University of Osnabrück

Institute of Cognitive Science

Utilizing Cross-Domain Cognitive Mechanisms for Modeling Aspects of Artificial General Intelligence

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Submitted By:

Ahmed Mohammed Hassan Abdel-Fattah

(Cairo, Egypt)

Supervisor:

Prof. Dr. ***Kai-Uwe Kühnberger***

Artificial Intelligence, Institute of Cognitive Science, University of Osnabrück

Osnabrück, 2014

A PhD thesis by:	Ahmed Mohammed Hassan Abdel-Fattah (born on 28–September–1976, Cairo, Egypt)
Department:	FB 8 (Cognitive Science)
Submitted:	3–February–2014
Defended:	27–March-2014
Supervisor:	Kai-Uwe Kühnberger (University of Osnabrück, Germany)
Reviewer:	Helmar Gust (University of Osnabrück, Germany)
Reviewer:	Pei Wang (Temple University, USA)

Gedruckt mit Unterstützung des Deutschen Akademischen Austauschdienstes (DAAD).

Nomenclature

ACRONYMS:

AGI	Artificial general intelligence
AI	Artificial intelligence
CB	Conceptual blending
CFB	Counterfactual blend (space)
CFC	Counterfactual conditional
CogSci	Cognitive science
CS	Computer science
GI	General intelligence
GOFAI	Good old-fashioned AI
HDTP	Heuristic-driven theory projection
HELCO	High entrenchment level concept
KB	Knowledge base
KR	Knowledge representation
KRR	Knowledge representation and reasoning
LEVCO	Low entrenchment level concept
NARS	Non-axiomatic reasoning system
SME	Structural mapping engine
SMT	Structure-mapping theory

NOTATIONS:

$\langle \square_1, \square_2 \rangle$	Ordered pair
$(\square_1, \square_2)$	Tuple / Unordered pair
b (or b_i)	Conceptual entities / Beliefs
B (or B_i)	Blend concept
e_V	Function for entrenchment values and levels
F^c	Representation of an arbitrary concept c
G	Generalization concept
$\mathbb{K}_B$	Knowledge base of conceptual entities
$\mathbb{K}_C$	Set of concept names / Lexicon
S	Source concept
T	Target concept

Abstract

In this era of increasingly rapid availability of resources of all kinds, a widespread need to characterize, filtrate, use, and evaluate what could be necessary and useful becomes a crucially vital everyday task. Neither research in the field of artificial intelligence (AI) nor in cognitive science (CogSci) is an exception (let alone within a crossing of both paths). A promised goal of AI was to primarily focus on the study and design of intelligent artifacts that show aspects of human-like general intelligence (GI). That is, facets of intelligence similar to those exhibited by human beings in solving problems related to cognition. However, the focus in achieving AI's original goal is scattered over time. The initial ambitions in the 1960s and 1970s had grown by the 1980s into an "industry", where not only researchers and engineers but also entire companies developed the AI technologies in building specialized hardware. But the result is that technology afforded us with many, many devices that allegedly work like humans, though they can only be considered as life facilitators (if they even do). This is mainly due to, I propose, basic changes on viewing what true essences of intelligence should have been considered within scientific research when modeling systems with GI capacities.

A modern scientific approach to achieving AI by simulating cognition is mainly based on representations and implementations of higher cognition in artificial systems. Luckily, such systems are essentially designed with the intention to be acquired with a "human-like" level of GI, so that their functionalities are supported by results (and solution methodologies) from many cognitive scientific disciplines. In classical AI, only a few number of attempts have tried to integrate forms of higher cognitive abilities in a uniform framework that model, in particular, cross-domain reasoning abilities, and solve baffling cognition problems —the kind of problems that a cognitive being (endowed with traits of GI) could only solve. Unlike classical AI, the intersection between the recent research disciplines: artificial general intelligence (AGI) and CogSci, is promising in this regard. The new direction is mostly concerned with studying, modeling, and computing AI capabilities that simulate facets of GI and functioning of higher cognitive mechanisms.

Whence, the focus in this thesis is on examining general problem solving capabilities of cognitive beings that are both: "human-comparable" and "cognitively inspired", in order to contribute to answering two substantial research questions. The first seeks to find whether it is still necessary to model higher cognitive abilities in models of AGI, and the second asks about the possibility to utilize cognitive mechanisms to enable cognitive agents demonstrate clear signs of human-like (general) intelligence. Solutions to cross-domain reasoning problems (that characterize human-like thinking) need to be modeled in a way that reflects essences of cognition and GI of the reasoner. This could actually be achieved (among other things) through utilizing cross-domain, higher cognitive mechanisms. Examples of such cognitive mechanisms include analogy-making and concept blending (CB), which are exceptional as active areas of recent research in cognitive science, though not enough attention has been given to the rewards and benefits one gets when they interact.

A basic claim of the thesis is that several aspects of human-comparable level of GI are based on forms of (cross-domain) representations and (creative) productions of conceptions. The thesis shows that computing these aspects within AGI-based systems is indispensable for their modeling. In addition, the aspects can be modeled by employing certain cognitive mechanisms. The specific examples of mechanisms most relevant to the current text are computation of generalizations (i.e. abstractions) using analogy-making (i.e. transferring a conceptualization from one domain into another domain) and CB (i.e. merging parts of conceptualizations of two domains into a new domain). Several ideas are presented and discussed in the thesis to support this claim, by showing how the utilization of these mechanisms can be modeled within a logic-based framework. The framework to be used is Heuristic-Driven Theory Projection (HDTP), which can model solutions to a concrete set of cognition problems (including creativity, rationality, noun-noun combinations, and the analysis of counterfactual conditionals).

The resulting contributions may be considered as a necessary, although not by any means a sufficient, step to achieve intelligence on a human-comparable scale in AGI-based systems. The thesis thus fills an important gap in models of AGI, because computing intelligence on a human-comparable scale (which is, indeed, an ultimate goal of AGI) needs to consider the modeling of solutions to, in particular, the aforementioned problems.

Keywords: Cognitive science, general intelligence, analogy-making, HDTP, conceptual blending, noun-noun composition, counterfactual conditionals, creativity.

Contents

Nomenclature i

Abstract iii

Contents v

List of Figures ix

List of Tables xi

I Foundational Backgrounds and Literature 1

1 Prolegomenous Remarks and Background 3

1.1 Cognitive Sciences 4

1.1.1 The Challenge of 'Labeling' 4

1.1.1.1 The Sloan Initiative's Report 7

1.1.1.2 The AI Debut as an Outmatching CogSci Discipline 9

1.1.2 Computational Processing of Cognitive Mechanisms 10

1.2 Artificial General Intelligence (AGI) 12

1.2.1 Roots of AI and Intelligent Agents 12

1.2.2 AI Reloaded: The AGI Debut 15

1.2.3 The Crossroad: A CogSci/AGI Compact Viewpoint 17

1.3 Representing Concepts for Computational Cognition 20

1.3.1 Concepts: Knowledge as Grouped Conceptual Entities 21

1.3.2 Levels of Representation 23

1.4 A General Overview of the Thesis 28

2 Analogical Reasoning 31

2.1 Cognitive Science, AI, and Analogy-Making 31

2.1.1 Using the Analogy Label in AI and CogSci . . . 32
2.1.2 General Motivating Examples . . . 34
2.1.3 The Structure-Mapping Theory (SMT) . . . 36
2.2 Computational Models for Analogy-Making . . . 41
2.2.1 Processes Involved in Computing Analogy . . . 42
2.2.2 Modeling Approaches . . . 45
2.3 Concrete Symbol-Based Systems . . . 47

3 A Logical Framework for Modeling Analogical Reasoning 53
3.1 Heuristic-Driven Theory Projection (HDTP) . . . 53
3.1.1 First- and (Restricted) Higher-Order Anti-Unification . . . 54
3.1.2 HDTP's Language: Conventions and Terminologies . . . 56
3.1.3 HDTP's Framework: Characteristics and Aspects . . . 59
3.2 Application in Analogy Domain Examples . . . 62
3.2.1 Two Classical Analogy Situations . . . 63
3.2.2 The Flower/Brain Metaphor . . . 67

4 Cross-Domain Reasoning via Conceptual Blending 73
4.1 General Assumptions and Basic Elements . . . 74
4.1.1 Conceptual Spaces and Frames . . . 75
4.1.2 Cross-Space Mappings . . . 78
4.2 A Cognitively Plausible View of Conceptual Spaces . . . 79
4.2.1 Characterizations of Concepts . . . 80
4.2.2 Representations of Concepts . . . 81
4.3 The Conceptual Blending Framework . . . 84
4.3.1 The Network Model: Constructing Blend Spaces . . . 85
4.3.2 An Overview of Former Accounts . . . 88
4.4 Challenges and Weaknesses of CB . . . 91
4.4.1 Principles of Optimality for Conceptual Integration . . . 92
4.4.2 What Concepts are "not" Blends? . . . 94
4.4.3 Strengths Vs. Weaknesses: Searching for a Missing Link . . . 96

II Applicability within Cognitively Inspired AGI Systems 97

5 Roles of Multifaceted Mechanisms in Logic-Based Computational Creativity 99
5.1 Introduction . . . 100
5.2 Forms of Creativity . . . 102

5.3 Creativity and Cognition . . . 103
5.4 Towards a Logic-Based Framework . . . 106
5.4.1 Roles of Analogy and Blending in Modeling Creativity . . . 109
5.5 Conclusive Remarks and Related Ideas . . . 111

6 Rationality-Guided Aspects of General Intelligence 115
6.1 AGI and Rationality . . . 116
6.2 Traditional Models of Rationality . . . 118
6.2.1 Some Rationality Challenges and Puzzles . . . 119
6.2.2 Classical Resolution Strategies of Irrationality . . . 121
6.3 Non-Standard, CogSci-Based Approaches . . . 122
6.3.1 Resolving the Selection Task by Cognitive Mechanisms . . . 122
6.3.2 Resolving the Linda Problem by Cognitive Mechanisms . . . 124
6.4 Modeling Rationality: Case Studies . . . 125
6.4.1 NARS: GI with Relative Rationality . . . 125
6.4.2 HDTP: GI-Based Rationality Through Analogy . . . 127
6.5 Conclusive Remarks and Related Ideas . . . 129

7 Concept-Based Interpretation of Novel Noun Compounds 131
7.1 Problem Importance and Challenges . . . 132
7.1.1 Motivations and Goals . . . 133
7.1.2 An Overview of Problem Challenges . . . 134
7.2 A Proposed Concept-Based Model . . . 136
7.2.1 Special Assumptions for Knowledge Acquisition in a Concept-Based Model . . . 137
7.2.2 Principles and Notations for a Concept-Based Model . . . 139
7.2.3 Development of Conceptual Knowledge Entities . . . 141
7.3 A Framework for Modeling Interpretations . . . 143
7.3.1 From HDTP to CB . . . 143
7.3.2 From CB to Interpretations . . . 146
7.4 Conclusive Remarks and Related Ideas . . . 148

