perspective. For the purpose of comparison, it is worth first presenting a set-theoretical definition of probability.

Let $\Theta = \{\theta_1, \dots, \theta_k\}$ be a set of all possible outcomes of an experiment. This set is called the *outcome space* (or sample space). Note an event A is any subset of Θ . Then a probability mass function (in a discrete distribution case) f is defined as a function from Θ to real numbers in $[0,1]$ such as:

$$\begin{aligned} f(\theta_i) &\geq 0 \quad (i = 1 \dots k), \text{ and} \\ \sum_{i=1}^k f(\theta_i) &= 1 \end{aligned} \tag{2.7}$$

That is, f assigns a value to each member of Θ . Then the probability function P can be defined in terms of f as follows:

$$P(A) = \sum_{\theta \in A} f(\theta) \quad (2.8)$$

It is important to note that this definition is equivalent to the one given in the previous section. Since propositions are statements about events, propositions can then be represented by subsets of Θ . In other words, propositions and subsets of Θ are equivalent.

In the theory of belief functions, Θ is called the *frame of discernment* instead of the outcome space. One of the fundamental concepts in the theory of belief functions is *basic probability assignment*, on which belief functions are defined. Different from the probability mass function, a basic probability assignment m is a function that maps from the power set of Θ , which is the set consisting of all elements of Θ plus all its subsets, to real numbers in $[0,1]$. That is, unlike f , which assigns a value to every element of Θ , m assigns a value to every subset of Θ . Similarly, m must satisfy the following two conditions:

$$\begin{aligned} m(\emptyset) &= 0, \text{ and} \\ \sum_{A \subset \Theta} m(A) &= 1 \end{aligned} \quad (2.9)$$

where $\emptyset$ is the empty set.

It is important to note that the basic probability assignment to A is intended to measure, based on a body of evidence, the amount of belief one commits to A but not to any proper subset of A . In other words, this basic probability assignment to A can be thought as a message. This message summarizes the available evidence and reveals that the outcome of the experiment is in A , but it provides no information that would indicate

that one element of A is more, less, or equally likely than another. By definition the amount of belief committed to the empty set ϕ is 0, and the sum of all the basic probabilities assigned to the subsets of Θ is 1.

While probability theory expresses the beliefs about propositions in terms of probability functions, the theory of belief functions, as the name suggests, uses belief functions (and plausibility functions) to measure beliefs. More specifically, a *belief function* BEL represents a lower bound of the belief that the outcome of the experiment will fall in a subset, and a *plausibility function* PL represents an upper bound of the belief that the outcome will fall in the subset. They are defined in terms of the basic probability assignment:

$$BEL(A) = \sum_{B \subseteq A} m(B) \quad (2.10)$$

$$PL(A) = \sum_{A \cap B \neq \phi} m(B) \quad (2.11)$$

To elaborate, consider a subset A . Any message $m(B) > 0$ such that $B \subseteq A$ provides evidence that the outcome lies in A . Therefore, the total amount of *belief* that an outcome in A occurs is the sum of messages in the set $\{B \mid B \subseteq A\}$. That is, $BEL(A)$ measures the total amount of belief committed to A .

Similarly, any message $m(B) > 0$ such that $A \cap B \neq \phi$ is not incompatible with the outcome lying in A . Thus the sum of messages in the set $\{B \mid A \cap B \neq \phi\}$ provides a measure of the total amount of *plausibility* committed to A . In other words, $PL(A)$ measures the degree to which the evidence fails to support “not A ” (therefore, A is plausible).

It is easy to see by definition that $PL(A) \geq BEL(A) \geq 0$, which justifies the decision to use a *belief interval* $[BEL(A), PL(A)]$ to represent the total amount of uncertainty concerning A . By definition one also has $BEL(A) + PL(\bar{A}) = 1$, where $\bar{A}$ is the complement of A with respect to Θ . It further demonstrates the relationship between belief and plausibility.

It should be clear by now that probability theory is a special case of the theory of belief functions. When only the singletons of Θ but nothing else are assigned non-zero basic probabilities by the basic probability assignment m , m in fact becomes the probability mass function f . In this case, it is easy to show that $BEL=PL=P$. That is, BEL and PL become the probability function. This special belief function is called the *Bayesian belief function*.

2.2.3.1. Simultaneous Belief and Ignorance

How does the theory of belief function solve the ignorance and imprecision problem? By allowing the assignment of basic probabilities to all subsets of Θ instead of only elements of Θ , the theory of belief functions offers a unique way to represent ignorance and imprecision. More specifically, to represent the ignorance and imprecision, all one needs to do is to specify a belief interval $[BEL, PL]$, to represent a lower bound and an upper bound of the probability that a proposition is true. Total ignorance is represented by a special *vacuous belief function* which has $m(\Theta)=1$. This belief function produces the largest belief interval $[0,1]$. In addition, precise uncertainty is achieved when the belief interval converges to a point, where one returns to probability theory.

Let us consider the previous coin-tossing example where probability theory encounters difficulties. When no information is available about the fairness of the coin, what is the probability that the coin will land heads next time? It is clear $\Theta = \{\text{heads}, \text{tails}\}$. Without further information, all one can do is to assign the following basic probabilities:

$$m(\{\text{heads}\})=0, \quad m(\{\text{tails}\})=0, \quad m(\{\text{head}, \text{tails}\})=1$$

Note that this is a vacuous belief function and represents total ignorance. These basic probability assignments translate into beliefs and plausibilities as follows:

$$BEL(\{\text{heads}\})=0, \quad BEL(\{\text{tails}\})=0, \quad BEL(\{\text{head}, \text{tails}\})=1$$

$$PL(\{\text{heads}\})=1, \quad PL(\{\text{tails}\})=1, \quad PL(\{\text{head}, \text{tails}\})=1$$

That is, the belief of the proposition that the coin will land heads next time can be represented by the belief interval $[0,1]$, which is equivalent to say that, “there is a 0-100% chance that the coin will land heads next time”.

Consider another example. An urn has 10 balls in it. There are 2 red, 3 green, and 5 yellow. Suppose one is told that (a) every red ball has a fortune cookie inside; (b) none of the green balls has a fortune cookie in it; and (c) no information is available about whether yellow balls contain fortune cookies or not. What is the probability of drawing a ball with a fortune cookie next time? Obviously probability theory encounters difficulties again. According to the theory of belief functions, $\Theta = \{\text{fortune}, \text{no fortune}\}$, and one can assign the following basic probabilities:

$$m(\{\text{fortune}\})=0.2, \quad m(\{\text{no fortune}\})=0.3, \quad m(\{\text{fortune}, \text{no fortune}\})=0.5$$

These again translate into beliefs and plausibilities as follows:

$$BEL(\{\text{fortune}\})=0.2, \quad BEL(\{\text{no fortune}\})=0.3, \quad BEL(\{\text{fortune, no fortune}\})=1$$

$$PL(\{\text{fortune}\})=0.7, \quad PL(\{\text{no fortune}\})=0.8, \quad PL(\{\text{fortune, no fortune}\})=1$$

This belief function is to say that , “there is a 20-70% chance that the ball drawn next time will have a fortune cookie inside”.

Comparing to probability theory, the theory of belief functions provides a simultaneous representation of belief and ignorance. The cost, however, is weaker decision making power. For example, suppose one has an action rule in the above fortune cookie example which says that an action should be taken if the probability is greater than 0.6. One would not be able to make such a decision easily in this case.

2.2.3.2. Evidence Combination

Unlike probability theory, where belief revision is achieved by Bayesian conditionalisation when some subset in the frame is established with certainty, in the theory of belief functions, belief revision is achieved by Dempster’s rule for combining two or more belief functions from distinctive bodies of evidence. This kind of evidence combination is more general in the sense that new evidence is not restricted to be a certain occurrence of an event – two or more belief functions can be combined.

More specifically, let BEL_1 and BEL_2 be two belief functions over the same frame Θ representing independent bodies of evidence. Then Dempster’s rule provides a way to generate the *direct summation* of the two belief functions, $BEL_1 \oplus BEL_2$, which represents the joint evidence.

Dumpster's rule consists of two steps. First, the convolution of BEL_1 and BEL_2 , $BEL_1 * BEL_2$, is calculated by:

$$m_1 * m_2(C) = \sum_{\substack{A, B \subseteq \Theta \\ A \cap B = C}} m_1(A) \times m_2(B) \quad (2.12)$$

where m_1 and m_2 are the corresponding basic probability assignments of BEL_1 and BEL_2 . It should be noted that the convolution of two belief functions usually does not result in a belief function since $m_1 * m_2(\phi)$ (where ϕ is the empty set) usually is not equal to zero. Such a belief function is referred to as an unnormalized belief function. In addition, for obvious reasons, the value $c = m_1 * m_2(\phi)$ is called the *conflict* between the two combined belief functions. It is a measure of the degree to which the evidence of the two belief functions conflicts.

The second step is normalization. This is achieved by dividing each non-zero point of $m_1 * m_2$ by $1 - m_1 * m_2(\phi)$, and then setting $m_1 * m_2(\phi)$ to be zero. That is,

$$m_1 \oplus m_2(A) = \begin{cases} 0 & \text{if } A = \phi \\ \frac{m_1 * m_2(A)}{1 - m_1 * m_2(\phi)} & \text{otherwise} \end{cases} \quad (2.13)$$

An interesting example of Dumpster's rule is given in Almond (1995). Suppose an object is to be classified into one of three categories: x , y , or z . There are two experts who do the classification independently. The first expert summarizes the evidence as follows:

$$m_1(\{x\}) = 1 - e, \quad m_1(\{y\}) = e, \quad m_1(\{z\}) = 0$$

where e is an arbitrary value in $[0, 1]$. The second expert's evidence is summarized as:

$$m_2(\{x\})=0, \quad m_2(\{y\})=e, \quad m_2(\{z\})=1-e$$

How does one join the two experts' evaluations together? One can use Dempster's rule to combine the evidence as follows:

$$m_1 \oplus m_2(\{x\}) = m_1 \oplus m_2(\{z\}) = \frac{0}{1 - [(1-e)e + (1-e)e + (1-e)(1-e)]} = 0$$

$$m_1 \oplus m_2(\{y\}) = \frac{e^2}{1 - [(1-e)e + (1-e)e + (1-e)(1-e)]} = \frac{e^2}{1 - (1-e^2)} = 1$$

This suggests that the object should belong to class y . This result makes sense since class y is supported somehow by both experts.

It is interesting to note that the conflict is equal to $1-e^2$. Thus, as $e \rightarrow 0$ the conflict increases, indicating that the conclusions are increasingly based on ruling out what the two experts disagree. On the other hand, as $e \rightarrow 1$, the conflict decreases, indicating that the conclusions are increasingly based on reinforcing the agreement of the two experts.

It is worth pointing out that although Dempster's rule only applies to belief functions over the same frame, algorithms for enlarging and refining a frame have been developed. Due to this kind of *frame projection*, Dempster's rule can be applied in more general situations (see Almond, 1995).

It has been shown that when combining two Bayesian belief functions, Dempster's rule becomes Bayes' Theorem (see Almond, 1995). In this sense, Bayes' Theorem is a special case of Dempster's rule.

2.2.4. Graphical Belief Modeling

Recent years have seen a rapid development of using various formal belief management techniques to solve real-world problems. Graphical belief modeling is of special interest.

Not surprisingly, current development of graphical belief modeling typically adopts one of the two techniques for belief management: probability and belief functions. When probability is used, the model is commonly called *Bayesian belief networks*, or simply *Bayesian networks* or *belief networks*. When belief functions are adopted it is generally called *graphical belief models*.

Although the primary belief management techniques are different, both Bayesian belief networks and graphical belief models make use of graphs. The benefits of graphs are twofold. On one hand, a graph provides an intuitive and easy way to describe models. On the other hand, a graph can easily and explicitly capture the structure of the problem. In most complex real-world problems, taking advantage of the structures of the problems is the only way to produce computationally tractable models.

It should be clear that belief functions represent both uncertainty and imprecision, and probabilities represent only uncertainty. Therefore, graphical belief models have a greater expressive power than Bayesian belief networks. A price is paid, however: graphical belief models are more complex, have a higher computational cost, and have weaker decision making power. In addition, methods for learning Bayesian networks from data have been developed recently; and no corresponding methods are available for graphical belief models. Due to these reasons, Bayesian belief networks become the

more commonly adopted models to solve real-world belief management problems. In this section, I give a brief review of Bayesian belief networks (for detailed descriptions of Bayesian belief networks and graphical belief models, see Pearl, 1988; Jensen, 1996; Almond, 1995).

A Bayesian belief network is a graphical model that encodes probabilistic relationships among a set of variables of interest. Intuitively, all one needs is a joint probability distribution over the set since all the probabilistic relationships among the variables are fully determined by this joint probability distribution. However, fully specifying a joint probability distribution is usually complex or impossible, especially when the variable set is large.

A Bayesian belief network solves this problem by analyzing and exploiting the structural relationships among the variables. One of the most important relationships is conditional independence. A variable A is said to be conditionally independent of a variable B given a third variable C if after knowing C , knowing B provides no additional information about A . More formally, we have:

$$P(A \mid B, C) = P(A \mid C) \quad (2.14)$$

Conditional independence allows the factorization of the joint probability distribution, which greatly simplifies the model specification and computation.

More specifically, a Bayesian belief network for a set of variables (continuous or discrete) consists of (1) a network structure that encodes a set of conditional independence assumptions about the variables, and (2) a set of local probability distributions associated with each variable. The network structure is a *directed acyclic*

graph. Each node in the network corresponds to a variable. Directed arcs between nodes usually represent causal relations, which are from parents (causes) to children (effects). The way to specify local probability distributions depends on individual nodes. If a node has no parent node, then its local probability distribution is an unconditioned probability distribution. If a node has parents, then its local probability distribution is the conditional probability distribution (conditioned on its parents).

In many cases, both the network structure and the local probability distributions are relatively easy to specify based on background knowledge about the problem of interest. Causal relations between variables are usually easy to be identified, which in turn often guarantee the conditional independence assumptions being satisfied (see Pearl, 1988). Local probabilities are typically determined based on previous experience or by asking an expert. Together, these components in fact define a joint probability distribution, which is desired.

Once a Bayesian belief network is constructed, it can be used to determine various probabilities of interest. Because a Bayesian belief network essentially defines a joint probability distribution over the set of variables, in principle one can use it to compute any probability of interest. In fact, using a Bayesian belief network to determine various probabilities that are not stored directly is one of the primary purposes of building it. In general, the computation of a probability of interest in a Bayesian belief network is known as *probabilistic inference*.

Various algorithms have been developed to conduct probabilistic inference efficiently. For example, one algorithm continuously uses Bayes' Theorem to reverse

arcs in the network until the answer to the query can be read directly from the graph (see Jensen, 1996; Heckerman, 1997). However, it has been shown that exact inference in an arbitrary Bayesian belief network for discrete variables is NP-hard (Cooper, 1990). Even approximate inference is NP-hard (Dragum & Luby, 1993). These results imply that for an arbitrarily complex network there is no efficient way to conduct probabilistic inference, either exactly or approximately. Fortunately, for many applications where structures are simple, efficient inference is possible. These conclusions also raise interesting questions about how people do probabilistic inference. Given that it is so hard, what kind of solutions are people pursuing when facing real-world problems? Related questions will be discussed in the next section.

To illustrate, let us use a simple Bayesian belief network to solve the famous Monty Hall puzzle. In this puzzle, a player has the chance to win some prize if he or she is lucky enough to find the prize behind one of three doors. The game goes like this:

1. The player is asked to select one of the three doors (without opening it).
2. Monty Hall (the host, who knows where the prize is) opens another door which does not contain the prize.
3. The player is given an opportunity to redo his or her selection (i.e., to select one of the two closed doors).
4. The player gets the prize if it is behind the door he or she finally selected.

The problem of the puzzle is: What should the player do in the second selection?

In other words, should he or she change the first selection or not? Some would say

that it does not matter because it is equally likely that the prize is behind the remaining two doors. A simple Bayesian belief network can be built to test this.

The network is shown in Figure 2.1. It has 3 nodes (or variables): *Prize* (the door which hides the prize), *First Selection* (the player's first choice), and *Open* (the door Monty Hall opens after the player makes the first selection). It is clear that all variables have the same outcome space, {door1, door2, door3}. The two directed arcs indicate that which door Monty Hall opens depends on where the prize is and which door the player first selects.

The local probability distributions consist of three discrete probability tables, which are shown beside each corresponding node in Figure 2.1. It is assumed that the prize is equally likely to be behind any door, and the player's first selection is equally likely to be any door as well. Furthermore, it is assumed that when Monty Hall has a choice, which occurs when the player's first selection happens to contain the prize, she chooses one of the remaining two doors without preference.

How does the network help the player make decisions? When the network is told the player's first selection (e.g., door1) and the door Monty Hall opens (e.g., door2), the network does complex probabilistic inference (see Pearl, 1988 for details) and makes the following conclusion: The probability that the prize are behind door1 is $1/3$, and the probability that the prize is behind door3 is $2/3$. This suggests that the player have the best chance of winning if he or she changes the first selection, a quite surprising answer.

Figure 2.1. A simple Bayesian belief network to solve the Monty Hall puzzle.

One of the common criticisms of Bayesian belief networks is, familiarly, that prior background knowledge is necessary to construct the networks. Not only do the causal relations between variables require prior knowledge to be constructed but so do the local probability distributions, knowledge that the constructors usually do not have. This criticism has been answered soundly recently due to an important development that allows Bayesian belief networks to learn from observed data (see Heckerman, 1997).

Various Bayesian learning techniques have been proposed to allow learning probabilities in a Bayesian belief network. The basic idea is to use data to update the probabilities of a given Bayesian belief network. As long as the distributions of the probability parameters are well specified, the posterior probabilities given data can be obtained via Bayes' Theorem. For example, when a variable is known to follow a binomial distribution (e.g., an event occurs or does not occur), it is common to assume the probability parameter of this variable follows a $\text{beta}(1,1)$ (i.e., uniform) distribution in the beginning. Suppose that one later obtains 100 observations, and the event of interest occurs in 20 of them. Then it has been shown that the parameter distribution can be updated to follow a $\text{beta}(1+20, 1+80)$ distribution, which suggests that the posterior probability of the event given previous observations is $(1+20)/(1+100) \approx 0.21$, a significant improvement from the previous uniform assumption.

An even newer development in the field allows learning both the network structures (i.e., causal relations) and local probabilities simultaneously, although it is highly experimental at present (see Heckerman, 1997).

It is worth pointing out that human beings as well constantly learn and improve themselves as more experience is gained. Learning combines prior knowledge with present observations to produce improved knowledge. It is surprising to note that this dynamic learning aspect of belief management is often ignored.

2.2.5. Summary

Uncertainty is usually inevitable in real-world problems. For those AI researchers who are interested in building artificial systems to deal with real-world problems quantitative uncertainty management is essential.

Uncertainty has many faces. On the one hand, it is well agreed that there is a distinction between imprecision and uncertainty. On the other hand, depending upon where uncertainty comes from, one can distinguish between external uncertainty and internal uncertainty.

The theory of probability is the oldest and best established formal method for uncertainty management. While for frequentists probability is an objective property of the physical world, for Bayesians probability is nothing more than a measure of subjective belief based on personal knowledge. Subjective probability has been the focus in the areas of theoretical economics, decision making, and artificial intelligence.

Probability theory represents uncertainty in a precise manner. It has difficulties to handle both imprecision and uncertainty simultaneously. One of the common criticisms of probability theory is that it cannot represent ignorance, a situation where no information is available for belief assignment.

Bayes' Theorem is often regarded as a general mechanism for belief revision. However, it has been shown that Bayes' Theorem is applicable only in very restricted conditions.

The theory of belief functions is an extension to probability theory. By defining a belief function and a plausibility function, the theory of belief functions uses a belief interval $[BEL, PL]$ to represent both imprecision and uncertainty simultaneously. In addition, Dempster's rule offers a more general way of doing belief revision through evidence combinations.

Although the theory of belief functions has more expressive power, it has a higher computational cost and can lead to weak decisions.

Graphical belief modeling has been proposed to deal with complex real-world problems by using graphs. Graphs make the explicit representation of problem structures easier. While graphical belief models based on belief functions have been developed, Bayesian belief networks, which are based on probability, are more extensively studied. Though the general problem of probabilistic inference is NP-hard, many problems can be solved efficiently. More importantly, various sophisticated techniques have been developed recently to allow Bayesian belief networks to learn from observations. Both network structures and local probabilities can be learned.

2.3. Order Effects in Human Belief Revision

Although both abstruse philosophical inquiries and formal mathematical specifications provide valuable insights and results about belief and belief revision, the

full spectrum of the problem cannot be fully understood without a thorough understanding of how people actually manage belief. It is the empirical psychological studies that try to investigate this aspect of the problem.

Quite interestingly, the dominant findings in empirical research suggest that observed human behaviors are not consistent with the predictions of normative postulates, such as formal logic and Bayesian probability rules. A large number of studies have demonstrated that people often systematically deviate from these normative postulates. With the assumption that these normative postulates prescribe how a reasonable individual should behave, these systematic deviations are often labeled as *cognitive illusions, biases, or fallacies* (e.g., Tversky & Kahneman, 1974; Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1996; Evans, 1989). Several well-known biases are: *base rate fallacy, conjunction fallacy, overconfidence, and belief bias*.

The order effect in human belief revision is yet another interesting finding in empirical research. By the similar standard, the order effect should also be called a bias since the normative postulates, in particular Bayes' Theorem, have no room for it – it simply violates commutativity. However, as many researchers have already pointed out, calling it a bias is nothing more than giving it a label, which provides no help to understand how and why the order effect occurs.

Whether the deviations of behaviors from normative prescriptions imply human irrationality is an interesting yet controversial issue. Drawing such a conclusion was indeed a fairly natural thing to do until it has been seriously questioned recently. A number of researchers have shown that the reports of errors and biases are often

misleading and exaggerated (e.g., Gigerenzer, 1994). It has also been argued that people may follow some other principles that are reasonable but not captured by any of these normative postulates.

In this section, I first briefly review previous studies on heuristics and biases, then I present the view of bounded human rationality. In the context of the controversy, I focus on the studies and theories of the order effect in human belief revision. I give a summary at the end of this section.

2.3.1. Heuristics and Biases

A large body of evidence has shown that observed human behaviors often deviate from normative prescriptions when people make judgments under uncertainty. The program of heuristics and biases, initiated by Kahneman and Tversky in the 1970s, systematically identifies various kinds of departures and holds that a small number of judgmental *heuristics*, which are loosely characterized as distinctive mental operations or shortcuts, are responsible for these deviations (see Tversky & Kahneman, 1974; Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1973, 1996). Regarding these deviations as human errors, Kahneman and Tversky (1973, page 237) made the following statements:

In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors.

With the great motivation to disclose the flawed aspects of human intelligence, researchers have identified many such “severe and systematic errors”.

Base rate fallacy is perhaps the most studied one. Base rate information expresses how likely an event occurs in general and thus describes a fundamental statistical property of the environment. A sufficient appreciation of base rate information is important for any reasonable individual to make judgments under uncertainty. However, studies have shown that people tend to be dominated by evidence at hand and completely ignore or significantly underweight base rate information even when base rate information is readily available. This phenomenon is often called base rate fallacy (for reviews, see Bar-Hillel, 1980, 1983). Base rate fallacy has been demonstrated by so many studies that it has been concluded that “the genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact” (Bar-Hillel, 1980, page 215).

In an interesting illustration of base rate fallacy, Eddy (1982) asked one hundred physicians the following “mammography problem”:

The prevalence of breast cancer is 1% (in a specified population). The probability that a mammography is positive if a woman has breast cancer is 79%, and 9.6% if she does not. What is the probability that a woman who tests positive actually has breast cancer?

A simple application of Bayes’ Theorem suggests that the answer is 8%. However, It turns out that most of the physicians were hopelessly lost – 95 out of 100 physicians estimated the probability to be about 75%, a significant misjudgment. The reason, as believed, is the lack of appreciation of the background base rate information. A false

positive in the mammography is a rare occurrence, but so is having breast cancer. In this case, a false positive is less rare than the cancer itself (9.6% vs 1%). This is why ignoring base rate information leads to a significant false alarm here.

Conjunction fallacy is another well-known bias. It refers to the phenomenon in which people judge that a conjunction is more probable than any one of the conjuncts (for a review, see Tversky & Kahneman, 1983). This is a fallacy since according to the conjunction rule in probability theory, the probability of a conjunction cannot be larger than the probability of any one of conjuncts, that is, $P(A \cap B) \leq P(A)$. One classic demonstration of conjunction fallacy is Linda problem (see also Kahneman & Tversky, 1996). Participants are given a description of a young woman, named Linda, who resembles a feminist, but not a bank teller. Participants are then asked to determine which of two hypotheses is more likely: (a) Linda is a bank teller, or (b) Linda is a bank teller who is active in the feminist movement. Although based on the probability rule, the probability of (b) can never be larger than that of (a), it is surprising to note that most participants (80% - 90%) showed conjunction fallacy by stating that (b) is more likely.