8 An Implementation-Oriented Explication of Analyzing Counterfactual Conditionals 151
8.1 '*Being Smart*': Essences and Mechanisms . . . 151
8.1.1 Counterfactual Conditionals (CFC) . . . 152
8.1.2 Analyzing CFCs by Humans and in Artificial Systems . . . 155

8.1.3 A Crisp View of Specific Treatments 155
8.2 A Tale of Two Multifaceted Mechanisms 157
8.3 Towards a Treatment Formalization: Constructing Counterfactual Blends . 159
8.3.1 Generalization and Structural Mapping 160
8.3.2 Reasonability Principles for Counterfactual Blend Construction . . 161
8.4 An HDTP-Based Explication 165
8.5 Conclusive Remarks and Related Ideas 169

9 Conclusions **171**

References **179**

List of Figures

1.1 Primary Cognitive Sciences and Their Interconnections 8
1.2 KR's Challenges for Modeling Higher Cognition 24
1.3 A Symmetric Hexadecimal Code 26
1.4 A Change in Representations Changes Modeling 27

2.1 Phases of Analogy-Making 45
2.2 A Geometric-Analogy Example 48

3.1 Examples of First-Order Anti-Unification. 55
3.2 Examples of (Restricted) Higher-Order Anti-Unification. 58
3.3 Multiple Least General Generalizations 59
3.4 HDTP's Overall Approach to Creating Analogies 61
3.5 The "*Heat/Water*-Flow" Analogy Situation 66
3.6 An Intelligence Aspect: Drawing Metaphorical Similarities 67
3.7 Illustrating the "*Flower/Brain*" Analogy 70
3.8 Analogical Transfer of Relations 72

4.1 Conceptual Spaces of Smullyan's Puzzle 78
4.2 Goguen's Spaces: "BOAT" and "HOUSE" 79
4.3 The Prototypical Network Model of CB 86
4.4 Solving Smullyan's Rate-Time Puzzle via CB 87
4.5 Goguen's Version of CB. 90
4.6 Goguen's HOUSEBOAT and BOATHOUSE Blends. 91

5.1 The Nested Doll Principle: Design Examples from Two Different Domains 105
5.2 Analogy in Creative Advertising 106
5.3 HDTP's View of CB 108

7.1 Conceptual Entities and Frames 141
7.2 Blending of Nouns using HDTP 144

7.3 SNAKE GLASS: A Noun-Noun Blend Space 146

8.1 Generalization of Similar Roles for CFCs 161
8.2 Counterfactual Blend Spaces 168

List of Tables

2.1 Particularly Related Computational Models of Analogy 52

3.1 Axiomatizations of the Rutherford Analogy Situation 64
3.2 A Generalization of the Rutherford Analogy Situation 65
3.3 Another Formalization of the Rutherford Analogy Situation 66
3.4 Axiomatizations of the Flower/Brain Analogy Situation 69
3.5 A Generalization of the Flower/Brain Analogy Situation 71

5.1 Examples of Manifestations of Creativity 103

6.1 Parts of (a) the Wason Selection Task and (b) the Linda Problem 120

7.1 Combination of Suggested Parts of Noun Axiomatizations 145

8.1 A List of Counterfactual Sentences 154
8.2 The Muddy-Children Puzzle (Simplified) 154

9.1 Explaining Functions of Intelligent Systems by Cross-Domain Mechanisms 176

Part I

Foundational Backgrounds and Literature

1

Prolegomenous Remarks and Background

The systematic study of the various characteristics that make us attribute intelligence to cognitive beings, and the investigation of possible ways in which their cognitive abilities could be (biologically or physically) functioning, and could thus affect their level of intelligence, have always attracted the interest of numerous people over thousands of years, and among several study areas. For centuries, people have always been (i) contemplating the quintessence of intelligence in cognitive beings, (ii) fascinated with the aspects that can be considered fundamental faculties for cognition and intelligence: what such aspects are, where they originate from, and how they may work, and (iii) trying to simulate and mimic such aspects in most of their various pieces of artwork, sometimes mixing fact and fancy (e.g. fiction stories and artificial artifacts).

This thesis tries to contribute to the systematic study and investigation of utilizing concrete aspects of cognition, by considering contemplation, fascination, and simulation of such aspects, based on a cognitive-scientific approach. Its overall goal is to affirmatively answer both the following questions:

1. Is it still necessary to suggest general representations and utilizations of higher cognitive mechanisms (in particular, in building computational models that better mimic facets of general intelligence), despite the wide availability of a variety of systems aiming to model and compute their own views of human-like artificial intelligence?

2. Is it feasible and possible to do so?

The rest of the current thesis will expand over this take-home messages throughout the chapters it contains, demonstrating that solution models to a range of cognition

problems need to (and can) be resolved by understanding and computationally utilizing multifaceted, cross-domain, cognitive mechanisms in the modeling.

1.1 Cognitive Sciences

Over the past two millennia or so, the scientific interest, in empirically studying thinking, intelligence, cognitive abilities, or the human mind as a whole, started as a mere contemplation of the properties of mind and matter.[1] But recently, the interest evolved to viewing cognition and intelligence as brain activities and sorts of (mental) calculations, which can potentially be realized by artificial, electronic computations.[2] This evolution of the scientific interest has sometimes undergone unpremeditated paths through disparate disciplines, which did not only seem orthogonal disciplines but their sources of insights also spread over a very wide range of scientific or artistic traits: from argumentative (the humanities, e.g. philosophy of mind) to abstract and formal (the natural sciences, e.g. quantum-based models of cognition, logic, and artificial intelligence; cf. Busemeyer and Bruza [2012]; Engesser et al. [2007]; Wichert [2013]).

1.1.1 The Challenge of 'Labeling'

Despite their inherent, common connection, especially of studying higher cognition, these disciplines have continued to develop independently, most of the time. But it turned out to be hard (for the interested researcher) to both:

1. follow the related advancements in all of these disciplines, whether in breadth or in depth, and

2. benefit from studying one discipline in advancing another (by providing new solution insights or methods to the the well-known problems in the latter, by means of applying the recent findings in the former).

Moreover, there has always been an issue of debate regarding the content semantics and the name labels of these disciplines. What made the challenges even more difficult (for interested scientists) is to give these disciplines recognizable labels. Not only titles

[1] This has been mainly led by the theoretical works of pioneering philosophers, linguists, and logicians; particularly the Greek philosophers' interest in deductive reasoning, the process by which one assumes some statements to be true and derives further statements logically from the assumptions [Stillings et al., 1995, pp. 1].

[2] This is mainly guided by the theoretical and practical works of contemporary psychologists, neurologists, computer scientists, and electronic engineers.

of semantically similar disciplines were not as much syntactically similar[1], but also disciplines with fixed labels had their main content, their set of problems, or their solution methodologies re-defined.[2] This evolution has finally given the birth of a new field that gathers the intersections of the disparate disciplines and connects their respective parts in harmony, under the umbrella of a united, scientific area that offered multidisciplinary ways of understanding the mind and cognition. This area has developed in the past few decades under the label of "*cognitive science*" (cf. Boden [2006a,b]).

In the widest sense, various disciplines in the arts and sciences are particularly relevant to the study of (the functionalities of) the human beings's higher cognition and intelligence. A few hundreds of entries in psychology, neuroscience, linguistics, philosophy, anthropology, education, computer science, ethology, among other disciplines, are given in Wilson and Keil [2001] to contribute altogether as encyclopedic definitions of the most important terms appearing in what collectively can be called the 'cognitive sciences'. On the one hand, many 'cognitive sciences' have collectively offered multidisciplinary ways to understand the mind and higher cognition during previous centuries. On the other hand, the intellectual origins and foundations of a unified, interdisciplinary, academic field have only recently been laid in the mid fifties of the twentieth century (cf. Thagard in Thagard [2005] and in [Frankish and Ramsey, 2012, pp. 65]; Miller in [Miller, 2003, pp. 141]; and Gardner in [Gardner, 1987, Ch. 2]). This was when adherents in various fields developed techniques, theories, and theoretical assumptions of cognition, based on complex representations and computational procedures of thinking (see e.g. the early seminal works of Newell and Simon in [Newell et al., 1963; Newell and Simon, 1963, 1972, 1976], and section 1.1.2). This newly-born field focused on a common set of mind-related problems that have already long-existed in various 'cognitive sciences'. What was new, however, is uniting their shared theoretical standpoints and common research strategies and objectives as a way of studying the mind, under the umbrella of one field; called 'cognitive science' since then.[3]

The title name of this unified science, 'cognitive science', as well as its existence and development, have been affected very much by its organizational origins, which go all the way back to the seventies when the Cognitive Science Society (CSS) was formed

[1] For example, "cognitive studies" and "information-processing psychology" were different labels for the same discipline (cf. [Miller, 2003, pp. 143]).

[2] Within the field of psychology, for instance, "behaviorism" became no longer the most important area as it was in the early 1950s (cf. Baum [1994]).

[3] Instead of always emphasizing that the interdisciplinary field of 'cognitive science' is embracing a multitude of 'cognitive sciences', it has become widely acceptable to replace the plural, 'sciences', with the singular, 'science', in particular when the interdisciplinarity of the field is of a more important concern than the involved discipline titles.

and the journal "Cognitive Science" began (cf. Bermúdez [2010]; Thagard [2005]). One dares to assume that if it were not for the establishment of the latter two organizational origins with their respective title names, the interdisciplinary field would have been named differently or have its name changed over time (as already was the case before their establishment: the field's name at Harvard, the stronghold of the leaders of the "cognitive revolution" (cf. section 1.1.2), was 'cognitive studies', but was 'information-processing psychology' at Carnegie-Mellon University [Miller, 2003, pp. 143]). Moreover, the label, 'cognitive science', began to spread during (and after) the seventieth, which urged the involved and interested leading researchers to propose their own definitions of the name (according to their perspectives or scopes of work) instead of actually proposing other names that define what this unified field is for them, or how they view it from where their work stands. Gardner reflected this situation by mentioning that "in the course of proposing and founding a new field of knowledge, many individuals will formulate their own definitions" [Gardner, 1987, pp. 5]. In a sense, this is fortunate because it could have been undesirable to have disagreeing titles that describe the same field of interest, especially after the field has gone a centuries-long way of evolution to become a unifying field of the cognitive sciences. But this also was unfortunate, in another sense, because the perspectives, the important problems, and the methods to tackle such problems, among other things, of such 'cognitive sciences' were very difficult to put together to form a single field. Moreover, 'cognitive science' means to a large group of people 'cognitive psychology', or even 'psychology'. This is not less of a misunderstanding of labels than that of what "*artificial intelligence*" (denoted *AI*, henceforth) is to them.