Another interesting bias is *overconfidence fallacy*, which refers to the tendency to overestimate one's confidence relative to the actual accuracy (for reviews, see Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; Erev, Wallsten, & Budescu, 1994). In a typical experiment, participants are asked to answer multiple-choice questions and state their probability (or confidence) that they have answered each question correctly. The proportion of correct answers in each of the confidence categories is calculated. The major finding is quite surprising: when participants express 100% confidence, the relative

frequency of correct answers is only about 80%; when participants express 90% confidence, the relative frequency of correct answers is only about 75%; when participants express 80% confidence, the relative frequency of correct answers is only about 65%, and so on. That is, values of confidence are systematically higher than the actual relative frequencies. This systematic discrepancy between mean confidence and overall accuracy is called overconfidence fallacy.

How do these biases occur? The standard conclusion from the heuristics and biases program is that “people lack the correct programs for many important judgmental tasks” (Slovic, Fischhoff, Lichtenstein, 1976). As a result, instead of adopting the proper statistical algorithms, people rely on non-statistical, quick and dirty heuristics, which often result in biases. For example, it has been argued that both base rate fallacy and conjunction fallacy are due to a *representativeness heuristic*, which refers to a tendency to make judgments based on similarity or correspondence. Therefore, in the breast cancer problem, the close correspondence between positive mammography and breast cancer easily leads to a high probability judgment. Similarly, in the Linda problem, the probability is assessed by calculating the similarity between the description of Linda and each of the alternatives, which clearly leads to the wrong selection. In the same vein, it is argued that overconfidence bias results from miscalibration, a failure to calibrate the scales for subjective probability and external frequencies.

As many researchers (e.g., Gigerenzer, 1994) have pointed out, the biases mentioned above are typically related to making judgments in terms of Bayesian probability.

However, it is important to note that biases also occur in other belief related judgmental tasks.

For example, Ross and Lepper (1980) found that people are reluctant to give up their beliefs even when their original evidential bases are completely destroyed or when the new evidence suggests that they are wrong. This is called the *belief preservation effect*. Evans, Barston, and Pollard (1983) discussed the *belief bias effect*, which refers to the situation where people tend to endorse conclusions which are a priori believable as valid and those which are a priori unbelievable as invalid, completely ignoring the logic of the problem. A closely related but distinct finding is *confirmation bias* (see Klayman & Ha, 1987; Klayman, 1995). Confirmation bias refers to a tendency to seek out evidence that confirms rather than disconfirms current beliefs.

Finally, people show *order effects*. The order effect is a striking finding in human belief revision. It suggests that people are sensitive to the temporal order of evidence presentation when revising their beliefs, even when the temporal order is meaningless. This is clearly a phenomenon that does not conform to the normative rules.

Human beings, as a species, are certainly not sure of the pureness and completeness of their rationality. They put it under skepticism whenever they have a chance. What do these robust findings of human biases tell about human rationality? For many researchers, it is quite natural and straightforward to draw a conclusion that people systematically deviate from rationality. In fact, this conclusion is so well publicized that “the psychology of reasoning had become nearly synonymous with the investigation of

the irrational” (Gigerenzer, 1994, page 130). For many others, however, such a conclusion is oversimplified and premature.

2.3.2. Bounded Human Rationality

The conclusion of human irrationality is based on the assumption that following normative postulates is *the* criterion for rationality. Unfortunately, this assumption has been seriously questioned.

It has long been argued that the functional view is more sensible for human rationality than the normative view (e.g., Simon, 1955; Gibson, 1966; Anderson, 1981; Marr, 1982). According to the functional view, people try to be as rational as possible. However, due to various cognitive constraints, such as limited cognitive capacities and inherent uncertainty, the best solution to a problem is usually not within reach or sight. Instead of struggling for that best solution, which usually entails a very high cost, people are usually satisfied with a “good enough” solution at a reasonable cost. In other words, people usually try to find a balance between expected gain and actual cost. From an ecological point of view, the act of seeking a maximum gain with an affordable cost is the very meaning of rationality, which unfortunately is not caught by normative postulates. To be distinguished from the normative sense of rationality, this kind of rationality is usually referred to as bounded rationality.

At least one argument for the bound rationality is speculative. While it is true that people fail on many normative tests of rationality, it is a striking fact that most AI systems, which can pass all these normative tests, are brittle. Meanwhile, it is these non-

normative and “irrational” human cognitive systems that survive in this world and have created splendid cultures and sophisticated sciences. Therefore, people simply cannot be irrational. Someday people may become mature enough to understand how rational they really are.

Using empirical evidence, such as all the biases mentioned above, to prove human irrationality also raises methodology questions. For example, Cohen (1981) argues that psychological experiments cannot be used to demonstrate irrationality. One main argument concerns the competence/performance distinction first proposed by Chomsky (1957). People do make mistakes in various experimental studies, but these mistakes may be attributed to performance factors. They do not necessarily indicate defects in the underlying competence for rational reasoning. In addition, it has also been noted that the experimental tasks used to demonstrate human biases are often arbitrary, artificial, confusing, framed, and generally unrepresentative of real life situations, which certainly opens a door for arguments like the one that people may not show these biases when they face real-world problems.

In recent years, the view of bounded rationality has become generally accepted (e.g., see Evans, Over, & Manktelow, 1993; March, 1994). The ACT-R theory developed by John Anderson (1990, 1991, 1993; Anderson & Schooler, 1991) is a representative theory that takes this bounded rationality view as one of its foundations. According to this theory, human cognition can be understood in terms of a rational analysis. This rational analysis can be illustrated at two levels. At the sub-symbolic level, ACT-R claims that the fundamental cores of human cognition are always adaptive in that they tune cognition

to the statistical structure of the environment. From this perspective, it claims that forgetting, for example, which is often taken as a strong evidence for the flawed human memory and an indication for human irrationality, is not a failure of memory but in fact a statistically tuned adaptation. The forgetting curve in memory mirrors the way information tends to lose its utility with time in the real world. Consequently, forgetting is indeed a rational thing to do.

At the symbolic and computational level, the ACT-R theory maintains that human behaviors are goal-oriented and can be predicted as the solutions to mathematical optimization problems. In ACT-R, one such problem, for example, is to decide which production rule to fire in a certain situation. Instead of conducting an exhaustive search to seek the best possible solution, ACT-R attacks this challenge by treating it as a mathematical optimization problem and seeking a good enough solution. More specifically, the theory claims that the production rule with the highest expected gain should be selected to fire. The expected gain of each production rule in a certain situation is determined by a quantity $PG-C$, where P is the probability that firing this production rule will achieve the desired goal; G is the worth of the desired goal; and C represents the cost to achieve the goal. With this kind of bounded rationality built-in, ACT-R has shown great modeling power. As a cognitive architecture, it has been supported by a large number of studies (see Anderson, 1993).

One big challenge for any theory that takes this bounded rationality view is to convincingly illustrate under what circumstances biases are sensible. Many studies have been carried out for this purpose.

Noting that normative postulates often assume a stationary and discrete environment, many researchers have argued that the environment is neither stationary nor discrete. People's behaviors may appear biased or deficient according to those normative postulates, but they are in fact very functional and optimal when a continuous and dynamically changing environment is assumed (e.g., Hogarth, 1981; Jungermann, 1983).

Systematic investigations have been carried out to determine the conditions under which the biases appear or disappear. For example, Gigerenzer, among others (Gigerenzer, 1991, 1993, 1994, 1996; Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996) argues that the human mind, as a frequency-monitoring device, simply is not good at those awkward subjective probability judgment tasks used in traditional laboratory research. People can assess relative frequencies reasonably well although they perform miserably in assessing subjective probability. Since subjective probability is not something people are equipped with, "biases are not biases" (Gigerenzer, 1991, page 86), and heuristics, as an unfalsifiable hypothesis, are meant to explain something that does not exist. It has been shown that all the biases, including the base rate fallacy, conjunction fallacy, and overconfidence, disappear or are significantly reduced when information is presented to participants in frequency format (e.g., 10 out of 100) instead of single-event subjective probability format (e.g., 10%) (Gigerenzer & Hoffrage, 1995).

The frequency idea has also been experimentally demonstrated by other studies. For example, it has been shown that, in most cases, people can automatically and accurately acquire and use base rate information (Hasher & Zacks, 1979, 1984; Medin & Edelson, 1988). In addition, from an ecological point of view, natural sampling theory (Kleiter,

1993) claims that there is a condition in which it is rational to neglect base rate information since the base rate actually does not enter into the estimated Bayesian probability. This condition is called “natural sampling”, which refers to the situation in which experience about classes and their members is acquired naturally. For example, without losing any information, Eddy’s (1982) “mammography problem” can be re-represented in a “natural sampling” form in the following story (see also Gigerenzer, 1995):

There is an experienced physician. In her career she has seen 1000 women, 10 of whom had breast cancer. Of those 10, 8 got positive mammography. Of those 990 who had no breast cancer, 95 got positive mammography. Now a new patient, with a positive mammography, comes to see her. What is the probability that she has breast cancer?

Under this condition, in order to solve the problem, all the physician needs is the number of cases that had both positive mammography and breast cancer (here, 8) and the number of positive mammography cases (here, 95+8). The probability is simply $8/(95+8) = 8\%$. It is clear that under this representation keeping track of the base rate of breast cancer is not necessary. It is argued that evolutionarily this is the most natural way people acquire information (see Gigerenzer, 1995).

Erev, Wallsten, and Budescu (1994) provide an interesting analysis of overconfidence. They found that depending upon different methods of data analysis, one can simultaneously obtain overconfidence and underconfidence. Consequently, the overconfidence bias is simply an artifact of the method of data analysis. In addition, due to the stochastic nature of human judgment, an individual whose true judgments are

accurate can appear to be under- or overconfidence simply as a function of random errors with which those judgments are expressed.

2.3.3. The Order Effect

As stated previously, belief revision is the process of revising the current beliefs in the light of new information. The ultimate reason for belief revision is that the dynamic changes in the environment make the current beliefs out-of-date or inappropriate. Both philosophical inquiries and formal quantitative investigations provide valuable insights about belief revision and propose normative rules that regulate how the rational belief revision should be done. Unfortunately, as the above review has suggested, people, though adaptive and intelligent, do not generally follow those normative rules. Very often they show biases, either by under-weighting some relevant information or by over-weighting some irrelevant information. Although explanations have been proposed, in many cases how and why people show these biases still remain controversial and mysterious. The order effect in human belief revision is one such case.

The order effect is a robust finding in situations where people do belief revision over time. When information relevant to the current situation becomes available piece by piece sequentially, it is essential for people to constantly revise their beliefs to reflect the changing situation. Although in many cases the temporal order of information presentation carries important information that is critical for solving the problem of interest, in many other cases, the temporal order is simply a byproduct of information distribution across time and carries no significant or meaningful message. Any

sensitivity or over-weighting to the temporal order of information presentation in the latter cases indicates an occurrence of the order effect.

In a classic demonstration of the order effect, Asch (1946) read to participants a number of discrete characteristics of a person. The participants were asked to describe the impression they formed. For one group of participants, the following series were read:

intelligent – industrious – impulsive – critical – stubborn – envious.

And the following series were read for the other group of participants:

envious – stubborn – critical – impulsive – industrious – intelligent.

That is, the two series are identical with regard to the content, differing only in the order of presentation. More specifically, the first series begins with qualities of high merit and proceeds to qualities with low merit. In the second series the order is reversed. The results showed a significant group difference. In general, the person described by the first series was judged more favorably than the person described by the second series. This is called the primacy effect. Clearly the order of the presentation makes the difference.

In a more recent study of human belief revision, Zhang, Johnson, and Wang (1997) used a tactical decision-making task and asked the trained participants to predict the true identity of unknown targets based on a sequence of two evidence items. They found the final judgments were significantly affected by the final piece of evidence. In other words, participants made more positive identifications when the final evidence is positive than when it is negative. This is called a recency effect.

Although order effects like these have been repeatedly found in many studies (e.g., Miller & Campbell, 1959; Anderson & Hubert, 1963; Anderson & Farkas, 1973; Rywick & Schaye, 1974; Adelman & Tolcott, 1993; Adelman, Bresnich, Black, Marvin, & Sak, 1996; Tubbs, Gaeth, Levin, & Osdol, 1993; for a review, see Hogarth & Einhorn, 1992), disagreement exists in what order effects really mean. One view, suggested by Hogarth and Einhorn (1992), asks the following straightforward question: “[if a person] processes the same information in a different order, would this affect the final judgment?” (Hogarth & Einhorn, 1992, page 2). In this view, therefore, the order effect is meaningful only in the situation where two identical (with regard to the content) sequences of information are presented in different orders and the final judgments are compared. More specifically, for example, if one is given two pieces of evidence (e.g., E+ and E-, and E+ is more positive than E-) in two different orders (i.e., E+ then E-, vs., E- then E+), the order effect is meaningful only when the final judgments from the two sequences are compared. Furthermore, if the E+E- sequence yields a more positive judgment than the E-E+ sequence, it is called *primacy effect*. Conversely, if the E+E- sequence yields a more negative judgment than the E-E+ sequence, it is called *recency effect*. If there is no difference between the two judgments, then no order effect occurs. Since this view defines the order effect as the difference between the final judgments resulting from two content-identical but order-reversed information sequences (e.g., E+E-, vs., E-E+), it is natural to label this view as the *difference view*.

The second view is advocated by Anderson (1981) and Schlottmann & Anderson (1995) in the framework of information integration theory. According to this view, the

difference of the final judgments between the two evidence sequences is neither a necessary nor a sufficient condition for the order effect. The order effect is essentially a position effect. In other words, the order effect occurs whenever an information item is weighted differently at different positions. Therefore, no comparison between two content-identical but order-reversed information sequences (e.g., E+E- and E-E+) is needed. For any single evidence sequence (e.g., E+E-), if the weights, which can be estimated based on performance data via techniques developed in the information integration theory, for the information items at different positions (e.g., E+ and E-) are different, then the order effect is observed. Furthermore, a primacy effect occurs when the weights for earlier items (e.g., E-) are larger than those for later items (e.g., E+), and a recency effect occurs when the reverse is true. This view can be naturally called the *position view*.

One major difference between the two views is the different and sometimes inconsistent indices of the order effect, which furthermore implies different understandings about what the order effect really is and how it occurs. Compared to the difference view, the position view specifically claims that order effects result from the different weights that are assigned to different positions in the evidence sequence. Therefore, larger weights to earlier information items produce primacy effects, and larger weights assigned to later information items produce recency effects. If the final judgments of two evidence sequences, which are identical with regard to the content but differ in the presentation order, are different, this is a consequence of this weight assignment mechanism. In contrast, the difference view is interested in the differences

between the final judgments per se. It seems that the difference view of the order effect provides a narrower definition in the sense that it cannot be applied to the simple primacy and recency effects that have been extensively studied in traditional experimental psychology.

2.3.3.1. Task Characteristics

Despite the existence of different views and indices of the order effect, one underlying question is common: How does the order effect occur? One strategy is to investigate the task or stimulus characteristics and determine in what situations the order effect occurs and in what situations it does not occur; and when it occurs, whether it is a recency effect or a primacy effect. Many of the previous studies adopted this strategy. Consequently many important task or stimulus characteristics have been identified.

Response mode — Step-by-Step (SbS) vs End-of-sequence (EoS). Hogarth and Einhorn (1992) distinguish two different response modes. In the SbS mode, participants are asked to evaluate and report their beliefs after the presentation of each information item. In the EoS mode, participants are only asked to give their final beliefs after all the information items have been presented, thus no intermediate beliefs are queried. In usual cases (e.g., the evidence sequence is not long and not complex), they find that EoS induces the primacy effect, and SbS induces the recency effect (e.g., Stewart, 1965).

Evidence encoding — Evaluation vs Estimation. Hogarth and Einhorn (1992) also distinguish evaluation tasks from estimation tasks. In an evaluation task, evidence is encoded as positive or negative with respect to the hypothesis under consideration.

Therefore, positive evidence (relative to the hypothesis) increases belief and negative evidence decreases it, irrespective of the level of current belief. As a result, evidence is considered as bipolar in nature (e.g., from -1 to $+1$). An example is that one encodes the sound of thunder as -0.4 with respect to the hypothesis that it is now a sunny day, indicating that the thunder is considered as negative evidence and will decrease the current belief. In contrast, in an estimation task, evidence is evaluated relative to the current level of belief. Sequential integration of information across time involves a moving average that is adjusted by new evidence in the direction in which this deviates from the present level of belief. Therefore, positive evidence (relative to the hypothesis) does not necessarily increase the current belief, if the current belief is already high enough. Similarly, negative evidence does not necessarily decrease the current belief, if the current belief is already low enough. In these cases evidence is encoded as another measure of believability in the hypothesis, and is therefore unipolar (e.g., from 0 to 1). An example is that one encodes the sound of thunder as 0.1 with respect to the hypothesis that it is now a sunny day, indicating that given the thunder it is very unlikely that it is sunny now. This estimation (i.e., 0.1) is averaged with the current belief to determine the final belief. It has been shown that, incorporated with other factors, the estimation/evaluation distinction induces different order effects.

Memory of evidence. Many studies are concerned with the relationship between belief revision and the verbal recall of evidence items. Anderson and Hubert (1963) conducted an impression formation experiment, in which participants were shown a sequence of eight trait adjectives attributed to a single fictional person. At the end of the

sequence, participants made two responses: the first was a rating of the person's personality (or likableness), and the second was a recall of the adjectives. They found a strong recency effect for recall but a primacy effect for the impression formation (see Figure 2.2). They concluded that "the formation of the impression involves a memory process which is distinct from, and not dependent on, the immediate verbal memory for the adjective just heard" (page 386).

Figure 2.2. Evidence for the independence of verbal memory and belief revision in impression formation. Recall curve shows strong recency over the last six serial positions. Judgment curve shows primacy. (Reproduced from N. H. Anderson, 1991, page 10).

Other researchers address the same question by selecting some variables that are known to be important in memory research and investigating their effects on belief revision. For example, Rywick and Schaye (1974) studied the effect of a secondary irrelevant task during the presentation of information and the effect of a delay after the information presentation. They found that an irrelevant task reduced primacy in both recall and belief revision, but a delay, which reduced recency of recall, had no influence on the belief order effect. This is consistent with Anderson and Hubert's independence conclusion. Hogarth and Einhorn (1992) summarized previous research and suggested that a long list of evidence sequence tends to induce primacy effects while a short list tends to induce recency effects. They argued that as more information is processed across time, people become less sensitive to the incoming information. Hovland (1957) found that primacy is associated with unfamiliar but not familiar materials. He argued that this is because that the familiar evidence that "has already had ample opportunity to form an opinion should be much less subject to persuasive influence" (p139). In contrast, Hogarth and Einhorn (1992) suggested that recency is associated with more complex (and unfamiliar) tasks. Miller and Campbell (1959) studied the effect of the time interval between evidence items. They found that long time intervals are more likely to induce recency, which they concluded was a support to the application of the Ebbinghaus curves of forgetting to the studies of persuasion.

Consistent vs inconsistent evidence. Whether the evidence sequence consists of consistent (i.e., all positive or all negative) or inconsistent (i.e., mixed positive and negative) evidence also makes a difference in order effects. Hovland (1957) found that

“when contradictory information is presented..., there is a pronounced tendency for those items presented first to dominate the impression received” (page 133). In contrast, Hogarth and Einhorn (1992) suggested that in evaluation tasks consistent evidence produces no order effects at all, whereas inconsistent evidence produces recency effects.

Frequency learning and order effects. It has been argued that human beings are more adapted to use frequency information but not single-event probabilities, especially when this information is acquired through experience instead of being instructed (e.g., Gigerenzer, 1994; Gigerenzer & Hoffrage, 1995). As reviewed previously, for example, by using the frequency format, it has been shown that many cognitive biases, such as base-rate neglect, conjunction bias, and overconfidence, have disappeared. Zhang, Johnson and Wang (1997) studied the relationship between frequency learning and order effects to see if learning frequency information can make the order effect disappear. Their conclusion is that although people can implicitly learn frequency information through experience, it can not fully eliminate the order effect.

2.3.3.2. Theories

Given that the order effect in human belief revision is such a robust finding, many theories have been proposed to explain how it occurs. In the following, I briefly describe the major theories. It is worth pointing out that some of the theories are barely complete in the sense that they can only explain part of the order effect (e.g., primacy but not recency, or recency but not primacy).

Normative theory. The normative theory here mainly refers to the Bayes Theorem. As mentioned before, the normative theory has no room for the order effect since it simply violates commutativity. Therefore, the order effect is something that should not happen and thus should be called a human bias.

Direction setup. Asch (1946) proposed an explanation for the primacy effect found in his now classic study. This explanation adopted some concepts from Piaget's constructionism. He argued that initial evidence items set up (or construct) a directed impression and later items only cause assimilative changes. More specifically, he said:

When the subject hears the first item, a broad, uncrystallized but directed impression is born. The next characteristic comes not as a separate item, but is related to the established direction. Quickly the view formed acquires a certain stability, so that later characteristics are fitted ... to the given condition.

Memory decay. Another explanation, by linking belief revision and memory, argues that order effects in belief revision represent order effects in memory (e.g., Miller & Campbell, 1959). Due to memory decay, previous evidence items get weighted less as time goes. This view can easily explain the recency effect but has difficulties for the primacy effect. This view is problematic since direct comparisons showed that beliefs were largely independent of recall of the evidence items (e.g., Dreben, Fiske, & Hastie, 1979; Risky, 1979; Surber, 1982; Anderson & Hubert, 1963). It has also been argued that recency is an effect of judgment processes but not recall processes (e.g., Anderson & Hubert, 1963; Schlottmann & Anderson, 1995).

Serial Integration Model. The serial integration model was proposed by N. H. Anderson et al. (Anderson & Hubert, 1963; Schlottmann & Anderson, 1995) in the

framework of information integration theory (Anderson, 1981, 1991). Compared to the theories above, the serial integration model provides a more thorough account for the order effect in human belief revision.

According to this theory, a response is produced through a serial integration of the information from stimulus elements in a weighted-sum form. In many cases, this weighted-sum takes the form of a running average to reduce memory load. Quantitatively:

$$r_i = w_1s_1 + w_2s_2 + \dots + w_is_i = (1 - w_i)r_{i-1} + w_is_i \quad (2.15)$$

where r denotes the judgment, w denotes the weight, which is a measure of the influence of the corresponding evidence, s denotes the evidence value, and i indexes the serial position. It is claimed that when the weights assigned to different positions are different, the order effect occurs. More specifically, primacy occurs when the weights of the earlier items are larger than the weights of the later items. Recency occurs when the later items are weighted more heavily than the earlier items.

Since the weights are not observable and the only available data are the final judgments, how to determine a series of weights from a single final judgment poses a problem. This problem is solved by carefully designing experimental conditions to match scale values across serial positions. Then each serial weight is simply proportional to the judgment difference at the position. When there are only two possible levels of evidence at each position (e.g., + and -), this can be easily seen when the above equation is transformed as follows:

$$r_{+i} - r_{-i} = w_i(s_+ - s_-) \quad (2.16)$$

where $r_{+,i}$ is the judgment at position i when the i^{th} evidence is +, and $r_{-,i}$ is the judgment when the i^{th} evidence is -. If the scale matches (i.e., $s_+ - s_-$ is a constant) then w_i is simply proportional to the judgment difference at position i . It is important to note the judgment curve in Figure 2.2 is derived in this manner.

How can the weights being assigned to different positions differ from each other? The model adopts an attention decrement hypothesis. In other words, an attention decrement process makes people pay less attention to successive items of evidence. Note that this model easily explains the primacy effect. It is argued that this attention decrement reduces the weight assigned to later items, which in turn produces the primacy effect.

Although it is argued that the recency effect is also a specification of this attention decrement process, the explanation is not straightforward. According to this explanation, human belief revision has a two-component structure: a surface component and a basal component (e.g., Anderson & Farkas, 1973; Schlottmann & Anderson, 1995). The surface component represents a short-term reaction (i.e., the recency effect) to the evidence just presented, and reflects the potential to counteract the effect of the past experience when the world changes. Only a small part of this immediate reaction is incorporated into the cumulative belief, with the rest disappearing as new information is received. The basal component, in contrast, represents a long-term and stable effect, the small part of the immediate reaction that lasts for a longer time and is incorporated into the cumulative belief. In general, due to the attention decrement and the fact that prior belief is based on increased amounts of information as the sequence grows longer, there

is an increase in the weight of the prior belief and a decrease in the weight of new evidence. If this decrease is just right, no order effect will be observed. A faster decrease will yield primacy and a slower decrease will yield recency. However, the theory does not specify what factors affect this decrease, thus leaving the theory in a non-testable status.