So, What is Cognitive Science? Broadly speaking, 'cognitive science' (usually: *CogSci*) can be viewed as a scientific field of an inter-, and multidisciplinary nature, where each of its themes may belong to the intersection of more than one discipline (of the 'cognitive sciences'). Definitions of 'cognitive science' have been proposed according to several scopes, interests, and in many ways. Moreover, when it comes to questioning the involved disciplines, scholars differ in their presentations, which may range from the more specific to the very broad or even gelatinous. Here are three presentation examples by three expert scholars:

1. Thagard introduces cognitive science as "the interdisciplinary study of mind and intelligence, embracing philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology" [Thagard, 2005, pp. ix].

2. Miller considers cognitive science as the "child product" of a time when "psychology, anthropology and linguistics were redefining themselves and computer science and neuroscience as disciplines were coming into existence" [Miller, 2003, pp. 141].

3. Gardner defines it as "a contemporary, empirically based effort to answer long-standing epistemological questions —particularly those concerned with the nature of knowledge, its components, its sources, its development, and its deployment" [Gardner, 1987, pp. 6].

Nonetheless, to give one precise definition of what CogSci is, seems very challenging, if at all possible.[1] The interested reader may refer to [Bermúdez, 2010; Boden, 2006a,b; Frankish and Ramsey, 2012; Gardner, 1987; Stillings et al., 1995; Thagard, 2005, to mention a few], where several, classic, and modern, definitions and views of CogSci are presented and argued in depth.

1.1.1.1 The Sloan Initiative's Report

Specifying the common problems that appear in CogSci takes one back in time to one of the most famous roots of stressing the importance and promise of the interdisciplinary field. Namely, the State of The Art Report commissioned by the Alfred P. Sloan Foundation: an unpublished[2] report that is usually referred to as SOAP (the report is cited here as Keyser et al. [1978]). The SOAP report gives both a serious attempt to describe the state of research in 'cognitive sciences', as well as a broad outline of the theoretical viewpoints and research objectives of scholars in these fields. Here is an attempt to introduce the rising field of 'cognitive science' by a number of leading scholars:

> "the *study* of the *principles* by which *intelligent entities* interact with their *environments*" [cf. Keyser et al., 1978, pp. 3, emphasis added].

Faced with the challenge to integrate many considerations into a coherent science of cognition, the State of The Art Committee sketched a handful of cognitive problems and a handful of possible approaches that necessitate interdisciplinary collaboration in their pursuit. The report factored the complicated problem of specifying the main 'cognitive sciences' into the simpler problem of looking for disciplines that "interact strongly with

[1] Only to recall, the thesis focusses on the systematic study of intelligence aspects and the investigation of the possible ways these aspects can function, which helps in artificially modeling or computing them.

[2] Many scientists contributed to the Sloan initiative's report, which is edited by Keyser, S.J., Miller, G.A., and Walker, E. in 1978. A scanned copy can be retrieved from the Cognitive Science Journal's archive webpage at http://csjarchive.cogsci.rpi.edu/misc.

each other but only weakly with everything else" [Keyser et al., 1978, pp. vi]. The basic criteria is that solutions to some problems of one particular discipline "depend critically on the solution of problems traditionally allocated to other disciplines" [Keyser et al., 1978, pp. vii].

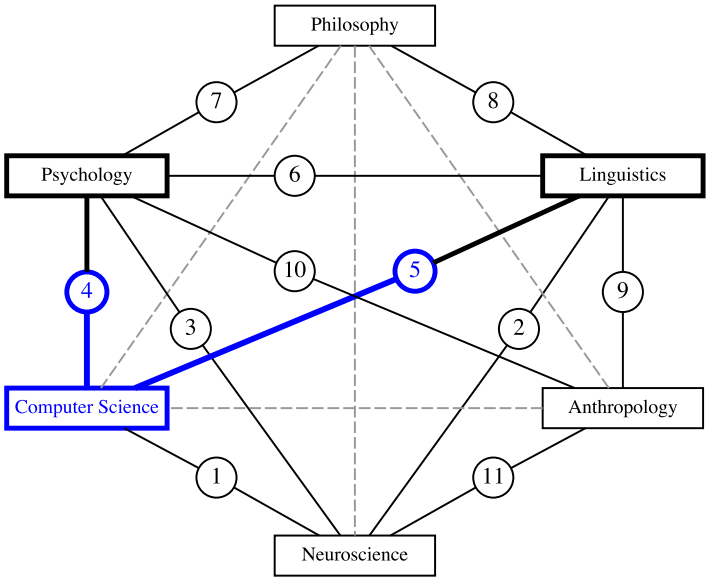


Figure 1.1: A reproduced copy of the original illustration given in [Keyser et al., 1978, pp. 4]. It depicts the 6 primary disciplines (in rectangles), and the 11 subdomains suggested in the SOAP report to constitute the cognitive sciences: 1. cybernetics, 2. neurolinguistics, 3. neuropsychology, 4. simulation of cognitive processes, 5. computational linguistics, 6. psycholinguistics, 7. philosophy of psychology, 8. philosophy of language, 9. anthropological linguistics, 10. cognitive anthropology, and 11. evolution of brain. Dashed lines indicate weaker interdisciplinary ties than those indicated by solid lines.

Since the time of the SOAP report, the 6 disciplines enclosed within rectangles in Figure 1.1 have been widely considered the primarily involved 'cognitive sciences'. The interconnections (or "major bonds") among these disciplines give rise to 11 well-defined areas of inquiry, each is called a "subdomain". The 6 primary disciplines and the 11 subdomains are seen as an integral part contributing to the established branch of cognitive science. A subdomain involves the intellectual and physical tools of the two disciplines this subdomain ties together. Subdomain 4, 'simulation of cognitive processes', for instance, is of a highly important concern to the thesis. The "simulation of cognitive processes has combined computer science and psychology in order to formulate explicit theories of thinking and problem solving" [Keyser et al., 1978, pp. 5].

1.1.1.2 The AI Debut as an Outmatching CogSci Discipline

The original SOAP report included 'computer science' (also CS), not AI, as one of the primary CogSci disciplines, whereas, in more recent references, websites, and precursors to the interdisciplinary branch, AI usually replaces CS in the corresponding re-produced hexagonal diagrams. For instance, 'computer science' appears in Keyser et al. [1978, pp. 4] and Miller [2003, pp. 143], whereas 'artificial intelligence' replaces it in Gardner [1987, pp. 37] and Bermúdez [2010, pp. 91]. Given that the seminal event for the birth of AI has already taken place in the summer of 1956, it means that more than two decades after the label 'artificial intelligence' took the chance to prevail, the leading scientists of the SOAP report chose 'computer science' as a primary cognitive science (cf. section 1.2.1). This may seem like a minor thing that could simply be a matter of taste in preferring one label over another, but, I hypothesize, it is not (at least according to the current chapter's prolegomenous context). On the contrary, this is a basic motivational observation that indicates how conceptions are changing over time and by usage, so that their development and entrenchment can be affected.[1] My main argument is that there is an apparent gradual change of the common (scientific) views, both about (i) the interrelation between AI and other disciplines, and about (ii) which of CS and AI appears to contribute more as an interdisciplinary field of study.

Back in the 1970s, CS was treated by the (scientific) community (in particular, the leading scholars who contributed to the SOAP report) as a cornerstone discipline that is as foundational and contributive to the 'cognitive sciences' as any of the other primary disciplines. This was the case because, similar to other primary CogSci disciplines, CS already contained mature directions of scientific interest, as well as major problems (and solution methodologies) related to studying human intelligence. In this sense, each primary discipline is an irreplaceable cornerstone. AI was seen as a rising subfield that could not have been considered as much foundational or irreplaceable. At that time, the (scientific) impact of AI on the (scientific) community, as a stand-alone discipline, may have ranged from very small to small, whereas that of any other primary CogSci discipline was remarkably high, I presume. Over time, AI has become more widespread, with a higher impact, creating ties with other CogSci disciplines. The same SOAP report's view of primary CogSci disciplines as those which "interact strongly with each other but only weakly with everything else" [Keyser et al., 1978, pp. vi] favors AI now as a more integrable CogSci discipline.

[1] This idea of "concept development" and "concept entrenchment" is related to one of the thesis' contributions, extensively elaborated on in Chapter 7.

1.1.2 Computational Processing of Cognitive Mechanisms

The general term "cognition", of natural (biological) entities, is customarily used for forms of "mental processing" that involve ongoing manipulation of "knowledge". This may include remembering, reasoning, learning, planing, decision making, problem solving, communicating deep ideas, among many other operations. Using both theoretical and empirical methods, cognitive scientists study all these higher mental "*processes*", and the underlying "*mechanisms*", whereas advocates of AI aim to apply the studies to computationally simulate the processes and mechanisms. "In cognitive psychology and cognitive engineering, cognition is typically assumed to be information processing in a participant's or operator's mind or brain" [Blomberg, 2011, pp. 85]. Cognition is usually the faculty for "processing" of "knowledge" and, in fact, the word "cognition" originates from the Latin *cognoscere* (co- + gnoscere); meaning: "to come to know". This additionally reflects the importance of discussing "knowledge" when discussing "cognition", and attracts one's attention to the importance of "processing", in general, and processing "knowledge" in particular (cf. section 1.3.1).

Having talked about the importance of knowledge processing, note that other CogSci disciplines may study the humans themselves, whilst AI should study and develop implementations. AI and CS can still work together as two sides of the same CogSci coin that connects 'computer science', 'linguistics', 'psychology', 'computational linguistics', and 'simulation of cognitive processes' in Figure 1.1. CS deals with "creating the right model for thinking about a problem and devising the appropriate mechanizable techniques to solve it" [Aho and Ullman, 1995, Ch. 1], which inescapably brings issues of knowledge representation and reasoning (KRR) to focus, when simulating and manipulating representations of higher cognitive mechanisms within computational models of cognition —a subject that is brought to discussion in sections 1.3, 4.2, and 7.2.

Once Upon a Time, There was a Revolution: "Behaviorism" (cf. Baum [1994]) was the most important area of psychology in the thirties and forties of the twentieth century, where (a school of) psychologists believed that it was only important to pay attention to actions and observable behaviors of people (and animals) which one could measure, not with unobservable events that take place in their minds (cf. [Skinner, 1984]). Behaviorists maintain that behaviors can be described scientifically without recourse either to internal physiological events or to "hypothetical constructs" such as "thoughts" and "beliefs", and made no distinction between animal behaviors, infant behaviors, and those behaviors of mature adults (cf. Baum [1994]; Gardner [1987]). Gardner indicates that the behaviorists do not talk about any kind of "internal representation", ideas, models

in the mind, or anything inside the black box called "the mind". Gardner also points out that cognitive scientists escaped this dead-end, where their idea was that people think, compute, solve problems, have images in their heads, and operate among these things —which is what thinking seem to be all about.