Anchoring and adjustment model. Based on the anchoring and adjustment heuristic first proposed by Tversky and Kahneman (1974), Hogarth and Einhorn (1992) suggested an anchoring and adjustment model to explain order effects. They argue that the observed complexity of behavior, such as order effects, can arise through the interaction of simple psychological processes with the demands of an infinitely varied environment. According to this model, belief revision is a sequential anchoring-and-adjustment process in which people adjust the current belief (the anchor) on the basis of how strongly new information confirms or disconfirms this belief. Quantitatively,

$$r_i = r_{i-1} + w_i (s_i - R) \quad (2.17)$$

where r_i is the degree of belief after the i^{th} evidence, r_{i-1} is the anchor, s_i is the information value of the i^{th} evidence, w_i is the adjustment weight for the i^{th} evidence, and R is the reference point, which can be either 0 or r_{i-1} . It is interesting to note that when R is r_{i-1} , which happens in the estimation mode, this model has the same form as that of the serial integration model.

However, different from the serial integration model, which claims that w is a function of the serial position and thus can be estimated from performance data, the anchoring and adjustment model assumes that the adjustment weight is a function of both

the anchor and the new evidence, but not the serial position. More specifically, when the impact of the new evidence is smaller than the reference point, the adjustment weight is proportional to the anchor. And when the impact of the evidence is larger than the reference point, the adjustment weight is inversely proportional to the anchor. In algebraic terms, it can be expressed as follows:

$$w_i = \begin{cases} \alpha r_{i-1} & \text{when } s_i \leq R \\ \beta(1 - r_{i-1}) & \text{when } s_i > R \end{cases} \quad (2.18)$$

where α and β are constants.

It is important to note that this way of determining the adjustment weights has a significant impact on how the order effect might occur. Consider the estimation case where R is r_{i-1} (i.e., the anchor). When the new evidence value is smaller than the anchor, which means we have negative evidence, the model suggests that the adjustment weight be proportional to the anchor. Therefore, when the anchor is large (i.e., we have a high degree of belief already), the weight is also large, which indicates that the current belief will be reduced a great deal. In contrast, when the anchor is small (i.e., the current belief is already weak), the weight is also small, which indicates that the current belief will not be reduced much. Similar adjustment occurs when the new evidence value is larger than the anchor, though the direction of adjustment is reversed.

It should be clear that it is this kind of contrast effect that results in the order effect. For example, the model predicts that a belief revision in a SbS mode for mixed evidence (some positive and some negative) will produce a recency effect. This is because earlier positive (negative) evidence makes the anchor high (low), and the later negative

(positive) evidence, due to the contrast effect, will have larger weights and thus dramatically counteract earlier belief. However, for similar reasons, the model predicts that a belief revision in a SbS mode for consistent evidence (all positive or all negative) will result in a primacy effect.

Summary. Viewed from a broader angle all these theories fall into the same category. That is, they all view human belief revision as a serial process of incorporating new evidence. Therefore, the primary focus is always on how to assess the impact of new evidence. This impact is usually represented in terms of different weights assigned to each new evidence item. For some theories, this difference in weights is due to memory decay (mainly for recency), and for others, this difference is due to attention decrement (mainly for primacy). However, the whole picture does not get any clearer because none of them can specify at what time which factor is important. In fact, it is more likely that it is the combination of all of these factors, but not any single one, that determine the impact of any new evidence.

2.3.4. Summary

While both philosophical and formal quantitative investigations provide valuable results about how people should manage belief, psychological studies focus on how people actually manage belief.

A large number of studies have shown that people in general do not follow normative principles when making judgments under uncertainty. Besides others, base rate fallacy, conjunction fallacy, and overconfidence bias all suggest that descriptive human behaviors

often deviate from normative prescriptions. It has been argued that these biases indicate that people rely on non-statistical heuristics due to the lack of correct statistical programs.

While it is natural to further argue that these biases imply human irrationality, many researchers disagree. Systematic investigations have been conducted to demonstrate when these biases appear and disappear. In addition, it has long been argued that people can satisfice. That is, instead of trying to find the best solution, which is often costly, people may settle for a good enough solution with a reasonable cost. The rational essence of seeking a balance between gain and cost is certainly not captured by normative postulates. This view is often called bounded rationality and is now generally accepted.

The order effect is a robust finding in human belief revision. It refers to the tendency to overweight irrelevant information, the temporal order of evidence presentation, when doing belief revision over time. More specifically, when earlier evidence is overemphasized, we have a primacy effect; and when later evidence is overemphasized, we have a recency effect. Normative principles, such as Bayes' Theorem, have no room for the order effect. But labeling it as a bias provides no help to understand how and why it occurs.

Previous studies on order effects reveal largely controversial results. On the one hand, different views exist on what order effects really mean. Is it a simple position effect due to differential weight assignments, or must it be assessed by comparing two content-identical but order-reversed evidence sequences? On the other hand, many different theories are proposed, among which the serial integration theory proposed by Anderson and his colleagues (1981, Anderson & Hubert, 1963; Schlottmann & Anderson,

1995) and the anchoring and adjustment model proposed by Hogarth and Einhorn (1992) are comprehensive and influential. Unfortunately, the two theories do not quite agree with each other in many aspects (see N. H. Anderson, 1991; Schlottmann & Anderson, 1995).

2.4. General Summary

Human belief is surely a fascinating thing. Belief shows itself everywhere, in both everyday life and scientific discovery. Because of belief, people can make judgments and take actions without truth-proved knowledge. Also because of belief, people can survive uncertainty and embrace the true meaning of life.

Philosophical inquiries of human belief make one important contribution. The coherence approach to belief justification suggests that belief can survive without solid foundational evidence. As long as one's beliefs are consistent and coherent with each other, they are justified. As a result, when a belief needs to be revised, the coherence approach commits to making minimal change to the current belief system while preserving coherence.

Quantitatively, belief is often summarized as a single probability number. However, the distinction between uncertainty and imprecision and the notion of ignorance suggest that belief has a complex multi-component structure, which is difficult to capture by probability. The theory of belief functions uses two numbers, belief and plausibility, to simultaneously express the degree of belief and ignorance.

Empirically, a large body of evidence has shown that people do not follow normative postulates when managing beliefs. Compared to the predictions from normative postulates, people often show biases. However, the view of bounded rationality suggests that people are statistically adaptive and follow some cost-effective principles that are not captured by normative postulates.

The order effect in human belief revision is a robust empirical finding. However, there is no single theoretical explanation that is universally accepted. Why and how the order effect occurs remains a mystery.

It is the aim of this dissertation research to develop a new approach, which emphasizes the coherence foundation, the multi-component structure, and the adaptive characteristic of human belief, to address these questions.

CHAPTER 3

AN EXPERIMENTAL STUDY

To recapitulate, literature has revealed three fundamental findings about human belief and human belief revision. First, human belief has a coherence foundation. Not only can a belief that has no sound reasons be justified, but also can a belief that has strong evidence be completely destroyed. New positive evidence does not necessarily reinforce a belief, and new negative evidence does not necessarily discredit a belief. Beliefs hold each other as a coherent system. Second, human belief has a multi-component structure. The uncertainty/imprecision distinction indicates that uncertainty has many faces, as does human belief. The notion of ignorance questions the intuitive one-dimensional probability summarization. A single probability number simply cannot capture all the important aspects of a belief. Finally, human belief has an adaptive characteristic. Beliefs are revised to accommodate new evidence. More importantly, in the long run beliefs are constantly shaped and tuned to reflect the statistic structure of the environment.

The order effect is a robust finding in human belief revision. It is of great importance to investigate how and why the order effect occurs in the framework established by these rich and sophisticated findings. Unfortunately, previous approaches to the order effect hardly take these findings into consideration. Those researchers intuitively feel that the order effect might be accounted for in a straightforward way by a weight assignment mechanism that weighs members of the evidence sequence differentially. They propose formula to quantitatively show how this can be done. However, they encounter great difficulties in fully explaining why weights have to be assigned in a particular way at a particular time. Consequently, they either propose one or more factors that they think are particularly important to determine the weight assignment (e.g., memory decay, or attention decrement, or simply the serial position itself), which of course often only accounts for a fragment of the order effect; or, they simply adopt arbitrary parameters in the weight assignment (e.g., α and β in Hogarth and Einhorn's (1992) Anchoring and Adjustment Model). Obviously, no theory is generally accepted.

The goal of this dissertation research is to demonstrate how the order effect might be more naturally explained on the basis of these established findings, both empirically (Chapter 3) and computationally (Chapter 4). In doing so, I hope to convey the message that the order effect in belief revision provides additional evidence for bounded human rationality.

This chapter is organized as follows. In section 3.1, I give a brief overview of what these fundamental findings imply about the order effect in human belief revision. In addition, I describe the specific hypotheses this empirical study is interested in. In

section 3.2, I present the experimental design and procedure. The experimental results are contained in section 3.3. And I give a discussion in section 3.4.

3.1. Overview

The fundamental problem in human belief revision is that one accommodates the new evidence in the context of current beliefs and settles in a new balance in the interaction between the beliefs and the evidence. Previous approaches often emphasize the evidence side of the problem and attempt to assess the impact of evidence in terms of external task or stimulus characteristics. However, it appears that belief itself and the interaction between belief and evidence may be more essential in the belief revision process, as implied by those important findings.

3.1.1. The Multi-Component Structure

The fundamental problem of human belief revision can be easily rephrased as follows: it is a problem to assess how easily current beliefs can be revised in the light of new information. From this perspective, the impact of the new evidence cannot be fully understood merely by external task characteristics. In order for the new evidence to make any impact, the nature of the current beliefs must be sufficiently appreciated. In other words, the current beliefs are a necessary condition for the new evidence to make any impact. Depending on the nature of the current beliefs, same evidence may have no effect, a small effect, or a large effect.

Let us consider the following example. After examining a patient with certain symptoms, both an experienced doctor and an entry-level medical school student believe that there is a 90 percent chance that the patient has disease *X*. Later a new medical test result comes out. After learning the new evidence, it turns out that the experienced doctor still believes that the patient has a high chance of having disease *X*, while the medical school student changes his opinion and suggests another disease.

What makes the difference? The multi-component structure of human belief implies an important probability/confidence distinction in describing the nature of a belief. And this example demonstrates it. Even though both the doctor and the medical school student have the same probability judgment (both are 90%), since the doctor is more experienced than the student, his judgment of the original disease is much more stable than that of the student. In other words, the student is less confident of his judgment than the experienced doctor.

Belief is never simple. The probability/confidence distinction plays a fundamental role in human belief revision. While the probability component of belief expresses the relative degree of belief, which is often the sole meaning in the usual use of the term “belief”, it is the confidence component of a belief that represents how easily the belief can be revised and altered. The confidence component of a belief is mainly determined by, among others, the amount of previous experience. A belief with no previous experience (e.g., total ignorance, established by the principle of indifference) has zero confidence and is easiest to alter. An example is that you believe, due to the principle of indifference, there is 50% chance a tossed coin will land heads up on the next toss when

you have no experience at all about the fairness of the coin. Now suppose you toss the coin 3 times and you get 3 heads in a row. Then you may begin thinking that this is an unbalanced coin and the probability you get a head next time should be greater than 0.5. On the other hand, a belief established by significant previous experience is committed with a confidence level and thus is hard to change. After tossing the coin 1000 times and observing 500 heads and 500 tails, you would also believe there is a 50% chance to get a head next time. In this case, however, 3 heads in a row will hardly make you to suspect the fairness of the coin. This belief has a much higher confidence and is difficult to alter.

It is interesting to note the functional role the parameters α and β play in Hogarth and Einhorn's (1992) Anchoring and Adjustment model. As mentioned in last chapter, there are two constants that help to determine the weight assignment of new evidence. According to the model, α and β ($0 \leq \alpha, \beta \leq 1$) represent subjects' sensitivity toward "negative" and "positive" evidence, respectively. "Small values of α and β imply low sensitivity to new information while large values of α and β imply high sensitivity" (page 15). It is easy to notice that parameters α and β are functionally closely related to the confidence component of the current beliefs in the sense that as the confidence goes up, the sensitivity to the new evidence might go down. However, the model does not realize that this sensitivity to new evidence is an essential property of the current beliefs and can be determined by examining the experience the beliefs are based on. Instead, the model argues that some individuals tend to view the negative evidence more seriously than others, and people sometimes view the negative evidence seriously and sometimes do

not. Therefore, the model claims that the two parameters actually represent some unidentifiable individual or context differences.

3.1.2. The Adaptive Characteristic

The adaptive characteristic of human belief implies that a belief is never static. It is constantly tuned to reflect the statistical structure of the environment. Tuning enables previous incorrect or inaccurate beliefs to be corrected. And tuning enables people to learn from experience and improve their confidence in dealing with future problems.

As stated in the last chapter, Bayesian learning is a sophisticated formal technique for learning statistical regularities based on observed data. It is interesting to note that Bayesian learning has been used to model human cognition. For example, the ACT-R theory, a successful cognitive architecture, extensively adopts Bayesian learning as one of its core learning mechanisms. Not only fundamental probabilities (e.g., the probability that a production rule succeeds) but also strengths of associations (e.g., associative strength between declarative chunks) can be easily learned from experience through the Bayesian learning technique.

Decades of empirical studies of human and animal learning has revealed yet another fundamental learning principle named the Rescorla-Wagner rule (Rescorla & Wagner, 1972; for a review, see Miller, Barnet, & Grahame, 1995). This rule regulates when the expectation differs from the actual occurrence how the underlying associative model responsible for the expectation should be tuned to become statistically more realistic. Gluck and Bower (1988) successfully showed how the Rescorla-Wagner rule can be

applied to enable a connectionist model to learn statistical causal relations from observations.

3.1.3. Summary

Human belief is coherent, complex, and adaptive. What does it tell us about the order effect in belief revision?

Interestingly, it implies that the order effect pattern may change with different levels of experience. The rationale is as follows. As one keeps interacting with the environment, one gains more and more experience. As a result, beliefs are gradually tuned to the statistical structure of the environment. And confidence increases as more experience is gained. Both factors will make one react to any new evidence more realistically rather than over-react or under-react. Since over-reacting and under-reacting are the fundamental causes of the order effect, this is to say that when one gains more and more experience about the environment, the order effect in belief revision tends to diminish and disappear.

The experiment presented next is designed to test this prediction.

3.2. Experiment

3.2.1. Task

A modified version of the CIC (Combat Information Center, Towne, 1995; see also Wang, Johnson, Zhang, 1998; Zhang, Johnson, Wang, 1997) computer simulator was used as the task domain. In a CIC task, the goal of the participant, acting as a

commanding officer of a naval ship, was to maintain the safety of his/her own ship by collecting information about various aircrafts in the radar area and accurately identifying their intentions. This is a belief revision task because valuable information related to the identification of an aircraft usually arrives sequentially, and the ability to properly incorporate the new information is essential to achieve the goal.

The CIC simulator used in the experiment is shown in Figure 3.1 in black and white. The simulator consists of a radar screen, which is the big dark region located at the top left part of the simulator. The small circle at the center of the radar indicates the position where the naval ship itself is located. The square with an arrow extended from it, which is yellow and can appear anywhere on the radar, represents the unknown aircraft. In this experiment, there is only one unknown aircraft in each scenario. The white rectangular region at the right side of the simulator is the message board, through which the participant can learn the new evidence. The buttons below the message board are used by the participant to take actions, including beginning a new scenario ("Next Game"), collecting information ("Check Route" and "Send Radio"), making decisions ("F" and "H", for friendly and hostile, respectively), or checking the true identity of the target ("True Identity"). Participants use the two scales at the lower left corner of the simulator to indicate their belief and confidence judgments.

Welcome to the Experiment

Hello
Captain!
Block#: 1
Game#: 1

Next Game

Check Route

Send Radio

How likely do you think it is that the target is friendly?
0 10 20 30 40 50 60 70 80 90 100

How confident are you of your judgment?
0 10 20 30 40 50 60 70 80 90 100

Decision:
F H

True Identity

Figure 3.1. The version of CIC used as the task domain.

In each scenario, the participant saw an unknown aircraft shown on the radar and tries to identify it. The aircraft was randomly positioned in the region within the “100” circle and the “200” circle. However, the heading of the aircraft, indicated by the arrow extended from the aircraft symbol, was always pointing to the ship. The participant could collect two pieces of information sequentially. One piece of information was about the route (R). The participant could check if the target was on a commercial air route, and the possible answer was either on or off. The other piece of information was self-identification (SelfID). That is, the participant could send radio warning to the target and ask for identification. The target either replied with a friendly identification (i.e., “This is a commercial airplane”), or simply did not respond. After obtaining both pieces of evidence, the participant was forced to make a two-alternative (i.e., friendly or hostile) judgment about the identity of the target. After the decision was made, the participant could request the true identity of the target if available.

A special arrangement of this experiment was that the participant was constantly required to answer two questions. The first question was “how likely do you think it is that the target is friendly?”, which clearly concerned the current degree of belief. This question was to be answered on a 0-to-100 scale, with 0 indicating impossible and 100 indicating completely sure. The second question was “How confident are you of your judgment?”, which concerned the confidence component of the current belief. Similarly, the participant answered this question on a 0-to-100 scale, with 0 for not confident at all and 100 for completely confident.

In each scenario the participant needed to answer the two questions three times. They were distributed as follows: 1) before any information was collected, 2) after the first piece of evidence was learned, and 3) after the second piece of evidence was learned but before the final 2-alternative judgment was made. Except for the first time (i.e., initial reporting), in which the scale was blank and the slider could not be seen (i.e., no anchor was visible), in the second and third times, the previous judgments were visible and the participant could do the reporting by simply moving the sliders to new locations.

In each scenario, the order of the information collection and the participant's actions were pre-determined. For example, one possible training scenario proceeded as follows: new game – belief judgment – confidence judgment – check route – belief judgment – confidence judgment – send radio – belief judgment – confidence judgment – decision – true identity. The participant was forced to follow this order. The order was enforced by colors. At any point of time, there was only one button (including the scales) that was activated and thus available. It had the red color. All other buttons were in gray color and thus unavailable. The participant could not proceed until the activated button had been taken care of. After that, the next available button became red and was activated.

3.2.2. Design

There were two types of scenarios (or trials) in the experiment. One was the standard training scenario. The other was the evaluation scenario. The main difference between the two was that in the evaluation scenario no true identity information was available, and thus the participant would not get feedback at the end. As shown later, since the

evaluation trials were located in critical positions during the training process, they provided a convenient means to easily evaluate the pattern changes of order effects.

The experiment adopted a 3x(4) factorial design. The between-subject independent variable was the ratio of total friendly targets to total hostile targets in the training samples. The ratio was 1:1 (i.e., equal number of friendly and hostile targets), 3:1 (friendly targets are three times as frequent as hostile targets), or 1:3 (hostile targets are three times as frequent as friendly targets). The purpose of this factor was to create environments with different statistical structures and test if participants could gradually tune their beliefs to capture these structures.

The experiment attempted to investigate how the patterns of order effects changed with training. The training was organized in blocks, which is the within-subject variable. Each participant performed four such blocks, with a two-minute break between consecutive blocks (see Figure 3.2).

Every block consisted of a training sub-block and an evaluation sub-block. Every evaluation sub-block had eight evaluation trials in it. The eight evaluation trials were constructed in the following way. There were two pieces of evidence (Route and SelfID), each of which had two possible values (“on” and “off” for Route, and “friendly response” and “no response” for SelfID), so there were 4 kinds of trials; and each piece of evidence could be collected first (this was different from the training trial, in which the evidence collection order was a between-subject variable), thus we had a total of eight trials. Participants were instructed to summarize their training experience in order to perform these evaluation trials (see Appendix A for instructions).

Each training sub-block consisted of 36 trials. The trial distribution is dependent on the friendly-hostile ratio and is shown in Table 3.1. Since a value of “on” for Route and a value of “friendly response” for SelfID are regarded as positive evidence for a friendly target, they are represented by “+”s in Table 3.1. Similarly the opposite values are represented by “-”s. The first row of the table indicates that when both pieces of evidence are positive, there are eight friendly targets and two hostile targets in the 1:1 condition; and there are twelve friendly targets and only one hostile target in the 3:1 condition; and there are four friendly targets and three hostile targets in the 1:3 condition. Such a trial distribution allows Route and SelfID information to be conditionally independent given the identity of the target. The details of conditional independence are provided in Appendix B.

Route	SelfID	1:1		3:1		1:3	
		friendly	hostile	Friendly	hostile	friendly	Hostile
+	+	8	2	12	1	4	3
+	-	4	4	6	2	2	6
-	+	4	4	6	2	2	6
-	-	2	8	3	4	1	12
Total		18	18	27	9	9	27

Table 3.1. The trial distribution used in the experiment.

Given the current training trial distribution, the theoretical base rates and conditional probabilities given evidence can be easily calculated. By theoretical base rates, I mean the built-in (due to this specific design) probabilities of friendly (or hostile) when no evidence is available. By theoretical conditional probabilities, I mean the built-in (due to this specific design) conditional probabilities of friendly (or hostile) given evidence. They are shown in Table 3.2.

	Friendly-Hostile Ratio		
	1:1	3:1	1:3
P(Friendly)	.50	.75	.25
P(Friendly +)	.667	.857	.40
P(Friendly -)	.333	.60	.143
P(Friendly ++)	.80	.923	.571
P(Friendly --)	.20	.429	.077
P(Friendly +-)	.50	.75	.25

Table 3.2. Theoretical base rates and conditional probabilities in the current design. P(Friendly) indicates the base rate. P(Friendly|+) is the conditional probability that the target is friendly given any one of positive evidence. P(Friendly|++) is the conditional probability that the target is friendly given that both evidence items are positive. The base rates and conditional probabilities for hostile identity are not shown but can be easily calculated (i.e., the corresponding complementary of each cell).

Each block thus had 44 trials (36 training trials and 8 evaluation trials). However, in order to get a measure of a participant's initial performance before any training, an additional evaluation sub-block was inserted before the training sub-block in the first block (see Appendix A for instructions), which made the first block to have 52 trials (8 initial evaluation trials and 44 ordinary trials). Consequently, the total number of trials for each participant was 184. The experimental design is shown graphically in Figure 3.2.

Figure 3.2. The experimental design.

To control the possible effect of the evidence collection order in training (i.e., which piece of evidence was collected first in a training scenario, Route or SelfID), half of participants always collected route information first, and the other half always collected the SelfID information first. In the example above, the route information is collected before the SelfID information. The evidence collection order in training served as a control variable.

3.2.3. Participants and Procedure

The participants were 140 undergraduate students in introductory psychology courses (Psychology 100) at The Ohio State University who participated in the experiment for partial fulfillment of the course requirements. They were randomly assigned to 3 base rate treatment (i.e., friendly-to-hostile ratio) groups, with at least 40 in each group. In addition, to control the possible effect of training order, half of the participants in each group were trained in the Route-SelfID order and the other half were trained in the SelfID-Route order (see Table 3.3). The trials in each sub-block were completely randomized for each participant.

The experiment was programmed in TCL/TK (Ousterhout, 1994). PCs running Windows 95 with 80486 processors and 14-inch color monitors were used to conduct the experiment. Since there were 20 such PCs in a computer lab, 20 participants could do the experiment independently and simultaneously. Participants used mouse input devices to make responses.

In each session, the experimenter read general instructions (see Appendix A) to the participants. Then the participants began the formal experiment. Each participant worked at his/her own pace. The whole session took about 1.5 hours.

3.3. Experimental Results

It is important to note that among the 140 participants, only 108 (about 77%) participants were counted as valid and were used for the analysis (see Table 3.3). 32 participants were considered invalid mainly because of obvious lack of cooperation. Details of the invalid 32 participants are provided in Appendix C.

The experimental results are reported in two parts, one concerning the training data and the other concerning the evaluation data. The statistical tuning of belief judgments and the pattern change of order effects are the major foci of the data analyses.

It is important to note that unless specified the belief judgments are reported in terms of friendliness on a 0-100 scale. Therefore, a belief judgment of 90 indicates that one is 90% sure that the target is friendly. In addition, in the current design the order effect is measured as the difference of final judgments between the negative-then-positive evidence sequence and the positive-then-negative evidence sequence (e.g., Route+ then SelfID- vs SelfID- then Route+, and Route- then SelfID+ vs SelfID+ then Route-).

Training Order	Friendly-Hostile Ratio		
	1:1	3:1	1:3
Route-SelfID	24 (18)	23 (17)	23 (21)
SelfID-Route	24 (19)	23 (17)	23 (16)
Subtotal	48 (37)	46 (34)	46 (37)
Total	140 (108)		

Table 3.3. The number of participants in each experiment condition. (The numbers in parentheses represent the number of valid participants, see Section 3.3 for details)

3.3.1. Training Data

Each participant was trained for 144 trials, which were divided into 4 blocks. In each trial, 7 data points were collected: initial belief and confidence judgments (belief0 and confidence0), belief and confidence judgments after the first piece of evidence (belief1 and confidence1), belief and confidence judgments after the second piece of evidence (belief2 and confidence2), and the final 2-alternative decision (decision).

Initial belief judgments (belief0) reflect one's basic recognition of the base rate information of the environment. It is of interest to investigate how the initial belief judgments changed with the training. The results, separated based on different base-rate conditions, are shown in Figure 3.3. The performance improvement is evident. The mean values of belief0 for the 1:1 condition, which has a theoretical base rate of 50%, are approximately constant (49.2, 52.0, 51.4, and 52.8, from block 1 to block 4, respectively). The mean values for the 3:1 condition, which has a theoretical base rate of 75%, gradually increase (54.5, 58.2, 64.1, and 72.2, from block 1 to block 4, respectively). In contrast, the mean values for the 1:3 condition, in which the theoretical base rate is 25%, gradually decrease (45.6, 41.2, 39.5, and 36.9, from block 1 to block 4, respectively). Linear regression analyses show that in all three base rate conditions the training trial number is a good predictor of the initial belief judgment. More specifically, the slope parameter estimates for 3:1 and 1:3 conditions (0.16 and -0.07 , respectively) are both significantly different from 0 ($F(1,142)=696.4$, $p<0.001$; and $F(1,142)=447.4$, $p<0.001$; respectively). Although statistics show that the slope parameter estimate for the 1:1 base rate condition is also significantly different from 0 ($F(1,142)=56.74$, $p<0.01$), it is very small (0.02). Overall, this pattern of results shows that as more training trials were performed, the initial belief judgments were getting closer and closer to the built-in theoretical base rates, indicating that participants gradually captured the statistical invariance and thus made more realistic initial judgments.

Figure 3.3. Initial belief judgments (belief0) across the 144 training trials. The corresponding locally-weighted regression (loess) curves² are also plotted. Based on the experimental data in all three base rate conditions. The numbers in parentheses indicate the corresponding theoretical base rates of friendly targets in terms of percentage.

² Loess fits a curve to the data points *locally*, so that at any point the curve at that point depends only on the observations at that point and some specified neighboring points. Because such a fit produces an estimate of the response that is less variable than the original observed response, the result is called a *smooth*. Loess is attractive since relies on the data to specify the form of the model (see MathSoft, 1997).

Did participants make more confident initial belief judgments with the amount of training? The results of initial confidence judgments (confidence0) are shown in Figure 3.4. Different from the initial belief judgments, the initial confidence judgments gradually increase with training for all three base rate conditions (i.e., from block 1 to block 4, 53.6, 57.7, 58.0, and 60.7 for the 1:1 condition; 52.2, 61.2, 66.7, and 74.5 for the 3:1 condition; and 49.7, 55.7, 57.8, and 58.6 for the 1:3 condition). Similar linear regression analyses show all three slope parameters (0.06, 0.20, and 0.08, for 1:1, 3:1, and 1:3 base rate conditions, respectively) are significantly different from 0 ($F(1,142)=195.6$, $p<0.001$; $F(1,142)=1421$, $p<0.001$; and $F(1,142)=237.7$, $p<0.001$; respectively). This pattern of results suggests that participants did make more confident belief judgments as more experience was gained through training.

The performance changes in terms of final decisions (decision) are shown in Figure 3.5, which depicts the proportion of friendly final decisions against the 4 training blocks. The gradual tuning of the decision behavior is clear. In the first block, the proportions of friendly decisions are 0.53, 0.60, and 0.37, for the 1:1, 3:1, and 1:3 conditions, respectively. Statistics show both 0.60 and 0.37 are significantly different from the corresponding theoretical values, $t(33)=-9.98$, $p<0.01$, and $t(36)=6.08$, $p<0.01$, respectively. In the fourth block, the proportions of friendly decisions (0.54, 0.74, and 0.28, for the 1:1, 3:1, 1:3 conditions, respectively) reach the theoretical probabilities ($t(36)=1.72$, $p=0.09$; $t(33)=-0.58$, $p=0.57$; and $t(36)=1.54$, $p=0.13$, respectively). An overall ANOVA shows that the interaction between the base rate conditions and the training blocks is significant ($F(6,315)=14.1$, $p<0.01$).

Figure 3.4. Initial confidence judgments (confidence0) across the 144 training trials. The corresponding locally-weighted regression (loess) curves are also plotted. Based on the experimental data in all three base rate conditions.

Figure 3.5. Proportion of friendly decisions across the four training blocks. Based on the experimental data in all three base rate conditions.

Figure 3.5 can be further decomposed based on different evidence sequences. They are presented in Figure 3.6, which depicts the proportion of friendly final decisions against the 4 training blocks, separated for the evidence sequences “++” (i.e., both evidence items are positive), “--” (i.e., both evidence items are negative), and “+-” (i.e., one evidence item is positive and the other is negative). These graphs show the gradual tuning of the final decisions conditioned on evidence. For the evidence sequence “+-”, in the first block, the proportions of friendly decisions are 0.53, 0.54, and 0.36, for the 1:1, 3:1, and 1:3 conditions, respectively. Statistics show both 0.54 and 0.36 are significantly different from the corresponding theoretical values, $t(33)=-8.31$, $p<0.001$, and $t(36)=3.26$, $p<0.01$, respectively. In the fourth block, the proportions of friendly decisions (0.48, 0.68, and 0.30, for the 1:1, 3:1, 1:3 conditions, respectively) approach the theoretical probabilities ($t(36)=-0.70$, $p>0.48$; $t(33)=-2.15$, $p>0.03$; and $t(36)=1.87$, $p>0.05$, respectively). An overall ANOVA shows that the interaction between the base rate conditions and the training blocks is significant ($F(6,315)=11.1$, $p<0.001$). A similar pattern is found for the evidence sequence “--”. In the first block, the proportions of friendly decisions are 0.09, 0.11, and 0.09, for the 1:1, 3:1, and 1:3 conditions, respectively. Statistics show both 0.09 and 0.11 (for the 1:1 and 3:1 conditions) are significantly different from the corresponding theoretical values, $t(36)=-3.69$, $p<0.001$, and $t(33)=-11.67$, $p<0.001$, respectively. In the fourth block, the proportions of friendly decisions (0.20, 0.35, and 0.08, for the 1:1, 3:1, 1:3 conditions, respectively) approach the theoretical probabilities ($t(36)=0$, $p>0.99$; $t(33)=-1.45$, $p>0.15$; and $t(36)=0.15$, $p>0.88$, respectively). An overall ANOVA shows that the interaction between the base rate

conditions and the training blocks is significant ($F(6,315)=5.07$, $p<0.001$). Finally, a similar pattern is also shown for the evidence sequence “+-“. In the first block, the proportions of friendly decisions are 0.96, 0.95, and 0.91, for the 1:1, 3:1, and 1:3 conditions, respectively. Statistics show both 0.96 and 0.91 (for the 1:1 and 1:3 conditions) are significantly different from the corresponding theoretical values, $t(36)=-13.36$, $p<0.001$, and $t(36)=18.14$, $p<0.001$, respectively. Visual inspection shows that the proportions of friendly decisions tend to approach the theoretical probabilities with training. However, statistics show the tendency is not significant (in the fourth block, the proportions of friendly decisions are 0.90, 0.92, and 0.72, for the 1:1, 3:1, 1:3 conditions, respectively; $t(36)=3.34$, $p=0.002$; $t(33)=-0.17$, $p>0.86$; and $t(36)=3.06$, $p=0.004$, respectively). As discussed later, this pattern may suggest that participants adopted a conservative strategy in this condition. An overall ANOVA shows that the interaction between the base rate conditions and the training blocks is significant ($F(6,315)=3.10$, $p<0.01$).

Figure 3.6. Proportion of friendly decisions across the four training blocks, separated for different evidence sequences. The dashed lines indicate the theoretical probabilities (see Table 3.2). Based on the experimental data in all three base rate conditions.

3.3.2. Evaluation Data

An evaluation sub-block contains 8 evaluation trials. Each participant performed the evaluation sub-block 5 times. They are, before any training (block0), and immediately after each training sub-block (block1 to block4). As a result, evaluation trials are the main tools to evaluate and test possible changes of the belief revision patterns with training (for more details, see the experimental design).

The belief revision pattern at the very beginning (i.e., block0) is shown in Figure 3.7. The order effect is evident (41.9 for positive-then-negative evidence sequence vs 50.8 for the negative-then-positive evidence sequence). Note that the data lines for evidence sequences “+-“ and “-+” form a fish-tail like pattern. This pattern indicates a recency effect in the sense that the final judgments are more affected by the latter evidence items. Statistics show that the recency effect is significant ($t(107)=4.12$, $p<0.001$).

Figure 3.7. The belief revision pattern based on the experimental data from the initial evaluation block (i.e., block0). “init”, “e1”, and “e2” indicate the time points before any evidence, after the first evidence, and after the second evidence, respectively. “+” denotes a positive evidence, and “-” denotes a negative evidence.

The belief revision patterns during the training (i.e., block1 to block4) for the base rate condition 1:1 are shown in Figure 3.8. First of all, the average initial belief judgment (i.e., before any evidence) is largely unchanged across the whole training session (50.0, 50.0, 49.5, and 53.3, from block1 to block4, respectively, and $F(3,108)=3.67$, $p=0.015$). Second, the order effect exists only in block1 (54.0 for the -+ evidence sequence vs 44.3 for the +- evidence sequence, $t(36)=3.92$, $p<0.001$). There is no order effect from block2 to block4 (48.9, vs 46.1, $t(36)=0.78$, $p>0.43$; 51.5 vs 48.1, $t(36)=0.76$, $p>0.45$; and 55.8 vs 52.1, $t(36)=0.99$, $p>0.32$; respectively). Finally, it is interesting to note that the areas inside the fish-tail like order effect patterns tend to become systematically smaller with the training. Since the pattern is approximately vertically symmetric, the area is proportional to the height of the fish-tail, which can then be roughly estimated in the following manner. Based on any one evidence sequence that is part of the fish-tail (e.g., the one with positive-then-negative evidence, which is also used in the following analysis), one can calculate the absolute difference between the initial and second belief judgment, and the absolute difference between the second and third belief judgment. Then the sum of the two absolute differences is a rough estimate of the area of the fish-tail. Quantitatively,

$$Area=|b0-b1| + |b1-b2| \quad (3.1)$$

where $b0$ is the initial belief judgment (i.e., before any evidence), $b1$ is the belief judgment after the first piece of evidence, and $b2$ is the belief judgment after the second piece of evidence. The estimated areas against the 4 blocks are plotted in Figure 3.9. The tendency to decrease is statistically significant (46.0, 39.0, 34.4, and 14.4, from

block1 to block4, respectively, $F(3, 108)=21.01$, $p<0.001$). This pattern indicates that participants fluctuated less in their belief judgments as more experience was gained, which further suggests that participants tended to be less sensitive to new evidence as confidence goes up.

The belief revision patterns for the base rate condition 3:1 are shown in Figure 3.10. First, different from the 1:1 condition, the average initial belief judgment (i.e., before any evidence) tends to increase with training (56.7, 59.0, 67.1, and 68.6, from block1 to block4, respectively). Statistics show that the difference is significant ($F(3, 99)=13.02$, $p<0.001$). Second, the order effect exists only in block1 (64.2 for the -+ evidence sequence vs 46.5 for the +- evidence sequence, $t(33)=7.66$, $p<0.001$). There is no order effect from block2 to block4 (63.1 vs 64.2, $t(33)=0.35$, $p>0.72$; 69.7 vs 72.7, $t(33)=0.85$, $p>0.40$; and 67.2 vs 65.9, $t(33)=0.49$, $p>0.62$; respectively). Finally, a similar pattern of decreasing areas inside the fish-tails was found. The same procedure mentioned above was used to estimate the areas, and the results are plotted in Figure 3.9. The tendency to decrease is statistically significant (47.3, 36.7, 23.8, and 26.7, from block1 to block4, respectively, $F(3, 99)=10.73$, $p<0.001$).

Figure 3.8. The belief revision patterns in the 1:1 base rate condition. Based on the experimental data from the four evaluation blocks (block1 to block4).

Figure 3.9. The area estimates across the four training blocks. Based on the experimental data from the four evaluation blocks (block1 to block4).

Figure 3.10. The belief revision patterns in the 3:1 base rate condition. Based on the experimental data from the four evaluation blocks (block1 to block4).

The belief revision patterns for the base rate condition 1:3 are shown in Figure 3.11. First, different from previous conditions, the average initial belief judgment (i.e., before any evidence) tends to decrease with training (47.8, 43.0, 41.3, and 40.2, from block1 to block4, respectively). Statistics show the decrement is significant ($F(3, 108)=14.51$, $p<0.001$). Second, the order effect exists only in block1 (49.7 for the -+ evidence sequence vs 43.3 for the +- evidence sequence, $t(36)=2.93$, $p<0.01$). There is no order effect from block2 to block4 (37.2 vs 37.1, $t(36)=-0.04$, $p>0.96$; 38.5 vs 34.1, $t(36)=1.25$, $p>0.21$; and 31.0 vs 31.6, $t(36)=0.20$, $p>0.84$; respectively). Finally, visual inspections indicate a similar pattern of decreasing areas inside the fish-tails. The same procedure mentioned above was used to estimate the areas, and the results are plotted in Figure 3.9. However, the tendency to decrease is not statistically significant in this case (34.0, 31.9, 33.2, and 26.4, from block1 to block4, respectively, $F(3,108)=1.27$, $p=0.288$).

To better compare the belief revision pattern changes in the three different base rate conditions, data from the two critical evidence sequences (+- and -+) in all three base rate conditions are combined in Figure 3.12. Both the statistical tuning and the disappearance of order effects are evident. The three sequences of belief revision patterns from the three base rate conditions tend to separate from each other as training progresses, approaching the corresponding theoretical probabilities. In addition, in all three base rate conditions, from block0 to block4, the areas inside the fish-tail patterns are getting smaller and order effect tends to disappear.

Figure 3.11. The belief revision patterns in the 1:3 base rate condition. Based on the experimental data from the four evaluation blocks (block1 to block4).

Figure 3.12. The belief revision patterns in all three base rate conditions. Based on the experimental data from all five evaluation blocks (block0 to block4). Only data from inconsistent evidence sequences (i.e., -+ and +-) are showed.

Figure 3.13 summarizes the belief judgments from the first and the last evaluation blocks and compares them with the built-in theoretical values. It is easy to identify that in general the belief judgments, both the initial unconditional beliefs and the conditional beliefs, gradually approach the theoretical values as the training continues.

Although the results presented above imply the potential role the confidence component plays, can it be explicitly accessed? The confidence judgments in all three base rate conditions are shown in Figure 3.14. First of all, it is important to note that the confidence judgments tend to increase as the training progresses. This can be seen clearly from the changes of average initial confidence judgments (i.e., before any evidence) across the 4 blocks. They are, from block1 to block4, 54.8, 56.5, 57.8, and 61.4, for the 1:1 base rate condition; 56.3, 62.2, 70.6, and 71.3, for the 3:1 base rate condition; and 53.0, 58.1, 58.6, and 61.4, for the 1:3 base rate condition. Statistics show that the confidence increments across the 4 blocks are significant in all three base rate conditions ($F(3,108)=2.89$, $p<0.05$; $F(3,99)=10.81$, $p<0.001$; $F(4,108)=6.50$, $p<0.001$; respectively). In addition, it is interesting to note that the confidence revision patterns are very different from the belief revision patterns discussed above. When the evidence items are consistent (e.g., “++” or “--”) the confidence judgment increases as more evidence items are collected (57.2, 70.0, and 85.7 for the 1:1 base rate condition; 64.9, 73.5, and 85.9 for the 3:1 base rate condition; and 58.1, 65.8, and 79.7 for the 1:3 base rate condition). Statistics show all the increments are significant ($F(2,72)=65.08$, $p<0.001$; $F(2,66)=44.31$, $p<0.001$; $F(2, 72)=62.40$, $p<0.001$; respectively). When inconsistent evidence items are collected (e.g., “+-” or “-+”), it seems they cancel each

other out and the confidence judgment stays constant (58.0, 70.5, and 71.2 for the 1:1 base rate condition; 65.4, 73.7, and 74.3 for the 3:1 base rate condition; and 57.5, 66.5, and 65.1 for the 1:3 base rate condition). Statistics show that the last two confidence judgments in each base rate conditions are not significantly different ($F(1,36)=0.17$, $p>0.68$; $F(1,33)=0.23$, $p>0.63$; and $F(1,36)=0.85$, $p>0.36$; respectively). The contrast between confidence revision patterns and belief revision patterns (i.e., belief decreases versus confidence increases in the “--“ evidence condition; belief changes dramatically versus confidence keeps unchanged after the second piece of evidence in the “+--“ and “--+“ evidence conditions) suggests that participants could distinguish between belief and confidence.

3.4. Discussion

In summary, the experiment was designed to test the belief revision pattern changes as more experience with the environment is acquired through interactive and extensive training. The results are quite promising. First of all, belief indeed adapts, as predicted. As participants performed more training trials, the belief judgments were gradually tuned to approach the theoretical probabilities. Not only did the unconditional initial belief judgments approach the theoretical base rates, a result that is consistent with many previous base rate learning findings, but also various conditional belief judgments gradually approximated the theoretical values. It is worth pointing out, however, that the belief judgments did not reach the theoretical values even in the fourth evaluation block, suggesting an overall conservative strategy.

Figure 3.13. From block1 to block4, belief judgments approach the built-in theoretical values. Based on the experimental data from the first and the fourth evaluation blocks (block1 and block4) in all three base rate conditions.

Figure 3.14. The confidence revision patterns in all three base rate conditions. Based on the experimental data from the four evaluation blocks (block1 to block4).

In addition, significant recency effects are found in the initial evaluation block before any training and in the evaluation block after the first block of training. As reviewed in chapter 2, a recency effect is a typical finding in such tasks with short lists of evidence and a step-by-step evaluation mode (see Hogarth & Einhorn, 1992; Zhang, Johnson, Wang, 1997).

The most important finding of this experiment, however, is the disappearance of the recency effect as more training trials were performed. The disappearance of the recency effect suggests that instead of over-reacting in the light of new evidence, participants made more proper and more realistic reactions. Proper reactions in the light of new evidence cannot come from any possible change of evidence since the implication of each evidence item in terms of the probabilistic prediction power is invariant during the whole training session. As suggested previously, one plausible explanation is that more experience was acquired during training, and the statistical tuning led participants to make more confident belief judgments, which eliminated over-reaction.

The hypothesis that confidence plays an important role in the disappearance of the recency effect has also been supported by another novel finding of the experiment. That is, as shown in Figure 3.12, if we look at the typical fish-tail patterns formed by the belief judgments in the “+–” and “–+” evidence sequences, it is easy to find that the areas inside the fish-tail patterns tend to get smaller across the four evaluation blocks. This result indicates that participants became less swayed by the new evidence and belief judgments tended to fluctuate less as more training trials were performed, which again implies the critical role confidence plays. As confidence goes up, over-reaction is avoided.

Whether participants can explicitly and accurately assess confidence is an interesting issue. The results shown in Figure 3.14 suggest that participants could, but only to certain degree. First of all, the contrast between the belief revision patterns and the confidence revision patterns implies that participant can indeed distinguish confidence from belief. The gradual increment of reported confidence along the whole training session is consistent with the prediction that confidence goes up as more experience is gained. However, the micro-pattern of confidence revision within a single trial (i.e., patterns of in Figure 3.14) raises a question. More specifically, it indicates that confidence increases with the number of collected evidence items when the evidence sequence is consistent, and confidence stays constant when the evidence sequence consists of interleaved positive and negative evidence items. Therefore, it implies that confidence is a function of both the amount of experience (in terms of the number of trials performed previously) and the type of received evidence items within a single trial. This is inconsistent with the original hypothesis that states that confidence is only affected by the amount of previous experience. The rationale behind the original hypothesis is that new evidence can only revise the current belief, and the confidence in the revised belief depends on the amount of previous experience with similar situations. Therefore, new evidence cannot affect both belief and confidence. In this sense, it seems that belief and confidence can be easily confounded and one's capacity of explicitly reporting confidence is limited. It is interesting to note that the notion of limited capacity of explicitly distinguishing confidence and belief is also supported by another interesting phenomenon in Figure 3.4. It seems that participants made higher confidence judgments

in the 3:1 base rate condition than in other base rate conditions, especially at the later stage of the training. This can be conveniently explained by the possible confusion in explicitly assessing belief and confidence.

It is interesting to note that whether the gradually tuning to approach the theoretical values in terms of final decisions (see Figure 3.5 and Figure 3.6) represents a kind of optimal (or rational) behavior is a controversial issue. Let us take the 3:1 base rate condition for an example. On the one hand, since in general friendly targets are more frequent than hostile ones in this condition, from the perspective of pursuing the maximal rate of correct decisions in the long run, it seems that making a friendly decision in every trial is the optimal strategy. This strategy suggests that to be optimal, the proportion of friendly decisions should be 100%, which is clearly not what participants did. On the other hand, since participants did not make decisions until they collected both pieces of evidence, one may suggest a different optimal criterion based on the trial distribution in Table 3.1. More specifically, when at least one piece of evidence is positive, since friendly targets are more frequent than hostile ones in these cases (12:1, 6:2, and 6:2, for evidence sequences “++”, “+-“, and “-+”, respectively), one should always make friendly decisions. On the contrary, when both pieces of evidence are negative, since hostile targets are more frequent than friendly ones in these cases (3:4 for evidence sequence “--“), hostile decisions should be made. As a result, one should make about 81% $((12+1+6+2+6+2)/36=81\%)$ friendly decisions in order to be optimal. It is clear from Figure 3.6 that this optimal strategy does not match participants' behavior either. Finally, in terms of the built-in theoretical base rates and conditional probabilities, derived from

Table 3.1 and presented in Table 3.2, it is clear that Figure 3.5 and Figure 3.6 show typical patterns of approximate *probability matching*, which refers to the phenomenon in which people tend to make a decision with probability p if a proportion p of the events turn out to be consistent with that decision. Probability matching is often characterized as irrational or sub-optimal, based on the optimal strategies mentioned above. However, it has been argued that probability matching is rational in terms of rational analysis (J. R. Anderson, 1991; Anderson & Lebiere, 1998), especially when a non-stationary environment is assumed.

It is important to note that one potential problem with the current trial distribution (i.e., Table 3.1) is that the total trial numbers in “++”, “+-”, “-+”, and “--” evidence combinations are not equal. This is true for all three base rate conditions. For example, in the 1:1 base rate condition, there are 10 trials in each of the “++” and “--” evidence conditions, and 8 trials in each of the “+-” and “-+” evidence conditions. This might be a potential confounded factor. However, it seems that the current trial distribution is the only distribution that satisfies most of the constraints, including conditional independence and the base rate constraint. Effects have been made to make the trial number difference in different evidence combinations as small as possible.

It is worth pointing out that each participant was trained in one evidence collection order (i.e., either Route-SelfID or SelfID-Route), but was tested in both orders in the evaluation trials. As mentioned earlier, the training order is a control variable. However, one may wonder the possible effect of this control variable. More specifically, one may ask, is there a difference between order effects derived from those evaluation trials that

have the same evidence collection order as the training trials and order effects derived from those evaluation trials that have the opposite evidence collection order as the training trials? Statistics show that the training order has no significant effects on the evaluation of order effects (see Appendix F for details). This result was also supported by a previous study (Zhang, Johnson, & Wang, 1997).

In next chapter, I show how a computational model, UECHO, is developed to model and further explain the experiment results.

CHAPTER 4

UECHO: A COMPUTATIONAL MODEL

Computational models are often very useful for theoretical investigations of mental processes (see Newell, 1990; Anderson, 1993; Levine, 1991; Thagard, 1996). Building a computational model contributes to theory evaluations in three ways. First, a computational model helps to show that the postulated representations and processes are computationally realizable. This is important because many algorithms that seem reasonable at first glance do not scale up to real-world problems. Second, in order to show not only the computational realizability of a theory but also its psychological plausibility, it is possible to apply a computational model qualitatively to various real-world situations and test if it achieves realistic and reasonable performance. An example is that a computational model should show the same kind of biases that people show. Third, to show a much more detailed fit between the theory and human thinking, a computational model can be used quantitatively to generate detailed predictions about human thinking, which can be compared with the results of psychological experiments. For example, if psychological experiments indicate that people make a certain percentage

of one class of errors, then the model should predict the errors at roughly the same percentage level. Cognitive theories by themselves are normally not precise enough to generate such quantitative predictions. A model can fill the gap between theory and observation.

In this chapter, I describe UECHO, a computational model developed to model human belief revision. UECHO takes a coherence view of human belief as its foundation. In addition, UECHO fully incorporates the multi-component structure and the adaptive characteristic of human belief. As a result, UECHO is expected to account for the full range of human belief revision and order effects. More specifically, UECHO should show order effects when experience is scarce. As UECHO learns and gains experience through interacting with the environment, order effects should become weaker and eventually disappear.