During the period of the Second World War, researchers "had developed a series of technologies that lent themselves to anthropomorphic description, and once the war ended these technologies inspired novel forms of psychological theorizing" (cf. [Agre, 1997, pp. 1]). Very few years after the war, when the use of electronic machines (a.k.a. computers) became more prevalent, a number of technologically sophisticated scientists, mainly cognitive psychologists, were ambitious enough to continue using their acquired knowledge of technical systems, and apply it to the study of the mind and behavior. Coming not only from cognitive psychology, but from philosophy, artificial intelligence, and other disciplines as well, the scientists styled themselves as "cognitive scientists". Their knowledge included (i) formal notions of theoretical machines (a.k.a. the "Turing Machines"; [Turing, 1936]) that could in principle carry out any possible calculation [Turing, 1950], as well as (ii) psychological notions, particularly neuropsychological syndromes, such as "aphasia (language deficit), agnosia (difficulty in recognition), and other forms of mental pathology consequent upon injury to the brain" (cf. [Gardner, 1987, pp. 22]). These scientists already had the chance to understand how computers work, and closely experienced how they can solve problems, which made them think about the human mind as being a certain kind of an information processing system (cf. [Gardner, 1987; McCorduck, 2004]).

According to how Gardner delivers it, a very important part of cognitive science was to point the science in a positive direction (rather than to a dead-end). This major paradigm shift of the twenties century came to be known as the "*cognitive revolution*". It was an intellectual movement in the modern context of a greater interdisciplinary research that is not restricted to cognitive psychology, but combines psychology with approaches developed within AI, CS, and neuroscience, for example. As Agre puts it, "[t]he new psychology sought to describe human beings using *vocabulary* that could be *metaphorically associated with technologically realizable mathematics*" [Agre, 1997, pp. 1; emphasis added]. As an example of describing human beings using this metaphorical vocabulary association, Agre pointed out that Miller[1] observed that "aiming a gun at a target by continually sensing the target's location" can be described, in human-like terms, as "pursuing a purpose based on awareness of the environment". Another exam-

[1]Miller was a key player in the cognitive revolution [Fellbaum, 2013, pp. 1] (together with Jerome Bruner and Noam Chomsky).

ple is that "new methods for signal detection" could be described as "making perceptual discrimination" (cf. [Agre, 1997]).

Studying "thinking", and the related "mental processes" in cognitive psychology, was the key idea in the cognitive revolution movement that started in the 1950s, and had become the dominant research line of inquiry in most psychology research fields by the 1980s. The cognitive revolution movement replaced "behaviorism" as the leading psychological approach to understanding the mind. Nowadays, it is both feasible and attractive to test posed hypotheses about how these mental processes function and, in addition, study computational frameworks that develop the functions in AI.

Transitioning: In this section, I briefly mentioned that the interdisciplinary field, concerned with studying thinking in cognitive beings, has finally been widely recognized as 'cognitive science'. This raised a challenge for the scientific community to agree on specifying and 'naming' the dispersed domains of the interdisciplinary field; a challenge that, to a great extent, has been resolved by the SOAP report. I also pointed out that, over time, AI become widely recognized as a primary CogSci discipline (that outmatched CS as a CogSci alternative). In the next section, I briefly discuss the historical challenge of 'fine-tuning' the AI field itself and its core contents to fit better within the CogSci domain (and to link to its sub-disciplines in a stronger way). Understanding this historical challenge plays an important role in building modern artificial computer systems that are cognitively inspired.

1.2 Artificial General Intelligence (AGI)

1.2.1 Roots of AI and Intelligent Agents

Artificial Intelligence: Although its label is now very widely and publicly known, AI does indeed suffer from not having commonly agreed-upon definitions or standard specifications of many of the terms it uses. There are even debates on what "intelligence" is, in the first place, and McCarthy himself, who introduced the AI label, points out that we cannot yet characterize in general what kinds of "computational procedures" we want to call intelligent (cf. McCarthy [1998]). He, nonetheless, defines AI as "the science and engineering of making intelligent machines, especially intelligent *computer programs*" [McCarthy, 1998, emphasis added], which is acceptable in the sense that the ultimate goal of AI was to study and develop "thinking machines": computer systems that possess human-comparable intelligence. Russell and Norvig also list eight definitions of AI,

organized into four categories, and laid out along the two dimensions of "thinking" and "acting" (cf. [Russell and Norvig, 2010, pp. 2]). In addition, other sources try to specify what AI does or how it operates, leaving "intelligence" itself aside as an issue of debate (mostly in psychological contexts about theories of intelligence).

The label "Artificial Intelligence" has been coined in the summer of 1956, where McCarthy first used it in the proposal for the Dartmouth conference, which conjectured that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [McCarthy et al., 2006, pp. 12]. The Dartmouth Summer Research Project on AI is considered the point in time to which the big-bang of AI is rooted. It was in this event that the proposal with introducing the term AI is credited [McCarthy et al., 2006; McCorduck, 2004], and the plan for next years of work in the field is foreseen by the pioneers who participated in the event. The Dartmouth conference itself "did not lead to any new breakthroughs", but it introduced "all the major figures" in AI to each other (cf. [Russell and Norvig, 2010, pp.18]), and listed main ideas and contributions in the field, such as Newell and Simon's "Logic Theorist" (cf. Newell and Simon [1956]) which was "received with interest" [McCorduck, 2004, pp. 104] and "stole the show" [Russell and Norvig, 2010, pp. 17] in the event. Other rooting ideas are also found in the collection of classic papers by pioneers, such as that of Turing's (the English mathematician and logician), and Marvin Minsky's (the American co-founder of the MIT's AI laboratory), who both were behind the pivotal advances in artificially simulating human thought processes with computers (cf. Feigenbaum et al. [1995]).

Intelligent Agents: The AI field's initial goal was to primarily focus on "the study and design of intelligent agents" [Poole et al., 1998, pp. 1]. But "what is exactly an intelligent agent?" is another (philosophical) question that is embarrassing "in just the same way that the question *what is intelligence?* is embarrassing for the mainstream AI community" [Wooldridge and Jennings, 1995b, pp. 116; emphasis original]. Nevertheless, it is customary to view an *intelligent agent* as anything (computationally) approximating a *system* that possesses the ability to *perceive* its surrounding *environment* (through sensors) and take *actions* (through actuators) to maximize its chances of success in *intelligently solving* a given problem.

Lots of texts in the literature give more elaborations on the definitions, properties, and attributes of intelligent agents, and the several applications they can be used in (cf. [Nilsson, 1998; Russell and Norvig, 2010; Wooldridge and Jennings, 1995a,b, to mention just a few]). Russell and Norvig, in particular, give more details about the

traditional depictions of different types and characterizations of agents, sensors, actuators, and environments (cf. [Russell and Norvig, 2010, §2.4]). Wooldridge and Jennings argue that AI researchers, in particular, use the term "agent" in a more specific sense that means a computer system, which is either conceptualized or implemented reflecting mentalistic notions (i.e. concepts that are more usually applied to humans, such as knowledge, belief, intention, and obligation) [Wooldridge and Jennings, 1995b]. In a sense, these are how "intelligent agents" may be viewed in this text. But one should note also that the set of concepts that are more usually applied to humans can be very large (or fuzzy). For example, Steels focusses on behavior-oriented AI and presents "self-preservation" as an emerging property of agents that adapt their behavior in interaction with dynamically changing environments (cf. Steels [1993]). Steels's notions are easy to present in deterministically formalized frameworks for physical agents, but would take cognitive scientists ages to agree on general specifications.

Artificially Intelligent Technology: Historically and scientifically, the cognitive revolution was behind the wide interest in AI and its flourishing as an interdisciplinary field. However, many AI projects concerned themselves with commercially producing physical, "intelligent devices" that allegedly work like humans. That was how the idea spread, especially when AI turned out to be more of an industry of "expert systems" than of a research field of an interdisciplinary nature (cf. [Russell and Norvig, 2010, §1.3]). The initial ambitions in the 1960s and 1970s had grown by the 1980s into an "industry", where not only researchers and engineers but also companies developed the AI technologies in building specialized hardware. Large consultant-based businesses of "knowledge engineering" also grew, which entailed extracting knowledge from human experts and encoding it in machine readable formats using specialized languages (e.g. LISP). "The AI industry boomed from a few million dollars in 1980 to billions of dollars in 1988 [...but soon] after that came a period called the "AI Winter", in which many companies fell" [Russell and Norvig, 2010, pp.24].

This has definitely a big effect on the meaning of the AI field, on the goals of its projects, on the ambitions of its adherents, and on almost every related aspect to AI as a scientific field (e.g. what criteria signify AI agents?). But in the very first years of the twenty-first century, notable technological achievements in the field started to show up again, and funding was increased again. IBM, for instance, has single-handedly been able to impress the global community, twice, by showing that the continuation along the technological line of improving AI by industrial projects is still plausible and impressive. IBM first succeeded in 2002 to present the first chess-playing computer,

DeepBlue [Hsu, 2002a], that could beat the human chess-master at the time. Then, in 2011, IBM succeeded again in delivering the question-answering system, Watson [Ferrucci et al., 2010], that could beat human participants in the American TV quiz show, *Jeopardy!*.

Artificially Intelligent Cognitive Agents: Can what Watson-like devices are really doing be considered "intelligent"? Would that be considered a theory of intelligence? According to the main objectives of studying human cognition and building human-comparable intelligent entities, and according to the position[1] taken in this thesis, this is definitely not the case. Without going into further traditional debates, let us not forget that people in the modern era are not astonished by the hands-on experience they already have with intelligent "technological" systems (e.g. Apple®'s Siri, Google®'s online services, Microsoft®'s 3D motion sensing input device: "Kinect", and smart phones and tablet devices, which are now considered utilities for mundane usage). Moreover, existing AI systems can already reason, plan, and perform actions, but their behavior may not be viewed as originally motivated by essential cognitive abilities that reflect one aspect of intelligence or another. Many of these systems neither focus on integrating human-comparable competencies nor on applying such competencies to a wider range of directions, but are usually designed to rather solve a specific task, and fail not only in solving another task but also in compatibly parsing that other task's input. Note that IBM's Watson and DeepBlue systems can neither deal with each other's main functionalities nor even parse each other's input. Their impressive success obscures that these systems clearly lack "general intelligent action" as suggested by Newell and Simon (cf. Newell and Simon [1976]). Moreover, "while both of these artifacts are intelligent simpliciter, they most certainly aren't general-intelligent" [Bringsjord and Licato, 2012, p. 26]. If we even go back to the roots of the AI field, the Dartmouth conference proposed that 'intelligence', as a central property of human beings, can be so precisely described that it can be simulated by a machine (cf. McCarthy et al. [2006]). But 'general intelligence', the integrated view, is not yet achieved, and is still among the AI field's long term goals. Here is exactly where the modern view of AI (e.g. AGI) plays the role.

1.2.2 AI Reloaded: The AGI Debut

Researchers use terms such as 'classical AI', 'narrow-AI', and 'good old-fashioned AI' (GOFAI) to indicate that they take a traditional approach to achieving human-like think-

[1] Particularly regarding the integration of representations and utilization of specific higher cognitive mechanisms in solving problems that signify aspects of intelligence.

ing. This traditional approach aims at producing systems demonstrating "intelligence" in very specific, highly constrained tasks, based only on symbolic "manipulation" of sorts of ideas, such as classical (propositional) logic and problem solving.

The modern approach to achieving AI by simulating cognition is mainly based on representations and implementations of higher cognition in artificial systems. Such systems are designed with the intention to be acquired with "human-like" level of "general intelligence", so that their functionalities are supported by results (and solution methodologies) from other CogSci disciplines, such as philosophy, psychology, or linguistics. Advocates to this latter approach adopt terms other than AI to refer to their research field. For example, '*strong-AI*', '*cognitive systems*', or '*artificial general intelligence*' (AGI).

The notion of "*general intelligence*", abbreviated 'GI' henceforth, does not only focus on treating "intelligence" as one factor of problem solving by machines; e.g. the way "intelligence" is treated by 'narrow-AI'. GI is, however, a general problem solving capability of cognitive beings that is:

1. "human-comparable": the required intelligence capability is very close to (i.e. exactly like, better than, or a little bit below) that of smart humans; and

2. "cognitively inspired": the required intelligence capability behaves in a way that simulates the functioning of higher cognitive mechanisms.

In this thesis, the use of the term "human-comparable" is preferred over "human-like" because (i) the modern direction of treating intelligence as a whole (i.e. the direction followed by advocates to strong-AI and AGI) aims at achieving humans' general intelligence or even beyond, and (ii) humans are not the only example of generally intelligent, cognitive beings. "Human-comparable" is less anthropocentric, since many living creatures (e.g. birds, dolphins, ants, etc.) are already known to show exciting types of GI (though humans are usually considered the best available exemplar of generally intelligent, cognitive entities). Having already given this distinction, note that all problems discussed in this thesis are restricted to, and based on, human cognition.

AI in this modern, cognitively inspired sense should not only reflect on such an ability that can be measured by IQ tests, for example, but should reflect on the ability to solve a multitude of problems the way cognitive entities (especially humans) do; where the solution, in essence, involves and utilizes the cognitive capacities (e.g. making analogies, blending concepts, taking a "rational" decision, and being "creative", to mention some of the concerned examples). There is indeed a loose relationship between GI and the notion of "g-factor" in psychology, which is an attempt to measure intelligence across various domains in humans (cf. [Goertzel and Wang, 2007, pp. 1]).

Artificial General Intelligence (AGI): Not only has AI always been a challenging domain for its advocates (who belong to a wide spectrum that ranges from daily users of specific computer-based applications to experts in advanced scientific research), but so has also the notion of "intelligence" raised several discussion debates about whether an uncontroversial jargon is available in the first place (cf. Legg and Hutter [2007]; Wang [2008]). Wang, for example, clarifies, analyzes, and compares five typical ways to define AI, in order to argue how these ways led the AI research to very different directions; most of them have trouble to give AI a proper identity (cf. [Wang, 2008]). As a lively new research area, AGI inherits all such kinds of challenges. Over the past decade, there have been many attempts to define GI as prescriptions to model and develop strong-AI systems. AGI is used here as a widely accepted term that stresses the general nature of the desired intelligence capabilities of the strong-AI systems being researched. The expanded versions of [Bach et al., 2012; Baum et al., 2010; Schmidhuber et al., 2011; Wang and Goertzel, 2012] contain overviews and discussions about AGI as one of the widest strong-AI terms in use. Unlike other scientific disciplines, a major part of research in AGI still focusses on attempts to reaching an agreement on defining general intelligence. This goes side by side with the challenges the AGI field faces in developing working models and systems.

With the research advancements in CogSci disciplines, researchers increasingly acknowledged the indispensability (and recognized the feasibility) of returning back to the original goals. In recent years, the AGI researchers have (willingly) focused on reaching that ultimate, original goal by exploring "all paths", including theoretical and experimental CogSci disciplines, using their interdisciplinary methodologies. There is no doubt that this task of confronting the more difficult issues of human-comparable intelligence is immoderately tough. The toughness however encouraged many in the field to remarkably focus on recovering the path to the original goals, so that AI really flourishes as an interdisciplinary CogSci. The earlier observation in section 1.1.1.2, about exchanging CS with AI in the primary cognitive sciences, provides a further witness to this development.

1.2.3 The Crossroad: A CogSci/AGI Compact Viewpoint

Cognitive Modeling, Architectures, and Agents: Cognitive scientists usually use the term "computational cognitive modeling" (or its variant, "computational psychology"; cf. Boden [1984]) to explore the essence of cognition and various cognitive functionalities. This exploration is achieved through developing detailed, process-based un-

derstanding, by specifying corresponding computational models (in a broad sense) of representations, mechanisms, and processes (cf. Sun [2008]; Sun and Ling [1998]). Computational cognitive modeling hypothesizes internal mental processes of human cognitive activities, and expresses such activities by computer program models, which often consist of many components. Some researchers even take the position that "computational cognitive models can be true theories of cognition" (cf. Sun [2009]).

A *cognitive model* tries to generally exhibit abstracted mock-ups that structurally capture basic characteristics of the cognitive entities, in order to explain the major features of cognition and the cognitive processes. In simple words, cognitive models:

1. elucidate what such major features are (i.e. aspects of GI),
2. predict how the major features and their underlying processes function, and
3. explain how two or more major features relate to, or interact with, each other in the model.

A *cognitive architecture*, then, allows modeling to be more beneficial, by proposing (artificial) computational processes that approximately act like the corresponding modeled cognitive entities.

Many cognitive scientists (predominantly cognitive psychologists) would agree that a cognitive architecture is more of a *theoretical* entity in CogSci than of a computational model of GI. In fact, a proposed architecture basically attempts to encapsulate an integrated, broad "theory of cognition" about the many aspects of human cognition and performance. The aspects, too, are usually grounded on experimental data from cognitive psychology. But what I want to emphasize more is that a proposed architecture should also be able to provide us with a way to build human-comparable computational models, that satisfy as many GI aspects as possible, based on the architecture's theory of cognition (and on solution models describing how to solve cognition problems).

There are already too many cognitive models and architectures for intelligent agents, which poses a challenge on its own in advancing the (restricted) field of interest. The many different views of the "theories" behind these models and architectures seem remarkably difficult to integrate. The majority of the AI history makes it obvious that focus has always been on simulating isolated components of intelligent agents (e.g. vision systems, theorem-provers, voice recognition, question-answering, etc.) rather than on whole agents (cf. [Russell and Norvig, 2010, pp. 59]). This inadvertently helped in widening the gap between narrow-AI and AGI, for instance, by continually modeling isolated aspects of intelligence that may be more difficult to amalgamate in later im-

provements of any single model or architecture (which could have basically been built on such isolated aspects).

On the one hand, I claim that specific cognitive mechanisms are more important as intelligence aspects than others, and are easier to embed into existing models to endow the models with these aspects. But existing models, on the other hand, still need to update their views about how GI aspects can be integrated. This view of whole-agent became widely accepted and a central theme in recent texts (cf. Nilsson [1998]; Poole et al. [1998]; Russell and Norvig [2010]). It takes us back to the roots of AI, and agrees with Newell and Simon's original ideas of implementing systems that are capable of "the same scope of intelligence as we see in human" [Newell and Simon, 1976, p. 116]. Examples of whole-agent cognitive modeling are "ACT-R" and "SOAR", which are among the currently popular cognitive modeling architectures (cf. Anderson et al. [2004]; Laird [2008, 2012]; Laird et al. [1987]).

Restricting Cognitive Modeling: The thesis neither intends to provide a general theory of cognition nor a complete cognitive model, but rather promotes and adopts the restricted view of CogSci and AGI that: models of artificial, computational systems can be designed in a way that does not only enable ingenious solutions of baffling problems (related to higher-level cognition), but can also be endowed with the ability to solve problems the way a generally intelligent, cognitive entity (like a human being) does. The main focus is not on proposing a new cognitive architecture. It is rather on extensively investigating specific, essential cognitive capacities of human beings that deliver solution models to important problems related to higher cognition (also cf. Abdel-Fattah and Schneider [2013]). The given solution models are intended to be based on cognitive scientific studies, and not on the design of one particular cognitive model or another (although they are based on a specific framework called "HDTP"; cf. Chapter 3). However, the solution models given throughout the thesis are as abstract and general as possible, so that their underlying ideas can serve in general computational models seeking to simulate the same GI aspects. Existing studies from various cognitive science disciplines will be used to consolidate the suggested solution models. Furthermore, the selected capacities are computationally plausible, which means they can be computed by developing computer programs for artificial systems. Their utilization will be founded on using specific representational and structural treatments (cf. sections 1.3.1, 3.1, and 4.2). But the overall ideas can still be utilized within general cognitive models or architectures like ACT-R, SOAR, or others.

1.3 Representing Concepts for Computational Cognition

In order to be cognitively inspired, an AGI model should endow an artificial agent with the ability to mimic human beings in several aspects, particularly sensing, (intelligently) thinking, and (intelligently) acting. The three (sensing, thinking, and acting) have to be connected by a central component of the agent (an analogue to the human mind) that would be responsible for utilizing the cognitive mechanisms by means of collecting and organizing the stream of information, usually called "beliefs", which the agent (continuously) perceive or (intelligently) conclude. Thus, agents in AGI models need to build, maintain, and continuously manage, a repository of beliefs, that correspond to the collected pieces of information and the valid inferences that may be (intelligently) concluded.

About Knowledge Representation: Despite the widely accepted fact that there is a fundamental, epistemological difference between "beliefs" and "knowledge", the repository of belief entities is usually labeled "knowledge base" and denoted "*KB*". Moreover, the notions of "beliefs" and "knowledge entities" will be used interchangeably throughout the text to reflect the same meaning.