The development of UECHO serves two purposes. First, it provides an alternative coherence-based processing model of human belief revision, which is beyond the traditional weight-assignment-based integration model. As I will show later, coherence-based models of human belief revision have been supported by many studies. In addition, UECHO suggests a rationality basis for order effects in human belief revision. Since UECHO is constructed based on rational postulates and is intended to capture what people ought to do (see Thagard, 1989, 1992a), the fact that UECHO naturally embodies order effects convincingly “debiases” these effects.

This chapter is organized as follows. In section 4.1, I provide a brief overview of the foundations of UECHO. The Theory of Explanatory Coherence and its connectionist

implementation, ECHO, are presented, and the limitations of ECHO are discussed. In section 4.2, the details of the UECHO implementation are described. In section 4.3, I present the modeling results and compare them with the experimental results described in Chapter 3. The final section contains discussions and conclusions.

4.1. Overviews

Human belief has, in addition to a multi-component structure and an adaptive characteristic, a coherence foundation. The coherence foundation of human belief implies that beliefs co-exist as a system by supporting each other. Belief is seldom isolated. The acceptability of any belief is potentially dependent upon its relations to other beliefs. Maintaining the coherence of the global system is the ultimate justification for any belief's acceptance or rejection.

The coherence foundation of human belief has a significant influence on human belief revision in general. On the one hand, a new piece of evidence may be simply ignored and the current beliefs are hardly revised if doing so results in a more coherent system than otherwise. This can happen even if the new evidence represents the results of observations. On the other hand, as long as the system coherence is improved, a new piece of evidence may completely destroy the current belief system and result in a paradigm shift, which has been hypothesized to occur in scientific revolution (Kuhn, 1970).

4.1.1. The Theory of Explanatory Coherence

The Theory of Explanatory Coherence, or TEC, is proposed by Thagard (1989, 1992a) as a model of human abductive reasoning, a process of inference to the best (i.e., most believable) hypothesis or hypotheses that can explain observed data (For general treatments of abduction, see Fann, 1970; Josephson & Josephson, 1994). As indicated by its name, TEC takes a coherence view of human belief in general and hypothesis evaluation in particular. According to the theory, it is the coherence based on explanatory relations that is fundamental to human belief evaluation. More specifically, the acceptability of a proposition depends on the explanatory coherence of the set of propositions to which it belongs. A proposition should be accepted if it is coherent with other beliefs, rejected if it is incoherent with other beliefs, and regarded neutral if it is neither coherent nor incoherent. This is true even if the proposition represents the observed data. In addition, a proposition that has the most explanatory coherence, evaluated based on all other related beliefs in the system, should be taken as the most believable proposition.

Explanatory coherence is one of the fundamental concepts in TEC. According to TEC, propositions *P* and *Q* *cohere* if there is some explanatory relation between the two. More specifically, propositions *P* and *Q* cohere explanatorily if one or more of the following conditions is satisfied:

- *P* is part of the explanation of *Q*;
- *Q* is part of the explanation of *P*;
- *P* and *Q* are together part of the explanation of *R*;

- *P and Q are analogous in the explanations they respectively give of some R and S.*

In addition, *incoherence* means more than just that two propositions do not cohere. Propositions *P* and *Q* *incohere* if they contradict each other or if they offer competing explanations.

How to define the notion of explanation raises a question. Disagreement exists about whether explanation is primarily deductive (e.g., *p* entails *q*), statistical (e.g., *p* is probabilistically associated with *q*), causal (e.g., *p* causes *q*), or linguistic (e.g., *p* is semantically related to *q*). TEC does not offer its own definition of explanation. Instead, it takes explanation as primitive and claims that it is compatible with any of the above accounts.

On the basis of the concept of explanatory coherence, TEC proposes seven principles that establish explanatory relations among propositions and regulate the global coherence of an explanatory system. Let *P*, *Q*, and $P_1 \dots P_m$ be propositions in an explanatory system *S*. Then the principles are as follows (Thagard, 1992a):

symmetry: If *P* and *Q* cohere, then *Q* and *P* cohere.

If *P* and *Q* incohere, then *Q* and *P* incohere.

explanation: If $P_1 \dots P_m$ explain *Q*, then:

For each P_i in $P_1 \dots P_m$, P_i and *Q* cohere;

For each P_i and P_j in $P_1 \dots P_m$, P_i and P_j cohere;

And the degree of coherence is inversely proportional to the number of propositions $P_1 \dots P_m$.

Analogy If P_1 explains Q_1 , P_2 explains Q_2 , P_1 is analogous to P_2 , and Q_1 is analogous to Q_2 , then P_1 and P_2 cohere, and Q_1 and Q_2 cohere.

data priority: Propositions that describe the results of observations have a degree of acceptability of their own.

Contradiction If P contradicts Q , then P and Q incohere.

competition: If P and Q both explain a proposition, and if P and Q are not explanatorily connected, then P and Q incohere.

acceptability: The acceptability of a proposition P in S depends on its coherence with the propositions in S .

Finally, TEC claims that the global explanatory coherence of a system is a function of the pairwise local coherence established based on the above principles. Pursuing the greatest explanatory coherence at the system level is the goal that a system attempts to reach.

Not surprisingly, TEC has also brought many criticisms (see the Open Peer Commentary in Thagard, 1989). For example, one criticism questions the coherence foundation of TEC. If sets of beliefs must be evaluated holistically, and everything is potentially related to everything else, then how can one make a reasonable judgment about which to accept and which to reject? TEC addresses this issue by arguing that not all beliefs are equally important and some are more fundamental than others. As suggested by the principle of data priority, a belief describing the results of observation has a degree of acceptability on its own. It is this evidence priority that makes efficient judgment possible. However, it is important to note that due to the fundamental

requirement of coherence, a proposition having some independent acceptability may still end up not accepted if it only coheres with propositions that are not themselves acceptable.

Not precisely defining “explanation” is another common criticism of TEC. For example, Achinstein (1989) argues that without a reasonable account of “explanation”, there is no guarantee that explanation insures acceptability. However, Thagard insists that a precise definition of “explanation” is out of sight. The best we can do is to describe features characteristic of typical explanations, such as that some explanations are deductive, and some explanations are causal.

Despite these criticisms, it is important to point out that TEC, as a theory of human belief evaluation, has gained much empirical support (e.g., Thagard, 1989, 1992a, 1992b).

4.1.2. ECHO

Though TEC is intrinsically interesting, it is too vague to make any quantitative predictions. ECHO, a connectionist implementation of TEC, fills the gap by demonstrating how TEC works computationally.

A connectionist model is typically composed of nodes, connections between nodes, and laws for the changes over time of the node activation and connection strengths (for general treatments of connectionism, see Rumelhart and McClelland, 1986; Quinlan, 1991; Levine, 1991; Anderson, 1995). In ECHO, propositions representing hypotheses and results of observation are represented by nodes. Coherent relations are represented by excitatory connections, which have positive weights so that the activation of one

proposition increases the activation of another. Incoherent relations are represented by inhibitory connections, which have negative weights so that the activation of one proposition decreases the activation of another. The principle of symmetry insures that these connections are symmetric. The activation of a node, ranging from -1 to 1 represents acceptability. Positive activation indicates the acceptance of the proposition represented by the node, negative activation indicates the rejection, and activation close to 0 indicates neutrality.

The principle of data priority is implemented in ECHO by connecting to each data node from a special evidence unit (SEU), which always has activation 1 . A positive weight of this connection gives each data node some acceptability on its own. The weights can be different. The higher the weight, the more enduring the associated data node is.

When the network is running, activation spreads, through the connections, from the SEU to the data nodes and then to other connected nodes. The network is holistic in that the activation of every node can potentially have an effect on every other node.

Synchronous activation updating occurs at each running cycle. More specifically, for each unit j , its new activation a_j is a function of both its activation at a previous time point and the net effect of all the units connected to it. Quantitatively, the activation of a node j is updated according to the following rule:

$$a_j(t+1) = a_j(t)(1 - \theta) + \begin{cases} net_j \times [\max - a_j(t)], & \text{if } net_j > 0 \\ net_j \times [a_j(t) - \min], & \text{if } net_j \leq 0 \end{cases} \quad (4.1)$$

where t represents time in the unit of cycle, θ is a decay parameter that decrements each node at every cycle, max and min are constants representing maximum (+1) and minimum (-1) activation respectively, and net_j is the net input to node j , which is defined as the weighted sum of activations from all the connected neighbor nodes. Quantitatively,

$$net_j = \sum_i w_{ij} a_i(t) \quad (4.2)$$

where w_{ij} is the weight of connection from node i to node j .

The network repeats the updating cycles until it converges, which happens when no node has changed activation larger than a specified threshold value. During this process, propositions that are incoherent die out and propositions that are coherent are strengthened. When the network settles, the most activated propositions represent the most plausible explanations.

A measure H of the global explanatory coherence of a system at time t is defined as follows:

$$H(t) = \sum_i \sum_j w_{ij} a_i(t) a_j(t) \quad (4.3)$$

Note that H is maximized when the system converges. More specifically, the convergence occurs: 1) if w_{ij} is positive, a_i and a_j should have the same sign and reasonably big absolute activation values (i.e., the activation of one excites the other); 2) if w_{ij} is negative, a_i and a_j should have the opposite sign and reasonably big absolute activation values (i.e., the activation of one inhibits the other). In this sense this measure

H has been called “harmony”. Thus ECHO stands for “Explanatory Coherence by Harmony Optimization”.

It is important to note that the impact of net effect depends on the current activation level. If the net effect is positive, the impact is constrained by the range $[a_j(t), \max]$. This means that, if the current activation $a_j(t)$ is already large, there is little room (i.e., $\max - a_j(t)$) for further improvement, no matter how strong the positive net effect is. In contrast, if the net effect is negative, then impact is constrained by the range $[\min, a_j(t)]$. In this case, if the current activation $a_j(t)$ is large, then $a_j(t) - \min$ is also large, which means the impact of net effect would be greatly magnified. This is called the *contrast effect*.

Parameters give ECHO numerous degrees of freedom. Some important ones include:

Tolerance Nodes with excitatory connections reinforce each other and nodes with inhibitory connections suppress each other. If inhibitory weight is low compared to excitatory weight, excitation will overcome inhibition thus incoherent hypotheses can all be activated. Tolerance refers to the absolute value of the ratio of excitatory weight to inhibitory weight. With high tolerance, the system will entertain competing hypotheses. With low tolerance, winning hypotheses suppress the losing ones. Typically ECHO is run with an excitatory weight of 0.04 and an inhibitory weight of -0.06, which lead to a tolerance of about 0.67.

Simplicity impact If multiple hypotheses need to be combined to explain a data node,

then the principle of explanation says that the excitatory connection between any one of the cohypotheses and the data node is inversely proportional to the number of cohypotheses. For example, if P explains E, then the connection weight between P and E is the default excitatory weight, say 0.04. However, if P and Q together explain E, then typically the connection weight between P and E becomes the default value divided by 2, that is, 0.02. The system can be made to emphasize simplicity and penalize complexity. This is controlled by the parameter of simplicity impact, as shown in the following equation:

$$weight(P, E) = \frac{\text{default weight}}{(\text{number of cohypotheses of } P)^{(\text{simplicity impact})}} \quad (4.4)$$

Therefore, the greater the simplicity impact, the lower weights for having more cohypotheses. The default simplicity impact is 1.

Skepticism The decay rate θ in the activation updating equation is called the parameter of skepticism. The higher this parameter is, the faster the node activation decays. If this parameter is very high, then no hypotheses will be activated. Thus the system is in a highly skeptical status.

Data excitation The priority of data nodes can be adjusted by varying the weights from the SEU to the data nodes. Data excitation is a parameter from 0

to 1 that provides these weights. Note that it is possible to set the weights for each data node differently to reflect the different importance. Typically, data excitation is set to 0.05.

Appendix D provides some important ECHO parameter setting used in this dissertation research.

One criticism of ECHO is that with so many degrees of freedom, it might seem that any arbitrarily desired behavior could be obtained. It has been argued that the arbitrariness is greatly reduced if a fixed set of default parameters applies to a large range of situations (Thagard, 1992a). In addition, it is possible to extend ECHO to allow the parameters being learned from examples.

Hoadley, Ranney, and Schank(1994) argue that due to attentional and memorial limitations people are seldom able to consider all the related propositions simultaneously, as suggested by ECHO. They proposed a variation of ECHO, named WanderECHO, that adopts a stochastic updating algorithm. Instead of updating every node at every cycle, WanderECHO stochastically selects and updates only one node at a cycle. However, the results show that WanderECHO is only a little bit better than ECHO in terms of matching human data. And in the limit, it can be shown that WanderECHO and ECHO calculate similar values.

Theoretically, ECHO satisfies several critical regularities in human belief evaluation. For example, ECHO prefers a hypothesis that explains more observations to the one that explain fewer. This is called *explanatory breadth*. ECHO prefers a simpler hypothesis to a more complex one. This is called *parsimony*. In addition, ECHO prefers a hypothesis

which itself is explained. This is called “*being explained*”. Finally, in ECHO the credibility of an observation also depends on its coherence in the system, suggesting that an observation could be disbelieved. This is called *data reliability*.

Empirically, ECHO has acquired much experimental support (see Thagard, 1992a). For example, ECHO has been successfully applied to scientific reasoning (e.g., Thagard, 1989, 1992a), legal reasoning (e.g., Thagard, 1989, 1992a), social interactions (e.g., Read & Marcus-Newhall, 1993), adversarial problem solving (e.g., Thagard, 1992b), and human experimental data modeling (Ranney & Thagard, 1988; Schank & Ranney, 1991, 1992).

4.1.3. Toward UECHO

Despite its success, ECHO has limitations. It is these limitations that prevent ECHO from becoming a general model of human belief revision, and motivate the development of UECHO in this dissertation research.

First of all, ECHO handles sequential belief revision in an unrealistic way (see also Schank & Ranney, 1991). ECHO assumes that all the relevant evidence is available at the very beginning and can be encoded into the network simultaneously. This is clearly not a valid assumption since in realistic situations evidence usually does not become available at the same time. The problem is even more serious when the order effect is concerned. Due to the lack of a dynamic belief revision mechanism, when new evidence becomes available, ECHO has to completely destroy the old network and re-construct a new one. Since the newly constructed network has to be re-run from scratch, useful

information prior to the new evidence is regrettably lost. This kind of “amnesia” is clearly not the way people do belief revision.

Second, ECHO does not handle learning. ECHO does not learn from its experience. Therefore, it assumes that all hypotheses are equally plausible and the strengths of all connections are fixed when they first enter the system, and they never change. This is not a valid assumption either. In many situations people do not regard all competing hypotheses equally plausible. In addition, the adaptive characteristic of human belief suggests that no explanatory relation be fixed. The interaction with the environment constantly tunes human beliefs and their relations.

Thagard (1989,1992a) insists that the probabilistic approach does not apply to the theory of explanatory coherence. His major objection to Bayesian probability repeats the standard debate between subjective probability and frequentism. He claims that in many theory evaluation situations, such as the choice between Darwin’s evolution theory and the creationism, empirical frequencies are not available; thus it is obscure to assign probabilities (both prior and posterior) to such theories. In addition, he argues against frequentism as well, and claims that “theoretical inferences, in science and ordinary life, require us to go beyond probabilistic reasoning based on frequencies” (Thagard, 1992a, page 93). These arguments are not convincing. As reviewed in Chapter 2, human belief features intrinsic uncertainty, and the Bayesian approach nicely captures the uncertainty aspect of human belief management, in both scientific reasoning and everyday life (e.g., Howson & Urbach, 1993).

ECHO's ignorance of probabilities has been criticized from a very different perspective. Van Overwalle (1998) argues that while humans are very susceptible to the covariation between hypotheses and evidence, which refers to the contingent correlation between hypotheses and evidence and is often specified in a probabilistic formulation, ECHO fails to be sensitive to such a covariation. He uses both simulation and mathematical proof to demonstrate how ECHO can arbitrarily deviate from the true hypothesis-evidence covariation. For example, he shows that ECHO can predict a positive explanatory activation even when the true hypothesis-evidence covariation is positive, zero, or negative. The same is true for negative explanatory activation. As a result, he comments that although ECHO might be able to choose the best explanation, it fails to indicate how plausible the chosen explanation actually is.

Finally, closely related to the lack of learning from the experience, ECHO does not distinguish between probability and confidence. In the light of new evidence, the activation values of all nodes are equally likely to be changed, no matter how much prior experience they are based on.

These limitations clearly cast doubt on ECHO as a general model of human belief revision. They also motivated the development of UECHO.

4.1.4. Summary

The Theory of Explanatory Coherence, proposed as a general framework for human belief evaluation, takes the coherence view of human belief as one of its foundations. According to TEC, the acceptability of a proposition depends on the explanatory

coherence of the set of propositions to which it belongs. A proposition, no matter whether it is a hypothesis or an observation, should be accepted if it is coherent with other beliefs, rejected if it is incoherent with other beliefs. Pursuing the greatest explanatory coherence at the system level is the goal that a system attempts to reach.

ECHO, a connectionist implementation of TEC, shows how TEC works computationally. In ECHO, nodes represent propositions, connections represent explanatory relations, and activation represents acceptability. The system pursues the highest harmony (explanatory coherence) by updating activation in a parallel constraint satisfaction manner. When the system converges, the most believable hypothesis set will defeat any competitor and pop out.

Although ECHO has gained much empirical support, they have serious limitations: (1) ECHO does not handle sequential belief revision; (2) ECHO does not learn from experience.; and (3) ECHO does not distinguish confidence and probability. All these limitations cast doubt on ECHO as a general model of human belief revision.

In the next section, I show how UECHO, an extension of ECHO, is developed to address these limitations and model the results of the experiment described in chapter 3.

4.2. UECHO

In this dissertation I propose UECHO to extend ECHO to adequately handle uncertainty. Thus UECHO stands for “Uncertainty-aware ECHO”.

More specifically, UECHO extends ECHO by making three major changes: 1) A sequential belief revision mechanism is added; 2) A learning mechanism is added to

allow statistical tuning; and 3) A new confidence dimension is added to each proposition. How these changes are implemented is illustrated next.

4.2.1. Sequential Belief Revision

First of all, a sequential belief revision mechanism is added to UECHO. Note that in ECHO sequential belief revision is impossible since ECHO resets its state of activation to default values before each run, and thus the valuable information from a previous run cannot be preserved to the next run when new evidence is added. In order for UECHO to handle sequential belief revision, the previous state of UECHO needs to be preserved when UECHO accommodates new evidence and continues to run. While this is fairly easy to do, it brings two additional considerations. First, this requires that the structure of the network be as complete and stable as possible. In other words, the network should encode as many pieces of relevant evidence as possible at the very beginning, regardless of their availability. As a result, when new evidence comes in, UECHO does not have to add an additional node and restructure the network, as is done by ECHO. Instead, all UECHO needs to do is to connect the corresponding node, which already exists in the current network, to the SEU, to mark its availability and independent acceptability. Intuitively, this is a more realistic way to represent explanatory coherence. The notion of *schema* (e.g., Bartlett, 1932) and numerous studies in Gestalt psychology all suggest that people seldom maintain partial network and constantly reconstruct the network. In addition, people often make predictions based on current beliefs. This is possible only with a complete network – a partial network may not have a relevant concept or event,

which is what a prediction is based on but often new and beyond the current context, encoded. It is important to note, however, that a network as complete as possible does not mean it can be beyond various cognitive constraints, such as the capacity limitation of working memory. This is especially true when UECHO is considered as a system that involves many implicit and sub-symbolic processes.

Second, since evidence does not become available at the same time and some comes before others, sequential belief revision requires the order of evidence presentation be represented. One way to differentiate evidence available at different time points is through the parameter *data excitation* (DE for short). While the most current evidence item has a full connection, represented by DE, to the SEU, a previous evidence item should be connected to SEU with a decayed strength, indicating that it is an earlier evidence item. This solution is naturally supported by a large body of literature on the memorial and attentional attenuation of lists of items with time (e.g., Rubin & Wenzel, 1996; Anderson, 1981; Netica, 1996). Note that this solution requires an additional parameter, denoted by d ($0 \leq d \leq 1$), to represent the decay rate. The connection strength from an evidence node to the SEU is, therefore, a function of both the decay rate d and the time interval since the evidence was presented. One choice of the exact function form is to let DE decay exponentially in the square root of time (see Rubin & Wenzel, 1996 for a review of how such a function is one of the best for memory retention). Quantitatively, let t be the time passed since the evidence was presented; the DE at time $t+1$ is determined by:

$$DE_{t+1} = DE_t \times (1 - d)^{\sqrt{t}} \quad (4.5)$$

4.2.2. Learning

A learning mechanism is added to UECHO. Ideally, the weights of the connections between evidence nodes and hypothesis nodes should reflect the summarized associations of co-occurrence. However, in ECHO, the weights are fixed according to two parameters, *excitatory weight* and *inhibitory weight*. Learning from experience in UECHO means that these weights are no longer fixed, and can be modified based on experience, so that these weights can reflect the corresponding associations of co-occurrence (see Figure 4.1). This learning ability is important for UECHO to tune itself to capture statistical structure of the environment and to reflect a more realistic covariation between hypotheses and evidence. One central issue is to select an appropriate weight-updating rule.

One elegant choice for weight updating is the Rescorla-Wagner rule (Rescorla & Wagner, 1972). The Rescorla-Wagner rule was first proposed to learn the association between unconditioned stimulus (US) and conditioned stimulus (CS) in classical conditioning. It has gained wide acceptance in other areas of psychology due to its elegance, simplicity, and ability to predict associative consequences of presenting cues. In addition, it has been shown that the Rescorla-Wagner rule is identical to the well-known delta rule or LMS (Least Mean Squares) algorithm that is widely used in connectionist modeling (see Sutton & Barto, 1981; Barto & Watkins, 1990).

Figure 4.1. A graphical representation of learning in UECHO. Panel A represents the original network, in which dashed lines indicate inhibitory relations and the strengths are determined by the parameter *inhibitory weight*, and solid lines indicate excitatory relations and the strengths are determined by the parameter *excitatory weight*. Panel B represents the network after learning. Note that the connection strengths have been tuned to reflect the corresponding associations based on experience.

The central idea of the Rescorla-Wagner rule is that learning occurs when events violate expectations. Therefore, the discrepancy between expectations and actual outcomes is used to adjust the association so that future expectations approach actual outcomes. Formally stated, the Rescorla-Wagner rule is written as follows:

$$\Delta w = \alpha(\lambda - V) \times \beta x \quad (4.6)$$

In this equation, λ represents the actual outcome, V represents the expectation, x represents the set of cues, and α and β are constants. This rule says that the change in any association strength is determined by the product of the discrepancy between expected and actual outcomes and the presenting cues. $\lambda - V$ is often called reinforcement and βx is called eligibility.

The choice of the Rescorla-Wagner rule is based on the following considerations. First, numerous studies have shown that the Rescorla-Wagner rule is good at learning associative strengths from experience and has been successfully applied to a wide range of domains from behavioral specifications to connectionist modeling (see Gluck & Bower, 1988; also see Miller, Barnet, & Grahame, 1995 for a review). Although it has been shown that the Rescorla-Wagner rule has its limitations, such as its incapacity of handling complex non-linear problems (e.g., Barto, 1990) and real-time problems (e.g., Sutton & Barto, 1990), it is chosen because 1) it demonstrates that a learning mechanism can be added to Echo; and 2) it is adequate to deal with the simple learning task used in this research. Second, on the surface it seems that ECHO and the Rescorla-Wagner rule are not compatible in the sense that ECHO emphasizes symmetry and the Rescorla-Wagner rule is directional. I argue that a directional learning rule is appropriate since 1)

it is a natural mapping to the directional causal relations we are modeling; and 2) even in ECHO symmetry is not absolute: it is the hypothesis that “causes” the evidence, but not vice versa.

Let us construct a UECHO model of the CIC task described in Chapter 3 and use it as an example to demonstrate how the Rescorla-Wagner rule is applied in UECHO. Remember that the goal of the CIC task is to determine an unknown target’s identity (friendly or hostile) based on information about Route (on or off a commercial air route) and SelfID (reply or no reply to radio warning). It is straightforward to construct a simple UECHO network for this task. We use two nodes to represent hypotheses (F for “the target is friendly” and H for “the target is hostile”), and four more nodes to represent potential evidence items (R+ for “the target is on a commercial air route”, R- for “the target is off a commercial air route”, ID+ for “the target replies the radio warning”, and ID- for “the target does not reply the radio warning”). Intuitively, we assume F and H are incoherent with each other. In addition, we assume F is coherent with R+ and ID+ but incoherent with R- and ID-, and H is coherent with R- and ID- but incoherent with R+ and ID+. These assumptions result in the respective excitatory or inhibitory connections. The network is shown in Figure 4.2.