A KB needs to consider the basic issue of how knowledge entities (i.e. beliefs that constitute the agent's knowledge base) are represented (cf. diSessa [1988]; Gärdenfors [1988]). *Knowledge representation*, or *KR*, is a supremely crucial notion, which the modeling of artificially intelligent agents, in particular, must consider very carefully (cf. Davis et al. [1993]; Gutiérrez [2012]). Indeed, KR is even considered a cornerstone AI sub-discipline, the research in which focusses on how to translate 'knowledge' from being completely 'available' in one medium to being approximately (or better yet, equivalently) 'describable' in another: from descriptions, by which knowledge about the surrounding world is present in the human mind, to other descriptive interpretations that an agent can use to account for (and act upon) the corresponding part of its environment.

KR's main job (with regard to cognitive mechanisms, agents, models, and architectures) should be to sufficiently approximate descriptions of needed knowledge entities, where the description of these entities (and the relationships among them) affect (and are affected by) the described part of the surrounding environment, with which an agent is concerned. Thorny discussions about identifying many of the terms in this latter view are unavoidable. Many questions arise too. For example, "how, and what, knowledge is described in the human mind?", "what description may be used for knowledge entities

in artificial agents?", "how much 'sufficient' is a sufficiently approximate description?", and "what part of a surrounding environment is an agent concerned with?". For a text of the current size, however, I would rather prefer to only consider needed identifications (and crisply mention related literature), and stress that this representation issue is inevitable to be encountered whenever it comes to solution models of cognitive problems (e.g. related discussions are given in sections 4.2 and 7.2).

1.3.1 Concepts: Knowledge as Grouped Conceptual Entities

To start with, the next few paragraphs clarify both the difficulty and importance of reaching a generally satisfying representation, by means of dealing with "concepts" as consisting basically of knowledge entities. The intention is to introduce the issue of concept representation as well as to support why the particular representational views given later are chosen (cf. sections 3.1, 4.2, and 7.2).

Concepts: Cognitive psychology assumes that cognition is information processing in one's mind, where the basic elements of thought are considered to be conceptual entities, or broadly *concepts* (cf. Murphy [2004]; Schank [1975]). All conceptual entities in memory can be interrelated to categorize conceptualizations of all the present elements in the world. This categorization usually results in forming a web or "ontology" of linked concepts that are grouped together. According to a specific "ontology", one concept can have links to higher level concepts of which the concept itself may be a part. A *wheel* can for example be a part of a *car*, so the concept WHEEL may have thus a link to the concept CAR, which in turn can be seen as a part of the concept VEHICLE.[1] Concepts may also have links in the other direction, namely to their constituent parts (and parts can have parts, and so on).

Concepts are considered here as "the glue that holds our mental world together [..., where] they tie our past *experiences* to our present interactions with the world, and because the *concepts themselves* are connected to our larger *knowledge structures*" [Murphy, 2004, pp. 1; emphasis added]. Therefore, conceptual entities will be basic elements on which all forthcoming discussions are based. However, the difference in the discussions between the "representations" of concepts and the concepts themselves should be clear. There is an extremely vital difference between the "thing-in-itself" and its representation, both on the (Kantian) philosophical level and the computational

[1] To distinguish a real *entity* from a corresponding representation of this real entity as a CONCEPT or constituting conceptual entities, slanted font will be used to typeset the *former* and small caps to typeset the label of the LATTER. This convention is also maintained throughout the rest of the thesis.

level. Form is not substance: e.g. the blueprint is not the house, and the receipt is not the dish (cf. [Fauconnier and Turner, 2002, pp. 4]).

Roles of KR and Conceptual Entities: In a seminal discussion of the KR essences, Davis et al. argue about five distinctly different roles that a representation plays, and claim that each of which places different and at times conflicting demands on the properties a representation should have (cf. [Davis et al., 1993] and [Wagman, 1996, Ch. 1]).

The very first role KR plays is being considered "most fundamentally a surrogate, a substitute for the thing itself, that is used to enable an entity to determine consequences by thinking rather than acting" [Davis et al., 1993, pp. 17]. I agree with Davis et al. that "all representations are *imperfect approximations* to reality" [Davis et al., 1993, pp. 19; emphasis added], which is precisely one of the major sources where error may come from. Nevertheless, this is also the exact reason why I find it better for a GI model to represent knowledge using conceptual entities. Remember that "conceptual entities" refer to existing knowledge pieces (or beliefs) in the agent's KB. These can be categorized to constitute an approximate interpretation that reflects a conceptual understanding (i.e. conceptualization) of an ideation or a conception. A representation of the latter is referred to, here, as a "concept".

By representing the KB of a cognitive agent using conceptual entities, any knowledge entity (i.e. beliefs) that the agent may acquire will affect the "understanding" (i.e. the representation) that the agent has already acquired about the specific "concept" (i.e. the concept that the belief contributes to its description). In general, a knowledge entity affects the relationships among the concepts that share such an entity too. Moreover, using conceptual entities helps in overcoming the challenge of how detailed should a representation be. The model designer will be responsible for deciding upon the granularity level for representing the KB, but this should not affect the overall KR description structure of the KB (whence, affecting the whole KR level). The presented conventions are needed for the discussions about conceptions and their corresponding concept representations, presented within the context of section 4.2. Parts of the ideas given in Chapters 7 and 8 will also be based on these discussions.

KR plays its second role according to Davis et al. as a set of "ontological commitments", where "selecting any representation [...] unavoidably [make] a set of decisions about *how* and *what* to see in the world" [Davis et al., 1993, pp. 19; emphasis added]. But then, they indicated that their choice of the phrase "ontological commitment is perhaps not precisely correct for what [they] have in mind here, but it is the closest available approximation" [Davis et al., 1993, pp. 32]. I argue that the use of con-

ceptual entities as KR entities collapses the first and second roles of KR (given in Davis et al. [1993]) into only one: an approximate surrogate of ontological commitments. In other words, the use of conceptual entities as a way of representation enables KR to achieve (at least) both its first and second roles in modeling aspects of GI: that is, *KR is an approximate surrogate which is, inherently, also a set of ontological commitments*.

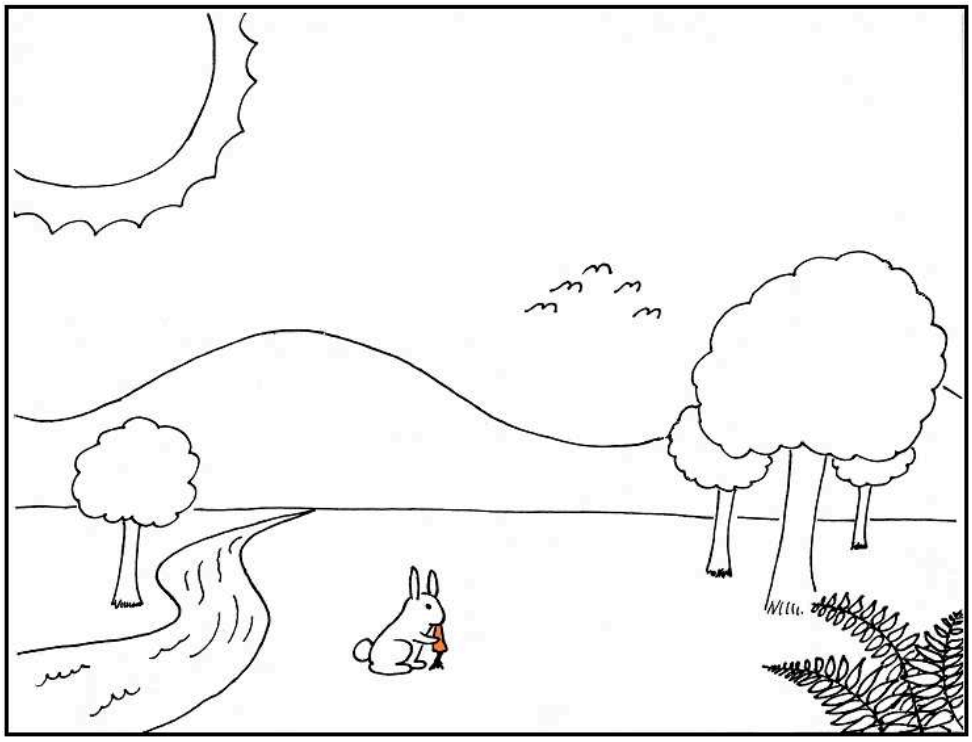
1.3.2 Levels of Representation

Consider the two pictures given in Figure 1.2, which contrast two representational views of the same scene. The pictures emphasize how the same (part of the) world can be represented based on completely different views. Human beings can recognize a wide range of possible representations (e.g. abstract or theory-driven), but artificial agents are not able to do exactly so —a situation that certainly requires a model designer to take several design decisions concerning the representation of knowledge.

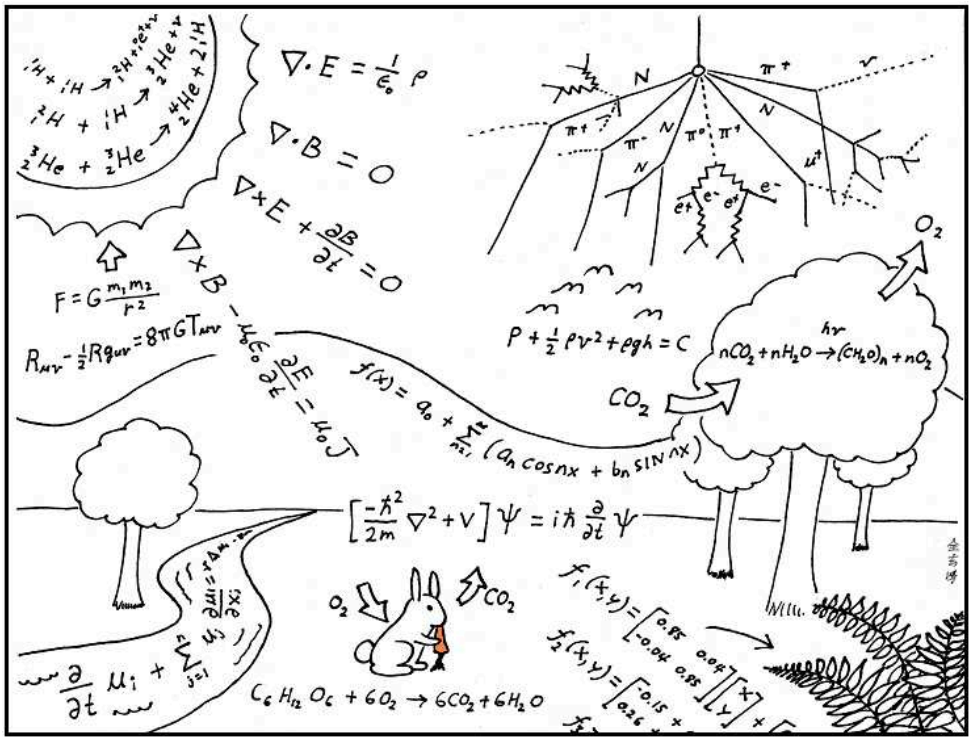
By looking at Figure 1.2a, humans would recognize a depiction of a natural landscape scene, with drawings that illustrate both animate objects (such as flying '*birds*' and a '*rabbit*' holding a '*carrot*') as well as inanimate objects (such as the '*carrot*' itself, a faraway '*mountain*', and the '*ground*', through which part of a flowing '*river*' is also illustrated). People can categorize different parts of the figure only by tracing some curves and lines in the figure (that is, a "processed sensing"). They understand for example that the upper half of the figure reflects the '*sky*', whereas the lower half of the same figure reflects the '*ground*', although both parts are not painted in any distinguishing colors. Also, they can even recognize a '*mountain*' between these two halves. In addition, at the top-left corner of the same figure, the drawing reflects the '*sun*' and an '*aura*' (that may be recognized as the '*corona*') by simply drawing a part of a circle surrounded by a part of another geometrical curve that looks like a hypocycloid of a bigger, unseen circle.[1]