Figure 4.2. The UECHO network for the CIC task. Inhibitory connections are represented by dashed lines and excitatory connections are represented by solid lines. In this specific example, it assumes that R+ and ID+ are observed and thus they are connected to the SEU. The numbers in parentheses are the corresponding activation values after the network settles. It is clear that the hypothesis of friendliness (with an activation of 0.64) beats the hypothesis of hostility (with an activation of -0.62).

It is important to note that the activation values need to be converted to subjective probabilities. They serve as measures of the degrees of plausibility of the corresponding propositions. In addition, these probabilities play the role of expectations (i.e., the quantity V in the Rescorla-Wagner rule). A logistic transformation is conveniently adopted (see Gluck & Bower, 1988). Quantitatively,

$$p = \frac{1}{1 + e^{-a}} \quad (4.7)$$

where a and p represent the activation value and the resulting probability measure, respectively, and τ is a constant. Such a mapping rule has the property that negative activation values are mapped to probabilities between 0 and 0.5, that 0 is mapped to 0.5, and that positive activation values are mapped to probabilities between 0.5 and 1. This mapping rule is shown graphically in Figure 4.3.

The network runs as follows. Given an episode, for instance, a target with both R+ and ID+ being true, UECHO runs and settles. Then UECHO converts the activation values (-1 to 1) of hypothesis nodes to subjective probabilistic beliefs (0 to 1), using the logistic function described above. In this specific case UECHO would report that a friendly hypothesis is more likely (see also Figure 4.2). Now, suppose UECHO receives some information indicating the true identity of the target, say it is indeed friendly. UECHO made the right choice and should be rewarded. The reward takes the form of updating the weights between the hypothesis nodes and the evidence nodes according to the Rescorla-Wagner rule, with a reinforcement of $1-p(F)$ (where $p(F)$ is the probability of F) for connections associated with node F and a reinforcement of $0-p(H)$ (where $p(H)$

is the probability of H) for connections associated with node H. The consequence is such that if a similar scenario is encountered in the future, UECHO should make the same decision, but with stronger degrees of belief. Similarly, if UECHO's decision turns out to be wrong, the weights should be changed correspondingly so that in a similar situation in the future, it will be less likely that UECHO will make the same mistake again.

Figure 4.3. The logistic mapping rule between activation values and probabilities, with τ being -3.2 .

It is important to note that the Rescorla-Wagner rule only calculates the changes of weights. The weights are actually changed in such a way that the resulting weights are always bounded in the range between two parameters, *inhibitory weight* (IW) and *excitatory weight* (EW), in the sense that IW and EW represent the strongest associations of not-co-occurrence and co-occurrence. Quantitatively,

$$w_{i+1} = w_i + \Delta w \times \begin{cases} (EW - w_i) & \text{if } \Delta w \geq 0 \\ (w_i - IW) & \text{otherwise} \end{cases} \quad (4.8)$$

The whole procedure is shown graphically in Figure 4.4.

4.2.3. Confidence

Finally, a new confidence dimension is added to the network. Consequently, each proposition contains not only an activation value that indicates its believability, but also a confidence value that denotes how much prior experience this activation value is based on. More specifically, as one solves more and more real-world problems, one's experience goes up, and so does one's confidence in future judgments. Confidence plays an important role in one's belief revision process. That is, the higher the confidence is, the harder it is to alter the belief. In other words, the impact of new evidence depends upon not only the strength of the new evidence and the current degree of belief, but also the ease with which the current belief is changeable.

It should be clear that confidence is a function of prior experience. The central issue is how confidence is represented.

Figure 4.4. The flowchart of associative learning in UECHO. Panel A represents the initial UECHO network with R+ and ID+ being true. Note that the SEU is not shown. Panel B represents the settled network. Note that the friendly node has a higher activation than the hostile node, represented by the degrees of grayness. Panel C indicates that a feedback of the true identity is provided (in this case it is friendly). Learning via the Rescorla-Wagner rule is shown in Panel D.

Netica, a commercial software for Bayesian belief networks, uses *experience* to represent confidence. According to Netica, *experience* is simply a measure of the number of relevant samples Netica has seen so far (Netica, 1996). More specifically, when a new case is encountered, the corresponding *experience* number is updated as follows:

$$experience_{n+1} = experience_n + 1 \quad (4.9)$$

where n is the number of cases that Netica have already seen. Note that this *experience* number is directly involved in the process of probability learning. Formally, the probabilities of relevant variables are changed according to the Bayesian learning technique:

$$probability_{n+1} = \frac{probability_n \times experience_n + 1}{experience_{n+1}} \quad (4.10)$$

The above equation clearly shows the important role confidence plays. When the *experience* number is large, the new case has little effect and the new probability is dominated by the previous probability value. On the contrary, when the *experience* number is small, the new case contributes considerably to the resulting probability.

The same notion of *experience* is adopted in UECHO. In the current case it represents the number of trials UECHO has performed or the number of scenarios UECHO has encountered. The key idea of confidence management in UECHO is to revise or tune relevant parameters based on the number of *experience*. More specifically, there are three mechanisms that are combined to enable UECHO to sufficiently handle confidence.

The first mechanism is related to the parameter of skepticism θ . As mentioned previously, θ represents the decay rate in the activation updating equation. The higher θ is, the faster does the node activation decay. If θ is very high, then no hypotheses will be activated, which in turn results in a highly skeptical system. Not surprisingly, confidence, based on experience, cures skepticism. Gradually tuning down θ as the experience number goes up is a natural way to represent confidence. A logistic function is conveniently adopted. Formally,

$$\theta_n = \theta_0 \times \left(a - \frac{1}{1 + e^{-b(n-c)}} \right) \quad (4.11)$$

where n is the experience number, a , b , and c are constants. This equation is shown graphically in Figure 4.5.

The second mechanism has to do with the parameters α and β in Hogarth and Einhorn's (1992) Anchoring and Adjustment model. According to the model, α and β ($0 \leq \alpha, \beta \leq 1$) represent one's sensitivity toward "negative" and "positive" evidence, respectively. Small values of α and β imply low sensitivity to new evidence while large values of α and β imply high sensitivity. Although Hogarth and Einhorn attribute the difference of sensitivity mainly to some unidentifiable factors of personality, the parameters α and β are functionally closely related to confidence in the sense that as the confidence goes up, the sensitivity to new evidence goes down. Therefore, gradually tuning α and β represents another aspect of confidence management.

Figure 4.5. The parameter of skepticism, θ , as a function of the *experience* number. As θ decreases, confidence increases. Based on equation 4.11, and $\theta_0=0.1$, $a=1.5$, $b=-0.032$, and $c=0$.

It is straightforward to embed the two parameters into UECHO's activation updating equation. Formally,

$$a_j(t+1) = a_j(t)(1-\theta) + \begin{cases} net_j \times \beta \times [\max - a_j(t)], & \text{if } net_j > 0 \\ net_j \times \alpha \times [a_j(t) - \min], & \text{if } net_j \leq 0 \end{cases} \quad (4.12)$$

As a result, ECHO's activation updating equation is a special case of UECHO's when α and β are both equal to 1.

How are the parameters α and β tuned? It is easy to see they should be functions of prior experience. In order to distinguish between sensitivity to positive evidence (β) and sensitivity to negative evidence (α), it is also necessary to further classify prior experience. More specifically, when prior experience shows that positive events, which are the events that are supported by positive evidence, are the majority, it is natural to apply a lower sensitivity to negative evidence. On the contrary, when prior experience shows that negative events, which are the events that are supported by negative evidence, are the majority, a lower sensitivity to positive evidence is reasonable. In addition, when positive and negative events are equal in numbers according to prior experience, no preference should be assumed. In all those cases, α and β decrease as experience increases. A similar logistic function is adopted. Formally,

$$\alpha_n = \alpha_0 \times \left(a - \frac{1}{1 + e^{-b \times \frac{n_p}{n} \times (n-c)}} \right) \quad (4.13)$$

$$\beta_n = \beta_0 \times \left(a - \frac{1}{1 + e^{-b \times \frac{n_n}{n} \times (n-c)}} \right) \quad (4.14)$$

where n_p and n_n denote the positive experience number and negative experience number, respectively, and a , b , and c are constant.

It is very important to note that this type of tuning reflects an emphasis on the base rate information. In other words, when in general a specific class of event is more frequent, the evidence that suggests other classes tends to be de-appreciated. This mechanism does not always produce reasonable behavior, especially in situations where the environment is not stationary. In these situations, sufficiently appreciating any evidence that contradicts the current belief is important.

The third mechanism for confidence management is related to the parameter *data excitation*. As mentioned before, data excitation is a parameter from 0 to 1 that provides the weights from the SEU to data nodes. The rationale is that these data nodes represent observed events and thus should have independent support. The hypothesis nodes, therefore, do not have such connections.

A unique feature of UECHO is that the hypothesis nodes can learn and remember their activation values. Therefore, unlike ECHO, which assumes all the hypotheses are equally plausible when no evidence is presented, in UECHO, this is true only in the case of ignorance. As experience is gained, the hypothesis nodes tune themselves to capture the statistical structure of the environment. As a result, they may acquire different activation values even when no evidence is available. This is just the essence of the base rate information. The base rate learning takes the form of Bayesian learning as follows (see also Anderson, 1993):

$$p_n = \frac{s + n_1}{t + n} \quad (4.15)$$

$$a_n = k \log\left(\frac{1 - p_n}{p_n}\right) \quad (4.16)$$

where p_n is the base rate in terms of probability, a_n is the base rate in term of activation, n is the total number of experience, n_1 is the number of experience in which the hypothesis is true, s and t are constants, representing prior true experience number and prior total experience number, respectively, and k is another constant. The logarithm function is the inverse of the logistic function mentioned before that is used to transform activation to probability.

This feature of UECHO that the hypothesis nodes are able to learn and remember their base rate information provides another way to represent confidence. That is, if as experience increases, the hypothesis nodes can construct non-zero connections to the SEU, then these connections are equivalent to the strengthened confidence. In other words, as experience goes up, the hypothesis nodes will gain independent support from the SEU. And the magnitude of the independent support, as a function of experience, can serve an indication of confidence.

The strengths of connections between the hypothesis nodes and the SEU, originally 0, are learned via Bayesian learning as follows:

$$w_{0i} = w_{i0} = \frac{s + (n_i - n_i)}{t + n} \quad (4.17)$$

where n_i is the true experience number of the i^{th} hypothesis node, n_i is the true experience number of all other hypothesis nodes, and s and t are constants. Normally s is

small (e.g., 0), and t is large (e.g., 1000), indicating a very low prior probability ($0/1000=0$) that a hypothesis node is independently supported. The logic is as follows. Suppose hypothesis i is more frequently true, then as experience increases, n_i will become larger than n_j . As a result of the tuning, w_{i0} will become larger as well, indicating that hypothesis i is getting a higher degree of independent support.

It is very important to note that the weights can become either negative or positive. Both are considered as reasonable confidence. When the weights are negative, the corresponding hypothesis nodes also have negative activation values. Taking together, the weights still lead to independent support.

4.2.4. Summary

UECHO extends ECHO by adding a sequential belief revision mechanism, a learning mechanism, and a confidence management system.

Sequential belief revision is easy to implement, although it requires the network be as complete as possible. In addition, the temporal difference of evidence items needs to be explicitly represented.

Learning in UECHO is implemented using the Rescorla-Wagner rule. The Rescorla-Wagner rule, as a good model for learning associative strengths, has been supported by many studies.

Confidence in UECHO is represented by three mechanisms. The decay parameter θ applies to every type of nodes. The parameters α and β focus on the evidence side of belief revision. Connecting the hypothesis nodes with the SEU is yet another way to

implement confidence. In all three mechanisms, confidence is a function of prior experience.

With these changes, UECHO is expected to serve as a general model of human belief revision. Next, I present how UECHO is used to model the experimental results described in Chapter 3.

4.3. Modeling Results

The software used in this modeling work was developed in C/C++ and TCL/TK on a Sun SparcStation 2 running SunOS 4.2.

The UECHO network that is used to model the CIC task is shown in Figure 4.2. Although it has been described previously, it is briefly repeated here for completeness. The network consists of 2 hypothesis nodes (F and H) and 4 evidence nodes (R+, R-, ID+, and ID-), with excitatory connections between F and R+, F and ID+, H and R-, H and ID-, and inhibitory connections between F and H, F and R-, F and ID-, H and R+, and H and ID+. In this problem it is convenient to think of the coherence relations in terms of causality. For example, it is reasonable to say that a friendly commercial target causes its being on a commercial route, or a hostile target causes its not responding to a radio warning. However, it is important to note that the causality is not absolute in the sense that friendliness does not entail on-route and hostility does not entail no-response. Covariation or probabilistic regulation represents an inevitable aspect of any causality in general.

The simulation adopted the same experimental design described in Chapter 3. 120 participants were simulated, with 40 in each base rate condition (i.e., 1:1, 3:1, and 1:3). With each base rate condition, half of the simulated participants received evidence in the order of Route-SelfID in the training blocks, and the other half received evidence in a reversed order. Each simulated participant first performed an initial evaluation block, which consists of 8 no-feedback trials. Then the simulated participant was engaged in the training session, which consists of 4 training blocks, with a standard evaluation block after each. A training block contains 36 trials, distributed according to the base rate condition (see Table 3.1). For each simulated participant, the order of trial presentation was completely random. Note that this randomness is the only difference among simulated participants in each base rate condition.

In each trial, the simulated participant had to provide an initial evaluation first (that is, the network was run without any evidence input). Then two pieces of evidence were presented sequentially, and the simulated participant had to provide evaluation after each (that is, the network was run after each piece of evidence). Finally a decision was made by choosing the hypothesis with a highest activation value. In a normal training trial, feedback about the true identity was then presented, and the simulated participant was supposed to learn via the learning mechanism. In addition, after each training trial, the simulated participant tuned itself according to the confidence management mechanisms. A segment of UECHO running trace is provided in Appendix E.

The training data and the evaluation data are reported separately. Similarly, the data analysis focuses on the statistical tuning of belief judgments and the pattern change of

order effects. Similar statistical techniques and conventions adopted in analyzing the experimental results are used here. The experimental results and the simulation results are compared whenever possible. In addition, it is important to note that comparable confidence data are not available in the simulation. It is difficult to compare the explicitly reported confidence data from the experiment with the confidence measures from the simulation, such as θ .

4.3.1. Training Data

Each simulated participant was trained for 144 trials, which were divided into 4 blocks. In each trial, 4 data points were collected: the initial belief judgment (belief0), the belief judgment after the first piece of evidence (belief1), the belief judgment after the second piece of evidence (belief2), and the final 2-alternative decision (decision).

The results of belief0, based on different base-rate conditions, are shown in Figure 4.6. The performance improvement is evident. The mean values of belief0 for the 1:1 condition, which has a theoretical base rate of 50, are approximately constant (49.8, 49.8, 50.0, and 50.0, from block 1 to block 4, respectively). The mean values for the 3:1 condition, which has a theoretical base rate of 75, gradually increase (56.2, 61.4, 66.0, and 68.0, from block 1 to block 4, respectively). In contrast, the mean values for the 1:3 condition, in which the theoretical base rate is 25, gradually decrease (43.5, 39.0, 34.9, and 32.7, from block 1 to block 4, respectively). Linear regression analyses show that in all three base rate conditions the training trial number is a good predictor of the initial belief judgment. More specifically, the slope parameter estimates for 3:1 and 1:3

conditions (0.11 and -0.10, respectively) are both significantly different from 0 ($F(1,142)=1876$, $p<0.001$; and $F(1,142)=1760$, $p<0.001$; respectively). Although statistics show that the slope parameter estimate for the 1:1 base rate condition is significantly different from 0 ($F(1,142)=17.37$, $p<0.01$), it is very small (0.002). Overall, the results indicate that the initial belief judgments are getting closer and closer to the built-in theoretical base rates with the training.

Comparing Figure 4.6 with Figure 3.3, which contains the corresponding experimental data, it is easy to find that the two are reasonably similar. Statistics show that the correlation for the 3:1 base rate condition is 0.88, with a 95% confidence interval of (0.84, 0.91), and the correlation for the 1:3 base rate condition is 0.83, with a 95% confidence interval of (0.77, 0.88). Due to the low variance of the simulation data, the correlation for the 1:1 base rate condition is low (0.02), with a 95% confidence interval of (-0.15, 0.18).

Figure 4.6. Initial belief judgments (belief0) across the 144 training trials. The corresponding locally-weighted regression (loess) curves are also plotted. Based on the simulation data in all three base rate conditions. The numbers in parentheses indicate the corresponding theoretical base rates of friendly targets in terms of percentage.

The performance changes in terms of final decisions (decision) are shown in Figure 4.7, which depicts the proportion of friendly final decisions against the 4 training blocks. The gradual tuning of the decision behavior is clear. The model tends to make more and more friendly final decisions in the 3:1 base rate condition (0.71, 0.80, 0.84, and 0.89, from block 1 to block 4, respectively, $F(3,117)=216.03$, $p<0.001$), and fewer and fewer hostile final decisions in the 1:3 base rate condition (0.31, 0.21, 0.19, and 0.15, from block 1 to block 4, respectively, $F(3, 117)=207.07$, $p<0.001$). In the 1:1 base rate condition, the proportions of friendly final decisions are relatively constant (0.51, 0.50, 0.50, and 0.48, from block 1 to block 4, respectively, $F(3,117)=1.98$, $p>0.12$). An overall ANOVA indicates that the interaction between the base rate conditions and the training blocks is significant ($F(6,351)=139.42$, $p<0.001$).

Although the general pattern in Figure 4.7 is similar to the relevant experimental results shown in Figure 3.5, the model makes more extreme decisions. For example, in the 3:1 base rate condition, the model makes 89% friendly final decisions in block 4, which is significantly greater than the theoretical value, 75%, ($t(39)=21.6$, $p<0.01$). Similarly, in the 1:3 base rate condition, the model makes 15% friendly final decisions in block 4, which is significantly smaller than the theoretical value, 25%, ($t(39)=-18.0$, $p<0.01$). On the contrary, as described in Chapter 3, human participants tend to match the theoretical values.

Figure 4.7. Proportion of friendly decisions across the four training blocks. Based on the simulation data in all three base rate conditions.

UECHO's extremism in final decisions is shown even more clearly when the final decisions are further decomposed based on different evidence sequences. They are presented in Figure 4.8, which depicts the proportion of friendly final decisions against the 4 training blocks, separated for the evidence sequences “++” (i.e., both evidence items are positive), “--” (i.e., both evidence items are negative), and “+-” (i.e., one evidence item is positive and the other is negative). For the evidence sequence “++”, in all base rate conditions except for the 1:3 base rate condition, UECHO made 100% friendly decisions in all four training blocks. In the 1:3 base rate condition, the proportions of friendly decisions decrease with training (1.00, 1.00, 0.94, and 0.78, from block1 to block4, respectively). Even in the fourth block, the proportion of friendly decisions (0.78) is far away from the theoretical probability, 0.57, ($t(39)=7.67$, $p<0.01$). An overall ANOVA shows that the interaction between the base rate conditions and the training blocks is significant ($F(6,429)=44.32$, $p<0.001$). A similar pattern is found for the evidence sequence “--”. In all but the 3:1 base rate conditions, UECHO made 0% friendly decisions in all four training blocks. In the 3:1 base rate condition, the proportions of friendly decisions increase with training (0, 0, 0.19, and 0.46, from block1 to block4, respectively, and the proportion of friendly decisions (0.46) is far away from the theoretical probability, 0.43. ($t(39)=0.82$, $p>0.41$). An overall ANOVA shows that the interaction between the base rate conditions and the training blocks is significant ($F(6,429)=104.98$, $p<0.001$). Finally, the extremism is shown also shown for the evidence sequence “+-”. In the first block, the proportions of friendly decisions are 0.52, 0.78, and 0.25, for the 1:1, 3:1, and 1:3 conditions, respectively. Statistics show all are

not significantly different from the corresponding theoretical values ($t(39)=0.81$, $p>0.42$, and $t(39)=1.88$, $p>0.06$, $t(39)=0.25$, $p>0.80$, respectively), indicating an approximate probability matching in the first block. However, as training continues, while the proportions of friendly decisions stay constant in the 1:1 base rate condition (0.52, 0.49, 0.49, and 0.45, from block1 to block4, respectively), the proportions tend to become far beyond the theoretical values in both 3:1 and 1:3 base rate conditions (0.78, 0.99, 1.00, and 1.00; and 0.25, 0.03, 0.01, and 0; respectively). An overall ANOVA shows that the interaction between the base rate conditions and the training blocks is significant ($F(6,429)=92.35$, $p<0.001$). Compared to Figure 3.6, these patterns suggest that UECHO may be sensitive to the optimal strategy, as discussed in Chapter 3.

4.3.2. Evaluation Data

Each simulated participant performed the evaluation block 5 times: before any training (block0), and immediately after each training sub-block (block1 to block4). Evaluation trials are the main tools to evaluate and test possible changes of the belief revision patterns.

The belief revision pattern at the very beginning (i.e., block0) is shown in Figure 4.9. A recency effect is visually evident (23.3 for positive-then-negative evidence sequence vs 76.7 for the negative-then-positive evidence sequence). Note that due to the lack of built-in noise (i.e., all the randomness in UECHO comes from the random order of training trials), all simulated participants obtained the identical initial belief revision data. As a result, statistics cannot be done to test the significance of the recency effect.

Figure 4.8. Proportion of friendly decisions across the four training blocks, separated for different evidence sequences. The dashed lines indicate the theoretical probabilities (see Table 3.2). Based on the simulation data in all three base rate conditions.

Figure 4.9. The belief revision pattern based on the simulation data from the initial evaluation block (i.e., block0). “init”, “e1”, and “e2” indicate the time points before any evidence, after the first evidence, and after the second evidence, respectively. “+” denotes a positive evidence, and “-” denotes a negative evidence.

Comparing to the corresponding experimental data shown in Figure 3.7, the general patterns are similar. Both reveal recency effects. However, the recency effect shown by human participants (42.4 vs 51.7, the difference is 9.3) is relatively mild, comparing to the one shown by the model (23.3 vs 76.7, the difference is 53.4). Therefore, without any training, it seems that the model makes more dramatic belief revision in the light of new evidence.

The belief revision patterns during the training (i.e., block1 to block4) for the 1:1 base rate condition are shown in Figure 4.10. First of all, the average initial belief judgment (i.e., before any evidence) remains constant across the whole training session (50.0, 50.0, 50.0, and 50.0, from block1 to block4, respectively, $F(3,117)=0.10$, $p>0.95$). Second, the magnitudes of order effects decrease from block1 to block4 (54.2 for the $-+$ evidence sequence vs 45.8 for the $+-$ evidence sequence, 51.5 vs 47.9, 50.3 vs 48.3, and 49.8 vs 49.0, respectively; the differences are 8.4, 3.6, 2.0, and 0.8, respectively). Due to small standard deviations, however, statistics show that all order effects are significant ($t(39)=16.1$, 10.3, 8.6, and 5.4, respectively, $p < 0.01$ in all cases). Finally, the areas inside the fish-tail like order effect patterns tend to become systematically smaller with the training. Similar techniques used in Chapter 3 are used here again to estimate the areas. The estimated areas against the 4 blocks are plotted in Figure 4.11. The tendency to decrease is statistically significant (35.2, 21.0, 15.3, and 10.8, from block1 to block4, respectively, $F(3, 117)=189.81$, $p<0.001$). This pattern indicates that as participants gained more experience they became less sensitive to new evidence and fluctuated less in their belief judgments.

Figure 4.10. The belief revision patterns in the 1:1 base rate condition. Based on the simulation data from the four evaluation blocks (block1 to block4).

Figure 4.11. The area estimates across the four training blocks. Based on the simulation data from the four evaluation blocks (block1 to block4).

The belief revision patterns for the 3:1 base rate condition are shown in Figure 4.12. First, the average initial belief judgment (i.e., before any evidence) tends to increase with training (58.8, 63.9, 67.2, and 69.0, from block1 to block4, respectively). Statistics show the difference is significant ($F(3, 117)=1476.08$, $p<0.001$). Second, the magnitudes of the order effects decrease from block1 to block4 (63.8, vs 54.7, 67.9 vs 64.4, 70.2 vs 68.6, 70.2 vs 68.9, respectively; the differences are 9.1, 3.5, 1.6, and 1.3, respectively). Due to small standard deviations, however, statistics show that all order effects are significant ($t(39)=30.0, 20.5, 17.4$, and 15.3 , respectively, $p < 0.01$ in all cases). Finally, a similar pattern of decreasing areas inside the fish-tails was found. The same procedure mentioned above is used to estimate the areas and the results are plotted in Figure 4.11. The tendency to decrease is statistically significant (30.7, 19.0, 14.1, and 11.9, from block1 to block4, respectively, $F(3, 117)=794.51$, $p<0.001$).

Figure 4.12. The belief revision patterns in the 3:1 base rate condition. Based on the simulation data from the four evaluation blocks (block1 to block4).

The belief revision patterns for the 1:3 base rate condition are shown in Figure 4.13. First, different from previous conditions, the average initial belief judgment (i.e., before any evidence) tends to decrease with training (41.4, 36.7, 33.6, and 32.7, from block1 to block4, respectively). Statistics show the decrement is significant ($F(3, 117)=2933.58$, $p<0.001$). Second, the magnitudes of order effects decrease from block1 to block4 (47.8 vs 36.7, 39.8 vs 33.7, 35.1 vs 31.5, and 33.7 vs 31.2, respectively; the differences are 11.1, 6.1, 3.6, and 2.5, respectively). Due to small standard deviations, however, statistics show that all order effects are significant ($t(39)=39.0, 26.3, 21.1$, and 21.8 , respectively, $p < 0.01$ in all cases). Finally, a similar pattern of decreasing areas inside the fish-tails was found. The same procedure mentioned above is used to estimate the areas and the results are plotted in Figure 4.11. The tendency to decrease is statistically significant (48.7, 37.0, 26.8, and 17.3, from block1 to block4, respectively, $F(3, 117)=193.66$, $p<0.001$).

For the purpose of better comparison of the belief revision pattern, changes in three different base rate conditions, data from the two critical evidence sequences (+- and -+) in all three base rate conditions are combined in Figure 4.14. Both the statistical tuning and the disappearance of order effects are evident. The three sequences of belief revision patterns for the three base rate conditions tend to diverge with the training, approaching the corresponding built-in probabilities. In addition, in all three base rate conditions, from block0 to block4, the areas inside the fish-tail patterns are getting smaller and order effect tends to disappear.

Figure 4.13. The belief revision patterns in the 1:3 base rate condition. Based on the simulation data from the four evaluation blocks (block1 to block4).

Figure 4.14. The belief revision patterns in all three base rate conditions. Based on the simulation data from all five evaluation blocks (block0 to block4). Only data from inconsistent evidence sequences (i.e., -+ and +-) are showed.

A combined comparison between the model's belief revision patterns, shown from Figure 4.10 to Figure 4.14, and human participants' belief revision patterns, shown from Figure 3.8 to Figure 3.12, reveals remarkable similarities. In both cases, from block1 to block4, the finely tuned initial beliefs, the decrements of order effects, and the decrements of the areas inside the fish-tail patterns, are evident.

It is worth pointing out that analyses were conducted to explore the possible effect of training order on belief revision patterns, as did in Chapter 3. The results show that the training order has no significant effects on belief revision and order effects, consistent with the experimental results (see Appendix F for details).

4.4. Discussion

In summary, the results from the UECHO simulation match the experimental results remarkably well, supporting UECHO as a model of human belief revision. First of all, UECHO is able to gradually tune itself to capture the statistical structures in the environment, including the base rate information and various conditional probabilities. Second, the dynamic belief revision mechanism enables UECHO to capture the order effect. Finally, through confidence management UECHO is capable of capturing the changes of belief revision patterns at different experience levels. The close match between the simulation results and the experimental results in the decrement and disappearance of order effects with the increase of experience suggests that UECHO's confidence management mechanisms are necessary and adequate.

Though in general the modeling results and the experimental results match remarkably well, it is important to note several places of mismatch. As mentioned previously, human participants usually show a general tendency of conservatism. This is evident in Figure 3.3. Belief judgments did not reach the theoretical values even in block 4. This is especially true in cases where the theoretical values are extreme. On the contrary, UECHO tends to make extreme judgments. As suggested in Figure 4.7, in the 3:1 base rate condition, UECHO made friendly final decisions 89% of the time in training block 4, which is significantly greater than the 75% theoretical value. Similarly, in the 1:3 base rate condition, UECHO made friendly final decisions only 15% of times in training block 4 while the theoretical value is 25%.

Once again, properly distinguishing theoretical values and optimal values is important. Although it is clear that UECHO is far away from “probability-matching” the theoretical values, it seems that UECHO made decisions following an optimal strategy. Take the 3:1 base rate condition for an example. As discussed earlier, the first optimal strategy would be ignoring evidence and making decision simply based on base rate information. Following this strategy, one would make friendly decisions 100% of time. The second optimal strategy is cue-sensitive. That is, one makes decisions with respect to different evidence combinations. Following this strategy, one would make 100% of friendly decisions when evidence sequences are “++”, “+-”, and “-+”; and make 0% of friendly decisions when evidence sequence is “--”. Overall, this strategy results in about 81% of friendly decisions. It seems that the second optimal strategy describes UECHO’s behavior reasonably well. The proportions of friendly decisions (0.71, 0.80,

0.84, and 0.89, from block1 to block4, respectively) approximate 81% better than they approximate 100% (the first optimal strategy) or 75% (the theoretical value). In addition, this hypothesis is strongly supported by the decision results separated by different evidence combinations (see Figure 3.6). In all but two conditions, UECHO showed exactly the optimal behavior. The two exceptions are, evidence sequence "--" in the 3:1 base rate condition, and evidence sequence "++" in the 1:3 base rate condition. In these two conditions, UECHO showed a little conservatism relative to the optimal strategy.

Another mismatch is revealed in the belief revision patterns resulting from the consistent evidence sequences. In this case, human participants usually made a relatively big further revision in the original direction after receiving the second piece of evidence. This is different from the results of the UECHO simulation. As shown in Figure 4.9, UECHO only made a relatively small further revision after receiving the second piece of evidence. In this case, it seems that UECHO shows conservatism.

As mentioned previously, UECHO, similar to ECHO, has many free parameters. These parameters give UECHO a large number of degrees of freedom. Although it by no means implies that UECHO can produce arbitrary behaviors, sometimes it does suggest that the parameter space needs to be finely explored in order to further understand the mismatches.

A comparison between Figure 3.3 and Figure 4.6 indicates the simulation data have much smaller variance. In the modeling work described above, the only variance that UECHO can show comes from the random order of trial presentation. In other words, all 40 simulated participants in each base rate condition were completely identical except

they performed trials in random order. This is different from the experiment with human participants, where individual differences were usually significant. The lack of other sources of random variance in the simulation is also responsible for the surprising statistical significance in the order effect testing. Even when the effect size is really small and thus not of practical significance, due to the small variance, it still produces statistical significance. In this sense, adding some random noise to UECHO's internal states to simulate individual differences might be a sensible thing to do.

A careful examination of equation 4.7 indicates that the probability measure p is only dependent upon the activation value a . As a result, it is possible that the sum of probabilities of complementary hypotheses is not equal to 1. For example, in the previous CIC network, it is possible that $p(F)+p(H) \neq 1$. Note that in many cases, $a(F) \approx -a(H)$, which results in that $p(F)+p(H)=1$. However, in some other cases, when the activation values of competing hypotheses do not follow that pattern, it can happen that the sum of probabilities is not equal to 1. In these cases, a normalization may be necessary.

The role that the decay parameter d (see Equation 4.5) plays in UECHO raises many concerns. Remember that parameter d decays the DE connections (i.e., from evidence nodes to SEU) with time. It serves the purpose of distinguishing evidence items that become available at different time points. More specifically, while the earlier evidence items have decayed DE connections, the most current evidence items obtain full DE connections. Parameter d happens (when d is less than 1) to play a role in the existence of the recency effect in the initial stage of training. It is this role that raises questions.

First of all, it may seem to people that it is this decay parameter that causes the recency effect. This is not the case, since decay does not ensure the recency effect. The same decay exists in the later stage of training, but the recency effect disappears. Second, earlier research suggests that memory decay is dissociated from the recency effect (e.g., Anderson & Hubert, 1963), does UECHO contradict this finding? The answer is no. Again, decay in UECHO is not the reason for the recency effect. In UECHO, it is not decay that causes the recency effect. Finally, Echo is claimed to prescribe what people should do, does the introduction of decay to UECHO destroy this property? I would argue – not really. Decay is a requirement of various cognitive constraints, such as the working memory capacity. It has been argued that memory decay is indeed a rational thing to do since in general information loses value with time (see J. R. Anderson, 1991).

The simultaneous demonstrations of the existence of the order effect and the disappearance of the order effect in a single UECHO framework of human belief revision has significant implications. The order effect exists when the relevant experience is scarce, and the order effect disappears when the relevant experience becomes rich. When the relevant experience is rich one has confident expectations and eliminates the need to over-react. When the relevant experience is scarce one has to sufficiently appreciate every single piece of information, because its relevance cannot be easily and accurately determined. In this sense, both the existence and the disappearance of the order effect are rational.

CHAPTER 5

GENERAL DISCUSSION AND CONCLUSION

Human belief and human belief revision are ubiquitous in everyday life and scientific discovery. The order effect, a phenomenon in which the final belief is significantly affected by the temporal order of evidence presentation even when the temporal order does not carry meaningful information, is a robust empirical finding in human belief revision. Normative theories, such as Bayes' theorem, have no room for the order effect since it simply violates the fundamental law of commutativity. Consequently, the order effect is generally regarded as a manifestation of human biases and an indication of human irrationality. It is the goal of this dissertation research to study how the order effect occurs and seek the rational foundations of the order effect.

A large body of previous research on human belief and human belief revision, in the fields of epistemology, artificial intelligence, psychology, and cognitive science, has provided both motivations and solid background knowledge for the current research to achieve such a goal.

First of all, decades of philosophical inquiries of human belief and human belief revision have revealed that human belief has a coherence foundation. The coherence foundation of human belief implies that beliefs can survive without solid foundational evidence. Beliefs co-exist as a system by supporting each other. The acceptability of any belief is potentially dependent on its relation to a whole complex of other beliefs. As long as one's beliefs are consistent and coherent with each other, they are justified. Maintaining the global system coherence is the ultimate justification for any belief's revision, acceptance, or rejection.

Second, the theory of subjective probability claims that the degree of belief can be quantitatively summarized by a single probability number. However, the distinction between uncertainty and imprecision and the notion of ignorance suggest that human belief has a complex multi-component structure, which is difficult to capture by a single number of probability. Research has long shown that besides the probability component, human belief has another intrinsic component – confidence. While the probability component of a belief expresses the relative degree of belief, which is often the sole meaning in the usual use of the term “belief”, it is the confidence component of a belief that represents how easily the belief can be revised and altered. The confidence component of a belief is mainly determined by the amount of previous experience. A belief established with little previous experience has a low level of confidence and is easy to alter, and a belief established with much previous experience has a high level of confidence and is hard to alter. Any thorough theory of human belief revision has to take the probability/confidence distinction into consideration.

Third, the environment always dynamically changes, so does human belief. Much previous research has demonstrated that human cognition in general and human belief in particular is constantly reshaped and tuned to reflect the statistical structures of the environment. Such a tuning enables people to learn from experience and improve their confidence in dealing with future problems.

In summary, previous research leads to the conclusion that human belief is coherent, complex, and adaptive. Such a conclusion motivates and guides both the experimental study and the computational modeling work described in the dissertation.

The experiment presented in Chapter 3 was designed to study how the different levels of experience (confidence) affect the belief revision patterns. A learning paradigm was naturally adopted. Participants' belief revision patterns were evaluated five times during the whole process of training. The statistical structure of the environment in terms of base rate information was manipulated. The results are quite interesting. First, the results show that the initial belief judgments (when no evidence has been presented) gradually approach the base rate with training, indicating that participants could gradually tune themselves to capture the fundamental base rate information. Second, though order effects exist at the early stage of the training process, they tend to decrease and disappear as the training continues to the later stage. Third, new evidence causes dramatic belief revision at the early stage of the training. Though the statistical structure of the training environment is kept constant during the training, the belief revision tends to be less dramatic at the later stage. In other words, participants tend to become less sensitive to

new evidence as the training goes on. Overall, the experimental results are consistent with the predictions.

The UECHO model presented in Chapter 4 is an indispensable part of the current investigation. The model attempted to model the experimental results and further expand the conclusions.

UECHO is a significant extension of ECHO, a connectionist model of human abductive reasoning. Similar to ECHO, UECHO takes the coherence foundation of human belief as one of its foundations. Compared to ECHO, UECHO makes three major changes to further incorporate the multi-component structure and the adaptive characteristic of human belief. First, a sequential belief revision mechanism is added to enable UECHO to dynamically revise beliefs. Second, a learning mechanism based on the Rescorla-Wagner rule is developed to enable UECHO to learn from experience and tune itself to capture the statistical structures of the environment. Third, a confidence dimension has been incorporated into UECHO. As a result, UECHO is expected to capture the sensitivity to the amount of previous experience in human belief revision.

The UECHO simulation results match the experimental results remarkably well. The simulation results reveal all the three major experimental findings. First, UECHO is able to learn from experience and gradually capture the theoretical base rate information. Second, the decrement and the disappearance of order effects with the training is demonstrated. Finally, compared to the early stage of the training, much less sensitivity to new evidence is revealed at the later stage of training, indicating the effect of the confidence management. Overall, the UECHO modeling results reveal and confirm the

following findings, 1) human belief has a coherence foundation; 2) human belief is adaptive; and 3) experience (confidence) plays an important role in human belief revision.

It is important to note the mismatches between the experimental results and the simulation results. The most significant mismatch comes from the final decision behavior. It seems that human participants followed an approximate probability matching strategy and simulated participants followed a cue-sensitive optimal strategy. As a result, simulated participants made more extreme decisions. This mismatch suggests that further tuning of the UECHO model is necessary.

One essential part of the model tuning is to add noise to the simulation process. Noise adds more variation to the model so it may partially cure the extremism in UECHO's decision behavior. In addition, more variation may also help to produce better fits in terms of belief revision patterns.

It should be clear that this dissertation research, including both the experimental study and the UECHO model, focuses only on the recency effect. It would be of great interest to explore how it can be tuned or extended to model other types of order effects, such as the primacy effect. Whether it can model the full range of order effects using the same mechanism is a strong test for UECHO as a general model of human belief revision.

As mentioned earlier, Bayesian belief network is a well established technique for artificial belief management and has been used in a wide range of application domains. It is interesting to study the difference between the Bayesian belief network and UECHO. Perhaps the most significant difference between the two is that UECHO takes a

coherence view of belief management and Bayesian belief network takes a foundational view. As a result, while in a UECHO network links are indirectional and circles are allowed, in a Bayesian belief network links are directions and the network must be acyclic. Despite the difference, it is important to for the two to learn from each other. For example, Bayesian learning techniques have been widely used in Bayesian belief networks, and they are also used in UECHO.

The key hypothesis of this dissertation research is that confidence plays an important role in human belief revision. More specifically, it has shown that as confidence increases the order effect disappears. However, as has been pointed out, the notion of confidence is not the only plausible explanation for the results. Other explanations are also possible. For example, it may be argued that the disappearance of order effects is simply due to learning (i.e., tuning to capture the statistical structures of the environment), and has nothing to do with confidence. This argument is problematic because 1) numerous previous studies have shown that human belief has a multi-component structure; 2) the experimental results show a correlation between the increment of confidence and decrement of the order effect; and 3) it is the confidence management techniques adopted in UECHO that enable UECHO to capture most of the essential data patterns. However, I do agree that a strong test of the role of confidence in human belief revision would require that we go beyond this kind of correlation test and study the causal relations between the two.

Finally, how do the UECHO modeling results reveal the rational basis of order effects? First of all, the fact that UECHO, constructed based on rational postulates and

intended to prescribe what people should do, naturally shows order effects (when the confidence level is low) convincingly “debiases” order effects. Second, the existence of order effects has ecological implications. UECHO reveals that order effects appear when the relevant experience is scarce, and order effects disappear when the relevant experience becomes rich. When the relevant experience is rich, one has confident expectations, which eliminate the need to over-react. When the relevant experience is scarce, one has to sufficiently appreciate every single piece of information since its relevance cannot be easily and accurately determined in the first place. In this sense, both the existence and the disappearance of the order effect are rational.

In conclusion, human belief has a coherence foundation, a multi-component structure, and an adaptive characteristic. Confidence plays an important role in human belief revision. Order effects in human belief revision occur where confidence is low and disappear when confidence goes up. UECHO provides a computational model of human belief revision and order effects.

APPENDIX A

EXPERIMENT INSTRUCTIONS

A.1. General Instructions

Welcome to the experiment. Please listen carefully to the following instructions.

In this experiment, you will act as a captain of a US Navy warship. As a captain, your responsibilities are many. One of them is to accurately identify the true intention of each target in your region. This decision is important because it is critical to the protection of your ship and its crew, and the maintenance of peaceable US relations in politically volatile regions. Your task in this experiment is to learn how to make such decisions. At the end of the experiment, you are expected to be an expert to make such decisions.

What you see on your screen is your radar system. The black region is your radar screen and you are positioned at the red spot in the middle. In each of the training trials you will see an unidentified target shown somewhere on your radar. Your task is to collect some relevant information and make a decision about its identity.

In this experiment, there are two pieces of information you can collect. One is the air route. You can check if the unknown target is on or off a commercial air route. Usually a target being on a commercial air route suggests it is a commercial airplane, although this may not always be the case. You check the route information by clicking the “check route” button on the right of your radar. The second piece of information you can collect is the self-identification. That is, you can send a radio warning to the target and ask it to identify itself. A hostile aircraft will usually ignore you and not send any response to you. A commercial aircraft will normally respond to you and identify itself as a commercial airplane. However, this may not always be the case. You send a radio warning by clicking the “send radio” button. After you have collected both pieces of information, you have to make a decision about the target’s true identity. Click “F” to say you think the target is friendly, and click “H” to say you think the target is hostile.

How do you get trained? You get trained by being told the true identity of each target you are working on. Sometimes you may find that you have made the right decision, and some other times you may find that you have made the wrong decision. Through this kind of feedback you learn to improve your performance. You can find out the true identity of the target by clicking the “true identity” button at the low-right corner.

In a real situation, you won’t get the two pieces of information at the same time. Instead, you get them one by one. This is also the case in this experiment. You get the route and radio information sequentially, one at a time. In some trials, you get the route information first and radio second. In some other trials, you get the radio information first and route second.

As a result of this sequential information acquisition, your belief of the target's intention may also change as you get more and more information about it. This belief revision is an important ability of a successful decision-maker. In this experiment, your belief revision process will also be recorded. This is done through the two scales at the lower-left corner. In the top scale, you are asked the question, "how likely do you think it is that this target is friendly"? You answer this question on a 0 to 100 scale. 0 means you completely believe it is hostile, and 100 means you completely believe it is friendly. You click anywhere in between to indicate your degree of belief. In the lower scale, you are asked, "how confident are you of your judgment"? You also answer this question on a 0 to 100 scale. 0 means you have no confidence in your belief judgment at all, and 100 means you are completely confident in your belief judgment. Make sure you understand the distinctions between the two. For example, you may have a belief judgment of 100 and a confidence judgment of 50, which means you believe for sure that the target is friendly, but you only have average confidence in it. As another example, you may have a belief judgment of 50 but a confidence judgment of 90, which indicates that you are very confident in your belief that the target has only a half chance to be friendly. Does anybody have a question? You will be asked to provide your belief and confidence judgments 3 times in each trial. They are, before you collect any information, after you collect the first piece of information, and after you collect both pieces of information.

In each trial, the order of the information collection and your reaction are pre-fixed. You must follow the order. The order is indicated by color. At any one time point, you can only work on one action, which is the element with the red color. All other actions

are disabled, which are in gray color. As you may notice, at this time, “new game” is your only available option, since it is red. All others are disabled. In order to go on, you must click “new game”. After you click “new game”, the next available option will become active and red. Please follow this red-leading action order. After you have finished the belief and confidence judgments, click the corresponding ok button to go on.

If you understand these instructions, please click the “new game” to go on. If you have any questions, please raise your hand.

A.2. Instructions before the initial evaluation sub-block

In the next 8 trials, you follow the same basic procedures as stated in the instructions except that you won't get any feedback about the true identity of each target. Since you haven't begun any training, try to perform these trials based upon any prior experience you may have. Press the button below to begin.

A.3. Instructions before the other evaluation sub-blocks

After the training trials so far, you have gained a lot of experiences about the environment you are working in. Please perform the following 8 trials based upon the experiences you have gained. Note you won't get any feedback about the true identity of each target. Also note that the order of evidence collection may be different. You should treat these trials as a test of your training performance. You may also treat these questions as if they are asked by a naive captain. You, the well-trained and experienced

captain, are teaching him/her by giving judgment based upon your training experiences.

Press the button below to begin.

APPENDIX B

CONDITIONAL INDEPENDENCE

Two sets of variables, A and B, are said to be (conditionally) independent given a third set C of variables if when the values of the variables C are known, knowledge about the values of the variables B provides no further information about the values of the variables A. Quantitatively:

$$P(A|B,C)=P(A|C)$$

In the current design (see Table 3.1), Route and SelfID are conditionally independent given identity, in all three base rate conditions. More specifically,

$$P(\text{SelfID+}|\text{Route+}, \text{friendly}) = P(\text{SelfID+}|\text{Route-}, \text{friendly}) = P(\text{SelfID+}|\text{friendly}) = 2/3,$$

$$P(\text{Route+}|\text{SelfID+}, \text{friendly}) = P(\text{Route+}|\text{SelfID-}, \text{friendly}) = P(\text{Route+}|\text{friendly}) = 2/3;$$

$$P(\text{SelfID+}|\text{Route+}, \text{hostile}) = P(\text{SelfID+}|\text{Route-}, \text{hostile}) = P(\text{SelfID+}|\text{hostile}) = 1/3,$$

$$P(\text{Route+}|\text{SelfID+}, \text{hostile}) = P(\text{Route+}|\text{SelfID-}, \text{hostile}) = P(\text{Route+}|\text{hostile}) = 1/3.$$

APPENDIX C

SOME DETAILS OF THOSE INVALID PARTICIPANTS

32 (about 23%) participants are considered as invalid due to lack of cooperation, which is not uncommon among Psychology100 students who participate in experiments for partially fulfilling course requirements. There are three major criteria for invalidation, all based on the experimenter's observations: 1) participants who dozed off or fell into sleep during the experiment; 2) participants who showed no interest in the experiment by engaging in other activities, including constantly talking to each other or reading newspaper; and 3) participants who used tricks (i.e., mechanically clicking the mouse and give each judgment the same value no matter what) in order to finish the experiment as quickly as possible. The number of participants excluded based on each criterion is listed in Table C.1. Examples of data patterns shown by eight participants who used tricks are shown in Figure C.1, in which the initial belief judgments are plotted along the 144 training trials. It is clear that the patterns are consistent with the experimenter's observations that these participants used tricks in the experiment.

	Base Rate			Total
	1:1	3:1	1:3	
1) dozing off or falling into sleep	2	1	3	6
2) engaging in other activities	3	4	2	9
3) using tricks	6	7	4	17
Total	11	12	9	32

Table C.1. The number of participants excluded from the data analysis due to various reasons.

Figure C.1. Examples of invalid participants who used tricks. The initial belief judgments are plotted against the 144 training trials.

APPENDIX D

UECHO PARAMETERS

D.1. Existing Parameters in ECHO

echo_N_EL is the maximum number of nodes in the model.

set echo_N_EL 60

echo_DATA_EXCITATION is the default connection strength from evidence nodes to the special evidence node (SEU).

set echo_DATA_EXCITATION 0.10

echo_EXCITATORY_WEIGHT is the default connection strength for a coherent relation.

set echo_EXCITATORY_WEIGHT 0.04

echo_INHIBITORY_WEIGHT is the default connection strength for an incoherent relation.

set echo_INHIBITORY_WEIGHT 0.04

echo_START_VALUE is the default activation value when the network is reset.

set echo_START_VALUE 0.01

echo_STOP_VALUE is the default threshold value. The network stops updating when no change is larger than this value.

set echo_STOP_VALUE 0.001

echo_THETA is the parameter of skepticism.

set echo_THETA 0.10

echo_ANALOGY_IMPACT is the parameter of analogy impact.

set echo_ANALOGY_IMPACT 1.00

```

### echo_MAX and echo_MIN limits the range of activation values.
set echo_MAX          1.00
set echo_MIN          -1.00

```

D.2. New Parameters in UECHO

```

### echo_DATA_DECAY regulates the decay of earlier evidence item
### (see Equation 4.5).
set echo_DATA_DECAY    0.4

```

```

### echo_ALPHA and echo_BETA are parameters in the Rescorla-Wagner rule
### (see Equation 4.6)
set echo_ALPHA         0.3
set echo_BETA          0.4

```

```

### echo_EXPTHETA is the parameter  $\tau$  used to convert from activation to probability
### (see Equation 4.7)
set echo_EXPTHETA      -3.2

```

```

### echo_THETADECAY is the parameter  $b$  used to tune echo_THETA
### (see Equation 4.11)
set echo_THETADECAY    -0.032

```

```

### echo_PRIOR is the parameter  $l$  in learning base rate information for hypothesis
### nodes (see Equation 4.15)
set echo_PRIOR         50.0

```


APPENDIX E

A SAMPLE OF UECHO SIMULATION TRACE

The following is a sample running trace from the UECHO simulation. The command is in the format of "e_traindisxx y z", where xx is the base rate condition ("11", "31", or "13"), y is the number of simulated subjects, and z is the number of blocks each simulated subject performs. For example, "e_traindis11 2 2" represents that two subjects are simulated in the 1:1 base rate condition, and 2 blocks of trials are performed by each subject.

```
% e_traindis11 2 2
```

```
SUBJECT# 1, TRAINING BLOCK# 1
```

```
There are 2 hypotheses. They are F, H.
```

```
There are 4 data points. They are E1, NE1, E2, NE2.
```

```
There are 36 training examples.
```

```
Training trial 1...
```

```
Looking for competing hypotheses...
```

```
iteration 4 change= 0.000891
```

SEU	E1	NE1	E2	NE2	F	H
1.00	0.01	0.01	0.01	0.01	0.00	0.00

```
Data set no. 1 is: NE1 with strength 1.000;
```

```
And here is the run.
```

```
Looking for competing hypotheses...
```

```
iteration 61 change= 0.000968
```

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.37	0.60	-0.32	0.32	-0.45	0.45

Data set no. 2 is: NE2 with strength 1.000;
Here is the final run.