Humans, who can recognize objects (in Figure 1.2a) such as '*sun*', '*rabbit*', '*river*', '*tree*', etc., must have already "learned" a lot in reality about such objects and their underlying conceptions to reach this level of identification. They must have already developed conceptualizations about what these objects are, what their main properties are, what their nouns or meanings imply, how they can be illustrated in drawings like the one in the figure, how they are linked, etc. This development of conceptualizations about objects is a process of developing understanding of the particular conceptions, as well as about the possible interrelations that lie within. Conceptions are not always

[1] In geometry, a hypocycloid is a special plane curve generated by the trace of a fixed point on a small circle that rolls within a larger circle.



(a)



(b)

Figure 1.2: KR is one of the biggest challenges for modeling higher cognition: The world may be seen abstractly or as theory-based. This picture is taken from Abstruse Goose's "World View" (cf. http://abstrusegoose.com/275).

developed in isolation, but also with possible relations to others. One recognizes that the (visible part of the) hypocycloid curve in Figure 1.2a represents the *corona* (of the *sun*), not because one directly has the belief entity that: "the *corona* is illustrated in figures by drawing a geometrical curve that looks like an *aura* in the form of a hypocycloid", but rather because one can "theorize" such an entity from one's previous experience in the KB; from the many "exemplars" of illustrations and images one has previously seen, in which rays that surround the *sun* are illustrated in an analogous way. This is highly related to well-known cognitive scientific views of concept representations, outlined in section 4.2.

Perhaps, a generally intelligent agent may have sensors to perceive the surrounding environment (or even take photo images of it, then store them in the KB). However, in almost all the cases, an agent is not expected to understand the environment by only storing percepts. Conceptual entities have to be arranged as KB contents through ontological relations that relate these entities. The agent would need to have internal information about the several knowledge entities, in order to grasp the kind of scenes given in Figure 1.2a. This is where the issue of using concepts as KR entities becomes very important for modeling GI aspects.

How Deep Should a Representation Approximate Reality? A direct answer to the question "how much to represent?" is always "it depends". It depends on storage capacity, on the processing speed, on the accepted degree of precision, on the importance of optimizing other resources, etc. The decision is crucial, however, because the choice always depends on the more important application or set of applications that need to be frequently performed. Implicitly, the decision also informs the KR engineer "what not to represent" as well. The same representation can be very beneficial in speeding up the computation of concrete processes and efficiently solving concrete problems, but very inefficient with respect to computing other processes and solving different kind of problems. Nonetheless, a decision must be taken anyway, because everything that follows (with respect to the utilization of the KB mentioned earlier) depends on it (also cf. section 4.2).

Figure 1.2b, for example, gives a representational view of the same illustration of Figure 1.2a, but uses a different KB that is richer in terms of representational formulae. On the one hand, the equations shown in Figure 1.2b seem inevitable if one needs not only to study surface features of the objects in the scene but also their deeper characteristics and the physical relations among them. Chemical reactions, for instance, are represented by equations that represent the interactions between objects: e.g. photo-

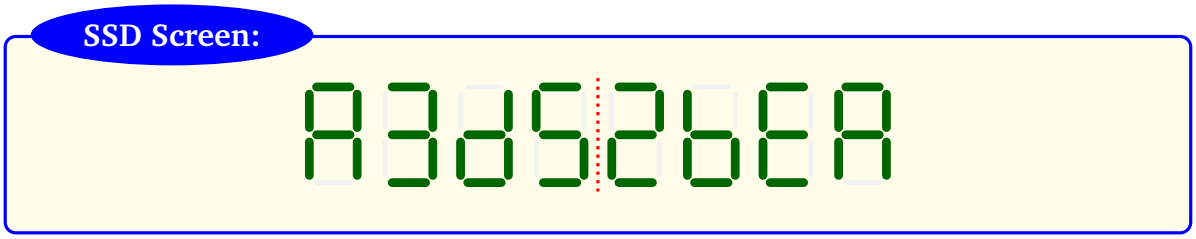


Figure 1.3: A symmetric hexadecimal code: "a3d52bea". The dotted, vertical line represents an imaginary axis of symmetry.

synthesis (by the plants in the scene) uses carbon dioxide (CO_2) to extract oxygen (O_2), whereas the breathing process (by the rabbit in the scene) does the opposite. Natural scientists (like physicists or applied mathematicians, in particular) may think that such lots of details are important for the representation and may in addition be able to identify other equations, such as:

1. the hydrogen fusion in the *sun* (hint: the equations to the top-left),
2. Navier Stokes continuum equations of flow (hint: the water flow of the '*river*'),
3. (biological) combustion of glucose (hint: the *rabbit* eating the *carrot*),
4. photosynthesis (hint: the *plants*),
5. Bernoulli's equation (hint: the flying *birds*),

as well as many other equations (like a Fourier series, a linear system of equations, Einstein's field equations, and Schroedinger's and Maxwell's equations). On the other hand, this formulae-rich representation seems to be only needed in a very limited number of applications, perhaps natural-scientific ones, if at all. But if the rich representation is needed, then it must be used. Otherwise, the difference between the approximate representation and the real existing knowledge could be satisfying. In arithmetic, for instance, the number system used for representing the numerals plays a differentiating role, simplifying some procedures and complicating others. What makes one system preferable over another is not only the way the system elements are represented, but also the processes that can be performed utilizing these corresponding representations. Again, it depends on the applications.

Representing the Environment: Representing the environment is another difficult task for a model designer, since it is of an utmost importance for achieving the functionality (i.e. processing) of some aspects of intelligence. Clark and Chalmers argue that not only can the environment sometimes be an important resource for cognition, but that

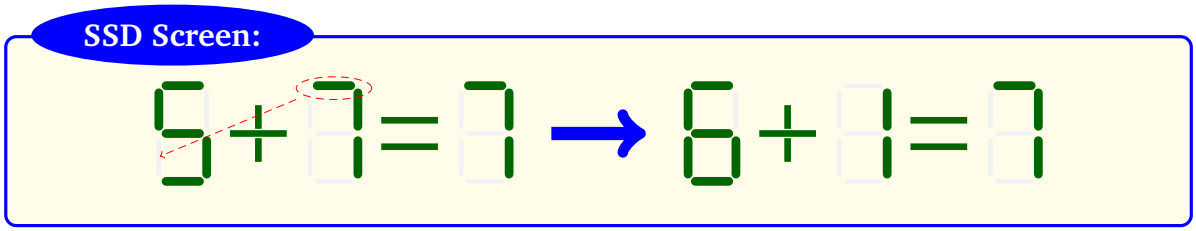


Figure 1.4: How a representation of the incorrect equation: "5 + 7 = 7" can be made into a representation of a correct equation: "6 + 1 = 7" by moving one "segment" from its "location".

bits of it sometimes actually become proper parts of the cognitive system (cf. [Blomberg, 2011, pp. 87–88]). In fact, thinking in terms of a specific representation for the environment may be "the" way to solve problems that can only be circuitously solved. Just think of how difficult it would be to recognize whether or not the hexadecimal codes that correspond to 10-digit decimal numbers, such as the codes "a3d52bea", "bbddbbdd", or "e5a25a23", are symmetric, without representing the hexadecimal codes on a 7-segment-display screen and defining the axis of symmetry to be vertical (cf. Figure 1.3). Now, if we assume the representation has been fixed to represent the conceptual entities as segments on 7-segment-display screens, would not it be considered intelligent if a cognitive agent can figure out how to correct the incorrect equation $5+7=7$ to $5+1=7$ by changing the location of only one segment (cf. Figure 1.4)?

A Summary: To sum up, the above discussion stresses KR as a fundamental issue that must be considered in modeling cognitive agents, while keeping in mind that KR is rather sensitive to many factors (such as the level the representations should be at). Representations in the KB are affected by the driving force to capture as equivalent interpretations of percepts (and their interrelationships) as possible. However, ways to represent knowledge varies depending on many factors, such as the richness level of details that is needed for the modeled applications and processes, as well as how these details are viewed in the first place. In modeling cognitive agents, the designer must take decisions in these regards from the beginning, because the decisions affect not only the details to be represented, but also the ways (or methods) to be followed, in order to model the needed processes. Moreover, in some cases we may not even be able to talk about modeling without also talking about a specific type of representation. The way any piece of information is represented affects almost every utilization to deal with such information, especially when it comes to devising ways of processing knowledge that contain many interrelated pieces of information.

1.4 A General Overview of the Thesis

All discussions have so far provided a collection of general backgrounds. One aim was to give directive hints to cover the restricted view of the thesis, which could help as an overview of related literature. Another aim was to mention selective ideas that will enrich discussions in forthcoming sections. Hopefully, this may also help to answer unasked, rather expected, questions. The remaining chapters in this (first) part of the thesis (namely, Chapters 2, 3 and 4) continue to give rather more specific and condensed backgrounds:

1. In the next chapter, Chapter 2, "analogical reasoning" is presented. It has a central position as a GI aspect that manifests cross-domain intelligent thinking. Examples of artificial systems of analogy-making will also be briefly mentioned, along with their main components and simulated processes.

2. A framework is important to, at least, affirm that a realization of computational solutions to the posited cognition problems is achievable. The selected framework is based on the powerful analogy engine, HDTP. An overview that focusses on HDTP is given in Chapter 3. (But the problems themselves are presented within the second part of the thesis.)

3. Another multifaceted, cross-domain mechanism that greatly helps in explaining various forms of GI is "conceptual blending" (or conceptual integration). Conceptual blending, which is fundamental to computational models of cross-domain thinking and creativity, is thoroughly elaborated on in Chapter 4. The elaboration is pivotal for all the chapters that follow.

The main contribution of the text is to confirm positive answers to the research questions whether multifaceted cognitive mechanisms are necessary, and if they can be utilized to endow cognitive agents with clear signs of GI manifestations in an integrable form of modeling. The second group of chapters gives more elaborate support to such affirmative answers. That part contains the following chapters, which exhibit main contributions and results of the study:

1. Chapter 5 is concerned with modeling manifestations of creativity, as a fundamental GI aspect that helps in stepping further towards achieving computational creativity in AGI models by means of cross-domain mechanisms. The major portion of this chapter is based on a copyrighted, published article that follows the same thematic research trace. (The presented text includes ideas and materials already co-authored as parts of Abdel-Fattah et al. [2012b].)

2. Given that "rationality" is an indispensable GI criterion of human beings (who are the best known generally intelligent exemplar), Chapter 6 puts an emphasis on the possibility to achieve human-style rationality in AGI systems. It also claims that classical systems fail to interpret deviations in such behavior precisely because they lack to utilize higher cognitive mechanisms in appropriate ways. The major portion of this chapter is based on copyrighted, published articles that follow the same thematic research trace. (The presented text includes ideas and materials already co-authored as parts of Abdel-Fattah et al. [2012a] and Gust et al. [2011].)

3. A proposed solution model of how to interpret novel noun-noun combinations, based on representing nouns as concepts and utilizing cross-domain mechanisms, is extensively discussed in Chapter 7. The major portion of this chapter is based on copyrighted, published articles that follow the same thematic research trace. (The presented text includes ideas and materials already co-authored as parts of Abdel-Fattah [2012] and Abdel-Fattah and Krumnack [2013].)

4. Chapter 8 identifies yet an additional benchmark aspect of GI: the capacity of humans to analyze counterfactual conditionals by utilizing cross-domain cognitive mechanisms. The chapter introduces the problem of analyzing counterfactual conditionals, emphasizing that such a crucial trait indicates intelligent and creative reasoning, and needs thus to be considered in modeling AGI systems. Therefore, Chapter 8 also explicates a method to utilize cognitive mechanisms in computationally analyzing counterfactual conditionals. The major portion of this chapter is based on copyrighted, published articles that follow the same thematic research trace. (The presented text includes ideas and materials already co-authored as parts of Abdel-Fattah et al. [2013a] and Abdel-Fattah et al. [2013b].)

Finally, some conclusive remarks and a summarized overview of the presented ideas are given in Chapter 9. In addition to discussions mentioned throughout the various chapters, this conclusive overview includes ideas and materials already co-authored as parts of Abdel-Fattah and Schneider [2013]; Martínez et al. [2011] and [Wang and Goertzel, 2012, Chapter 12]. The chapter quickly elaborates on future research subjects that may be needed to render some broad views more concrete.

2

Analogical Reasoning

Broadly speaking, *analogy making* is an act of comparing potentially dissimilar things, typically referred to as domains, by highlighting the respects in which they may be perceptually, or functionally similar. What is being reflected by this contrasting comparison is an *analogy*; and any type of thinking that relies upon it is usually termed *analogical reasoning*. Reasoning based on analogies is one of the most fundamental cognitive abilities to human thought (cf. Kokinov and French [2003]) and, arguably, to some non-human animals (cf. Bartha [2013]). In this chapter, the basic cognitive aspects of making analogies and its modeling approaches are introduced and discussed.

2.1 Cognitive Science, AI, and Analogy-Making

Analogies play a significant key role in a wide range of problem-solving contexts and in learning, with analogical reasoning providing a heuristic way of reasoning that differs from ways of exact, logical reasoning. This cognitive ability to reason analogically allows a cognitive reasoner to selectively retrieve needed information from the reasoner's memory (or knowledge base), based on a matched situation to the reasoner's current situation. This facilitates the drawing of inferences that the reasoner thinks are specifically related to the current circumstances, because the very idea of making analogies depends on previously encountered experiences (e.g. the knowledge available in, or retrieved from, the reasoner's memory or KB) and not on deterministic, formal rules. Many forms of analogical reasoning have the potential to be formalized and modeled. In fact, various modeling approaches and systems for computing analogies already exist. Computational modeling provided valuable sources of insights that led to a deeper understanding of analogy and the roles it plays in human cognition (cf. [Gentner and Forbus, 2011, pp. 266] and [Hofstadter and Sander, 2013]).

2.1.1 Using the Analogy Label in AI and CogSci

Analogies have been studied from a variety of perspectives. This gives different senses of usages and interpretations of what the term "analogy" itself means, as well as variations of the exact roles played by analogical reasoning in cognitive scientific studies. Three modes of using the analogy label are usually mentioned (cf. Indurkhya [1989]):

Analogy by Rendition: The term "analogy" can be used to render a less-familiar object as being alike in an uncommon way to another more familiar object, or vice versa. A characteristic of this usage mode of analogy is that similarity between objects do not originally exist prior to discovering (or, in fact, synthesizing) an interpretation. In cases of analogy by rendition, no further (actual) inferences about any of the objects can be drawn from the analogy. For example, the words "*desserts*" and "*stressed*" can be seen similar according to the serendipitous observation that they are the same if spelled backwards. Going further by inferring "*that is why we eat desserts when we feel stressed*" is at least as synthetic as the observation. To the best of my knowledge, analogy by rendition has not been explicitly addressed by computational models of AI so far, but systems do exist that exhibit features very much like analogy by rendition (cf. [Indurkhya, 1989, pp. 226]).

Proportional Analogy: A "*proportional analogy*" indicates the proportion of two pairs of four general objects A, B, C, and D in a statement: "'A' is to 'B' as 'C' is to 'D'", denoted "$(A : B) :: (C : D)$". An analogical reasoning problem in this usage mode tries typically to find a possible D that balances the proportion suggested by the analogy "$(A : B) :: (C : D)$". For example, "*dog is to bark as cat is to* ... (what?)" or "*sun is to the solar-system as nucleus is to* ... (what?)", and so on. Proportional analogies are probably closest to the original Greek meaning of *analogia* —(ana- + logia) that means "proportion"— which seems to give an explanation of why 'discovering an analogy' means 'discovering a relative proportion', and then applying it (e.g. to find a D alternative; also cf. Figure 2.2 on page 48). For proportional analogies, there do exist computational systems that can model it. Reitman et al.'s information processing model of thinking can be historically considered the first computer program that solves proportional analogies (cf. Reitman et al. [1964]). Ideas from Evans's system, dedicated entirely to solving geometric proportional analogies (cf. Evans [1964]), are crisply presented in section 2.3.

Predictive Analogy: This is the particular mode of analogy-based reasoning that the majority of analogy-making systems are studying and modeling. Unlike propor-

tional analogy, which is considered an end in itself, the predictive mode of analogies usually serves as *a heuristic in some larger context of problem solving* (cf. [Indurkhya, 1989, pp. 226]) —and will be utilized as a heuristic in the rest of the thesis' contexts as well.

A predictive analogy plays at least two crucial roles in cognition, since it typically helps in enriching knowledge about a recent conceptualization (referred to as the target) by:

1. observing possible similarities with an already-known conceptualization (called the base or source) retrieved from memory, and then
2. trying to transfer new information from the already-known to the newly encountered.

The ordering in the latter enrichment process is specific, which is what distinguishes predictive analogies from analogy by rendition: in the latter it is neither the case that similarities are always observed *before* information is transferred, nor is it always the case that information is transferred *from* the already-known to the recently encountered. Moreover, observing 'possible' similarities when working out predictive analogies is an indication that other possibilities of similarities may be drawn in a reasoning process (e.g. by having a different context or by changing the reasoner's KB). This allows predictive analogies to provide more "predictions" of similarities between the given conceptualized domains than the mere "rendering" in analogy by rendition (which seems intentional or synthetic in its nature). Throughout the rest of this text, it will become more evident that these predictions, in turn, play other crucial roles in modeling clear signs of cognition and GI.

An obvious aspect of higher-level cognition would be the ability of a cognitive agent to figure out whether or not the same 'system of relations' holds among the entities across their domains (cf. Gentner [1983]; Jee et al. [2013]). The domains are always assumed to be structured representations of conceptual entities in some KR language (cf. section 1.3.1 and section 4.2). Typical cases of analogy involve two of such structured domains or concepts:

1. one, that is referred to as the "*source*" (or the base) domain, about which more is already known, and
2. the other, that is called the "*target*" domain, which indicates a recently encountered situation or a new experience.

Intelligent reasoning allows humans to compare the underlying relations between entities in each domain, roughly highlighting distinctly notable (or unnoticeable) similarities (cf. Gentner [1989]; Gentner et al. [1993]). "For instance, when a person is told that the interior of the earth is like a peach, they are unlikely to assume that a substance with the taste, color, and texture of a peach can be found beneath the earth's crust. Rather, they understand the message to be that comparable spatial relationships hold within each of the two objects" [Jee et al., 2013, pp. 2–3]. The following discussions include somewhat extensive overviews of why analogy is important to cognitive scientists, and how prominent AI models succeed in computing several forms of it.

2.1.2 General Motivating Examples

As an important element of cognitive development, humans develop their ability to make analogies from the very early ages, in parallel to developing their comprehension of every conceptualization they encounter over time —an issue that further supports selecting analogy-making to serve as the basis for numerous other kinds of human thinking (cf. Chapters 5, 6, 7, and 8). Starting with "the simple ability of babies to imitate adults and to recognize when adults are imitating them", progressing to "children's being able to recognize an analogy between a picture and the corresponding real object", and, ultimately, culminating in "the adult ability to make complex analogies between various situations" (cf. [Kokinov and French, 2003, pp. 113]).

People employ analogies in many real-life situations that range from solving proportional analogies (e.g. the type of exercises in proportional analogy that became a standard part of various intelligence tests; cf. [Indurkhya, 1989, pp. 217]), to understanding deep, metaphorical, pictorial conceptualizations (cf. Figure 3.6). In fact, many cognitive processes involve analogy-making in one way or another, and this happens in a variety of situations such as:

- diagnosing an illness of a patient by a doctor, based on prototypical symptoms or on blood, or urine tests,
- founding a new formal or scientific theory, such as proving a new mathematical formula (cf. Kerber [1989]), or solving a problem based on the way a similar one was previously solved[1],
- writing a poet, composing a song, or creating any new piece of art (e.g. the shapes presented in Penrose and Penrose [1958], and Escher's implementation of them

[1] A simple example in education would be, e.g. the practicing of mathematical induction by students in discrete mathematics lessons.

in [Seckel, 2004, pp. 92] and [Seckel, 2004, pp. 91]; also see Abdel-Fattah and Schneider [2013] and Chapter 5