Looking for competing hypotheses...

iteration 21 change= 0.000923

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.36	0.48	-0.38	0.60	-0.48	0.48

Now, Echo should update itself, and here is how....
output node F should be -1.000, is -0.478 (p=0.178).
output node H should be 1.000, is 0.478 (p=0.822).
after 1 trials, theta is 0.0992
after 1 trials, HE_beta is 0.5000
after 1 trials, HE_alpha is 0.9975

Training trial 2...

Looking for competing hypotheses...

iteration 8 change= 0.000886

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.00	0.01	-0.00	0.01	-0.01	0.01

Data set no. 1 is: E1 with strength 1.000;
And here is the run.

Looking for competing hypotheses...

iteration 72 change= 0.000955

SEU	E1	NE1	E2	NE2	F	H
1.00	0.40	-0.26	0.12	-0.20	0.20	-0.32

Data set no. 2 is: E2 with strength 1.000;
Here is the final run.

Looking for competing hypotheses...

iteration 30 change= 0.000931

SEU	E1	NE1	E2	NE2	F	H
1.00	0.27	-0.26	0.40	-0.28	0.24	-0.36

Now, Echo should update itself, and here is how....
output node F should be -1.000, is 0.237 (p=0.681).
output node H should be 1.000, is -0.365 (p=0.237).
after 2 trials, theta is 0.0984
after 2 trials, HE_beta is 0.5000
after 2 trials, HE_alpha is 0.9974

Training trial 3...

Looking for competing hypotheses...

iteration 10 change= 0.000887

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.01	0.01	-0.01	0.01	-0.02	0.01

Data set no. 1 is: NE1 with strength 1.000;
And here is the run.

Looking for competing hypotheses...

iteration 65 change= 0.001000

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.24	0.39	-0.16	0.10	-0.28	0.18

Data set no. 2 is: NE2 with strength 1.000;
Here is the final run.

Looking for competing hypotheses...

iteration 31 change= 0.000982

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.23	0.26	-0.26	0.40	-0.33	0.21

Now, Echo should update itself, and here is how....
output node F should be -1.000, is -0.335 (p=0.255).
output node H should be 1.000, is 0.214 (p=0.665).
after 3 trials, theta is 0.0976
after 3 trials, HE_beta is 0.5000
after 3 trials, HE_alpha is 0.9973

Training trial 4...

Looking for competing hypotheses...

iteration 11 change= 0.000998

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.01	0.01	-0.01	0.01	-0.03	0.02

Data set no. 1 is: E1 with strength 1.000;
And here is the run.

Looking for competing hypotheses...

iteration 72 change= 0.000990

SEU	E1	NE1	E2	NE2	F	H
1.00	0.39	-0.25	0.10	-0.17	0.18	-0.28

Data set no. 2 is: E2 with strength 1.000;
Here is the final run.

Looking for competing hypotheses...

iteration 31 change= 0.000991

SEU	E1	NE1	E2	NE2	F	H
1.00	0.26	-0.24	0.40	-0.26	0.22	-0.33

Now, Echo should update itself, and here is how....
output node F should be 1.000, is 0.216 (p=0.666).
output node H should be -1.000, is -0.333 (p=0.256).
after 4 trials, theta is 0.0968
after 4 trials, HE_beta is 0.8071
after 4 trials, HE_alpha is 0.9885

Training trial 5...

Looking for competing hypotheses...

iteration 12 change= 0.000964

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.01	0.01	-0.01	0.01	-0.02	0.02

Data set no. 1 is: NE1 with strength 1.000;
And here is the run.

Looking for competing hypotheses...

iteration 66 change= 0.000945

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.33	0.54	-0.26	0.23	-0.39	0.35

Data set no. 2 is: E2 with strength 1.000;

Here is the final run.

Looking for competing hypotheses...

iteration 153 change= 0.000981

SEU	E1	NE1	E2	NE2	F	H
1.00	0.08	0.15	0.51	-0.28	0.22	-0.25

Now, Echo should update itself, and here is how....

output node F should be 1.000, is 0.223 (p=0.671).

output node H should be -1.000, is -0.253 (p=0.308).

after 5 trials, theta is 0.0960

after 5 trials, HE_beta is 0.9056

after 5 trials, HE_alpha is 0.9733

...
...
...

Training trial 36...

Looking for competing hypotheses...

iteration 5 change= 0.000924

SEU	E1	NE1	E2	NE2	F	H
1.00	0.00	0.01	0.00	0.01	-0.01	0.01

Data set no. 1 is: E1 with strength 1.000;

And here is the run.

Looking for competing hypotheses...

iteration 68 change= 0.000987

SEU	E1	NE1	E2	NE2	F	H
1.00	0.58	-0.30	0.10	-0.10	0.26	-0.27

Data set no. 2 is: E2 with strength 1.000;

Here is the final run.

Looking for competing hypotheses...

iteration 37 change= 0.000953

SEU	E1	NE1	E2	NE2	F	H
1.00	0.42	-0.27	0.58	-0.27	0.30	-0.31

Now, Echo should update itself, and here is how....

output node F should be 1.000, is 0.296 (p=0.721).

output node H should be -1.000, is -0.314 (p=0.268).

after 36 trials, theta is 0.0740

after 36 trials, HE_beta is 0.8610

after 36 trials, HE_alpha is 0.8610

SUBJECT# 1, EVALUATION BLOCK# 1

Base Evaluation

Looking for competing hypotheses...

iteration 2 change= 0.000971

SEU	E1	NE1	E2	NE2	F	H
1.00	0.01	0.01	0.01	0.01	0.00	-0.00

Evidence Sequence E1-NE2

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
Looking for competing hypotheses...

iteration 67 change= 0.000991

SEU	E1	NE1	E2	NE2	F	H
1.00	0.59	-0.28	0.10	-0.09	0.26	-0.26

Data point no. 2, with data_excitatory 0.100, added at time_step 1.
But how about previous data points?...

Data point no. 1, with data_excitatory 0.040, added at time_step 0.
Looking for competing hypotheses...

iteration 77 change= 0.000982

SEU	E1	NE1	E2	NE2	F	H
1.00	0.37	-0.18	-0.18	0.55	0.07	-0.07

Evidence Sequence E2-NE1

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
Looking for competing hypotheses...

iteration 79 change= 0.000983

SEU	E1	NE1	E2	NE2	F	H
1.00	0.13	-0.13	0.57	-0.24	0.22	-0.22

Data point no. 2, with data_excitatory 0.100, added at time_step 1.
But how about previous data points?...

Data point no. 1, with data_excitatory 0.040, added at time_step 0.
Looking for competing hypotheses...

iteration 90 change= 0.000983

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.25	0.57	0.32	-0.10	-0.15	0.15

Evidence Sequence NE1-E2

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
Looking for competing hypotheses...

iteration 68 change= 0.000974

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.28	0.58	-0.10	0.09	-0.25	0.25

Data point no. 2, with data_excitatory 0.100, added at time_step 1.
But how about previous data points?...

Data point no. 1, with data_excitatory 0.040, added at time_step 0.
Looking for competing hypotheses...

iteration 87 change= 0.000988

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.16	0.36	0.55	-0.20	-0.02	0.03

Evidence Sequence NE2-E1

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
Looking for competing hypotheses...

iteration 81 change= 0.000974

SEU	E1	NE1	E2	NE2	F	H

1.00 -0.13 0.12 -0.25 0.57 -0.21 0.21
 Data point no. 2, with data_excitatory 0.100, added at time_step 1.
 But how about previous data points?...
 Data point no. 1, with data_excitatory 0.040, added at time_step 0.
 Looking for competing hypotheses...
 iteration 87 change= 0.000970

SEU	E1	NE1	E2	NE2	F	H
1.00	0.58	-0.25	-0.09	0.32	0.17	-0.17

Evidence Sequence E1-E2

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
 Looking for competing hypotheses...
 iteration 67 change= 0.000991

SEU	E1	NE1	E2	NE2	F	H
1.00	0.59	-0.28	0.10	-0.09	0.26	-0.26

Data point no. 2, with data_excitatory 0.100, added at time_step 1.
 But how about previous data points?...
 Data point no. 1, with data_excitatory 0.040, added at time_step 0.
 Looking for competing hypotheses...
 iteration 37 change= 0.000955

SEU	E1	NE1	E2	NE2	F	H
1.00	0.42	-0.25	0.58	-0.25	0.30	-0.30

Evidence Sequence E2-E1

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
 Looking for competing hypotheses...
 iteration 79 change= 0.000983

SEU	E1	NE1	E2	NE2	F	H
1.00	0.13	-0.13	0.57	-0.24	0.22	-0.22

Data point no. 2, with data_excitatory 0.100, added at time_step 1.
 But how about previous data points?...
 Data point no. 1, with data_excitatory 0.040, added at time_step 0.
 Looking for competing hypotheses...
 iteration 38 change= 0.000989

SEU	E1	NE1	E2	NE2	F	H
1.00	0.59	-0.28	0.40	-0.21	0.30	-0.30

Evidence Sequence NE1-NE2

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
 Looking for competing hypotheses...
 iteration 68 change= 0.000974

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.28	0.58	-0.10	0.09	-0.25	0.25

Data point no. 2, with data_excitatory 0.100, added at time_step 1.
 But how about previous data points?...
 Data point no. 1, with data_excitatory 0.040, added at time_step 0.
 Looking for competing hypotheses...
 iteration 37 change= 0.000952

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.25	0.42	-0.26	0.58	-0.29	0.29

Evidence Sequence NE2-NE1

Data point no. 1, with data_excitatory 0.100, added at time_step 0.
Looking for competing hypotheses...

iteration 81 change= 0.000974

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.13	0.12	-0.25	0.57	-0.21	0.21

Data point no. 2, with data_excitatory 0.100, added at time_step 1.
But how about previous data points?...

Data point no. 1, with data_excitatory 0.040, added at time_step 0.
Looking for competing hypotheses...

iteration 39 change= 0.000984

SEU	E1	NE1	E2	NE2	F	H
1.00	-0.29	0.59	-0.22	0.39	-0.30	0.30

SUBJECT# 1, TRAINING BLOCK# 2

...
...
...

SUBJECT# 2, TRAINING BLOCK# 1

...
...
...

SUBJECT# 2, TRAINING BLOCK# 2

...
...
...

APPENDIX F

THE EFFECT OF TRAINING ORDER

The belief revision patterns are further analyzed based on whether the evidence collection order in evaluation is consistent with the evidence collection order in training. These results are shown from Figure F.1 to Figure F.4. Statistics show no new findings except those that have been reported in Chapter 3 and Chapter 4 where no such distinction is made. In other words, no effect of training order is found.

Figure F.1. The belief revision patterns in all three base rate conditions. Based on the experimental data when the training order is consistent with the evaluation order.

Figure F.2. The belief revision patterns in all three base rate conditions. Based on the experimental data when the training order is inconsistent with the evaluation order.

Figure F.3. The belief revision patterns in all three base rate conditions. Based on the simulation data when the training order is consistent with the evaluation order.

Figure F.4. The belief revision patterns in all three base rate conditions. Based on the simulation data when the training order is inconsistent with the evaluation order.

REFERENCES

- Achinstein, P. (1989). Explanation and acceptability. *Behavioral and Brain Sciences*, 12(3), 467-468.
- Ackermann, R. (1972). *Belief and knowledge*. Garden City, NY: Doubleday & Company.
- Adelman, L., Bresnick, T., Black, P. K., Marvin, F. F., & Sak, S. G. (1996). Research with patriot air defense officers: Examining information order effects. *Human Factors*, 38(2), 250-261.
- Adelman, L., & Tolcott, M. A. (1993). Examining the effect of information order on expert judgment. *Organizational Behavior and Human Decision Processes*, 56, 348-369.
- Alchourron, C., Gardenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction functions and their associated revision functions. *Journal of Symbolic Logic*, 50, 510-530.
- Almond, R. G. (1995). *Graphical belief modeling*. London: Chapman & Hall.
- Anderson, J. A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396-408.

- Anderson, N. H. (1981). *Foundations of information integration theory*. New York, NY: Academic Press.
- Anderson, N. H. (1991). *Contributions to information integration theory, Volume I: Cognition*. Hillsdale, NJ: Erlbaum.
- Anderson, N. H., & Farkas, A. J. (1973). New light on order effects in attitude change. *Journal of Personality and Social Psychology*, 28, 88-93.
- Anderson, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior*, 2, 379-391.
- Asch, S. E. (1946). Forming impression of personality. *Journal of Abnormal and Social Psychology*, 41, 258-290.
- Bar-Hillel, M. (1980). The base rate fallacy in probability judgments. *Acta Psychologica*, 44, 211-233.
- Bar-Hillel, M. (1983). The base rate fallacy controversy. In R. W. Scholz (Ed.), *Decision making under uncertainty*. Amsterdam: North-Holland.
- Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. London: Cambridge University Press.
- Barto, A. G. (1990). Connectionist learning for control: An overview. In W.T. Miller, III, R.S. Sutton, & P.J. Werbos (Eds.), *Neural networks for control*. Cambridge, MA: MIT Press.
- Barto, A. G., & Watkins, C. J. C. H. (1990). Learning and sequential decision making. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks*. Cambridge, MA: MIT Press.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis, (2nd ed)*. New York, NY: Springer-Verlag.
- Bhatnagar, R. K., & Kanal L. N. (1986). Handling uncertain information: A review of numeric and non-numeric methods. In L. N. Kanal & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence*. Amsterdam, The Netherlands: Elsevier Science.
- Britannica Online, <<http://www.eb.com:180/cgi-bin/g?DocF=micro/60/74.html>>.
- Cheeseman, P. (1985). In defense of probability. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*. American Association for Artificial Intelligence.

- Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.
- Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Science*, 4, 317-370.
- Cohen, P. R. (1985). *Heuristic reasoning about uncertainty: An artificial intelligence approach*. Boston, MA: Pitman Publishing.
- Cooper, G. (1990). Computational complexity of probabilistic inference using Bayesian belief networks (Research note). *Artificial Intelligence*, 42, 393-405.
- Cornman, J. W. (1978). Foundational versus nonfoundational theories of empirical justification. In G. S. Pappas & M. Swain (Eds.), *Essays on knowledge and justification*. Ithica, NY: Cornell University Press.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1-73.
- Dagum, P., & Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60, 141-153.
- De Finetti, B. (1974). *Theory of probability: A critical introductory treatment*. New York, NY: John Wiley & Sons.
- Diaconis, P., & Zabell, S. (1982). Updating subjective probability. *Journal of the American Statistical Association*, 77(380), 822-830.
- Dixon, S. E. (1994). *Belief revision: A computational approach*. Unpublished Ph.D. Thesis, Department of Computer Science, University of Sydney.
- Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, 12, 231-272.
- Doyle, J. (1992). Reason maintenance and belief revision. In P. Gardenfors (Ed.), *Belief revision*. New York, NY: Cambridge University Press.
- Dreben, E. K., Fiske, S. T., & Hastie, R. (1979). The independence of evaluative and item information: Impression and recall order effects in behavior-based impression formation. *Journal of Personality and Social Psychology*, 37, 1758-1768.
- Dubois, D. & Prade, A. (1988). *Possibility theory: An approach to computerized processing of uncertainty*. New York, NY: Plenum Press.
- Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgment under uncertainty: Heuristics and Biases*. Cambridge, NY: Cambridge University Press.

- Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3), 519-527.
- Evans, J. (1989). *Bias in human reasoning: Causes and consequences*. London: Erlbaum.
- Evans, J., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, 11(3), 295-306.
- Evans, J., Over, D. E., & Manktelow, K. I. (1993). Reasoning, decision making and rationality. *Cognition*, 49, 165-187.
- Fann, K. T. (1970). *Peirce's Theory of Abduction*. The Hague, Martinus Nijhoff.
- Gärdenfors, P. (1990). The dynamics of belief systems: Foundations vs coherence theories. *Revue Internationale de Philosophie*, 172, 24-46.
- Gärdenfors, P. (1992). Belief revision: An introduction. In P. Gärdenfors (Ed.), *Belief revision*. New York: NY: Cambridge University Press.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology*, Vol 2. Chichester, England: Wiley.
- Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives*. London: Routledge.
- Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is relevant for psychology and vice versa. In G. Wright & P. Ayton (Eds.), *Subjective probability*. New York, NY: Wiley.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky (1996). *Psychological Review*, 103(3), 592-596.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency format. *Psychological Review*, 102(4), 684-704.
- Gluck, M.A., & Bower, G.H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3), 227-247.
- Harman, G. (1973). *Thought*. Princeton, NJ: Princeton University Press.

- Hasher, L., & Zacks, R.T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, 108(3), 356-388.
- Hasher, L., & Zacks, R.T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologists*, 39(12), 1372-1388.
- Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1, 79-119.
- Hoadley, C.M., Ranney, M., & Schank, P. (1994). WanderECHO: A connectionist simulation of limited coherence. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 421-426). Hillsdale, NJ: Erlbaum.
- Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, 90, 197-217.
- Hogarth, R. M. & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1-55.
- Hovland, C. I. (1957). *The order of presentation in persuasion*. New Haven, CT: Yale University Press.
- Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach*. Peru, IL: Open Court.
- Jeffrey, R. (1968). Probable knowledge. In I. Lakatos (Ed.), *The problem of inductive logic*. Amsterdam: North-Holland.
- Jensen, F. V. (1996). *An introduction to Bayesian networks*. New York, NY: Springer Verlag.
- Josephson, J. R., & Josephson, S. G. (1994). *Abductive inference: Computation, philosophy, technology*. New York, NY: Cambridge University Press.
- Jungermann, H. (1983). The two camps of rationality. In R. W. Scholz (Ed.), *Decision making under uncertainty*. Amsterdam: North-Holland.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and Biases*. Cambridge, NY: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.
- Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, 11, 143-157.

- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103(3), 582-591.
- Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77(3), 217-273.
- Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan.
- Klayman, J. (1995). Varieties of confirmation bias. *The Psychology of Learning and Motivation*, 32, 385-418.
- Klayman, J., & Ha, Y-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2), 211-228.
- Kleiter, G. D. (1993). Natural sampling: Rationality without base rates. In G.H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology*. New York: Springer-Verlag.
- Klir, G. J. (1994). The many faces of uncertainty. In B. M. Ayyub & M. M. Gupta (Eds.), *Uncertainty modelling and analysis: Theory and application*. Amsterdam, The Netherlands: Elsevier Science.
- Kuhn, T. S. (1970). *Structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.
- Levine, D. S. (1991). *Introduction to neural and cognitive modeling*. Hillsdale, NJ: Erlbaum.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under certainty: Heuristics and bias*. New York, NY: Cambridge University Press.
- Lindley, D. (1994). Foundations. In G. Wright & P. Ayton (Eds.), *Subjective probability*. New York, NY: Wiley.
- March, J. G. (1994). *A primer on decision making: How decisions happen*. New York, NY: Free Press.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.
- MathSoft (1997). *S-PLUS 4: Guide to statistics*. Seattle, Washington: MathSoft.
- Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, 117(1), 68-85.

- Miller, N., & Campbell, D. T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurement. *Journal of Abnormal and Social Psychology*, 59, 1-9.
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3), 363-386.
- Moon, A. M., & Graybill, F. A. (1963). *Introduction to the theory of statistics*, (2nd ed). New York, NY: McGraw-Hill.
- Neapolitan, R. E. (1989). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York, NY: John Wiley & Sons.
- Netica. (1996). *Netica application for belief networks and influence diagrams: User's guide (Version 1.03)*. Norsys Software Corp. <<http://www.norsys.com>>.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Ousterhout, J. K. (1994). *Tcl and the Tk toolkit*. Reading, MA: Addison-Wesley.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann Publishers.
- Price, H. H. (1969). *Belief*. New York, NY: George Allen & Unwin Ltd.
- Quine, W. V., & Ullian, J. S. (1978). *The web of belief*. New York, NY: Random House.
- Quinlan, P. (1991). Connectionism and psychology: A psychological perspective on new connectionist research. Chicago, IL: The University of Chicago Press.
- Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naïve physics. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Read, S. J. & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3), 429-447.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York, NY: Appleton-Century-Crofts.
- Riskey, D. R. (1979). Verbal memory processes in impression formation. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 271-281.

- Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under certainty: Heuristics and bias*. New York, NY: Cambridge University Press.
- Ross, L., & Lepper, M. R. (1980). The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder (Ed.), *Fallible judgment in behavioral research: New directions for methodology of social and behavioral science, Vol. 4*. San Francisco, CA: Jossey-Bass.
- Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103(4), 734-760.
- Rumelhart, D. & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, 2 vols. Cambridge, MA: MIT Press.
- Rywick, T., & Schaye, P. (1974). Use of long-term memory in impression formation. *Psychological Reports*, 34, 939-945.
- Savage, L. J. (1972). *The foundations of statistics*. New York NY: Dover Publications.
- Schank, P., & Ranney, M. (1991). The psychological fidelity of ECHO: Modeling an experimental study of explanatory coherence. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 892-897). Hillsdale, NJ: Erlbaum.
- Schank, P., & Ranney, M. (1992). Assessing explanatory coherence: A new method for integrating verbal data with models of on-line belief revision. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 599-604). Hillsdale, NJ: Erlbaum.
- Schiller, F. C. S. (1979). *Problems of belief*. New York, NY: George H. Doran Company.
- Schlottmann A., & Anderson, N. H. (1995). Belief revision in children: Serial judgment in social cognition and decision-making domains. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5), 1349-1364.
- Schmitt, F. (1992). *Knowledge and belief*. New York, NY: Routledge.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69, 99-118.

- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1976). Cognitive processes and societal risk taking. In J. S. Carrol & J. W. Payne (Eds.), *Cognition and social behavior*. Hillsdale, NJ: Erlbaum.
- Sosa, E. (1980). The raft and the pyramid: Coherence versus foundations in the theory of knowledge. *Midwest Studies in Philosophy*, 5, 3-25.
- Stewart, R. H. (1965). Effect of continuous responding on the order effect in personality impression. *Journal of Personality and Social Psychology*, 1, 161-165.
- Surber, C. F. (1982). Separable effects of motives, consequences, and presentation order on children's moral judgments. *Developmental Psychology*, 18, 257-266.
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2), 135-170.
- Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian Reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks*. Cambridge, MA: MIT Press.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12(3), 435-502.
- Thagard, P. (1992a). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.
- Thagard, P. (1992b). Adversarial problem solving: Modeling an opponent using explanatory coherence. *Cognitive Science*, 16(1), 123-149.
- Thagard, P. (1996). *Mind: Introduction to cognitive science*. Cambridge, MA: MIT Press.
- Towne, D. (1995). *CIC: Tactical Decision Making (Version 2.0)*. Unpublished manuscript. Behavioral Technology Laboratories, University of Southern California.
- Tubbs, R. M., Gaeth, G. J., Levin, I. P., & Osdol, L. (1993). Order effects in belief updating with consistent and inconsistent evidence. *Journal of Behavioral Decision Making*, 6, 257-269.
- Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 91, 293-315.
- Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. New York, NY: Chapman and Hall.

- Wang, H., Johnson, T. R., & Zhang, J., (1998). UECHO: A model of uncertainty management in human abductive reasoning. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Wang, P. (1993). Belief revision in probability theory. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann Publishers.
- Wolterstorff, N. (1996). *John Locke and the ethics of belief*. New York, NY: Cambridge University Press.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353.
- Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1), 3-28.
- Zhang, J., Johnson, T.R., & Wang, H. (1997). The relation between order effects and frequency learning in tactical decision making. *Thinking and Reasoning*. 4(2), 123-145.

IMAGE EVALUATION TEST TARGET (QA-3)

APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved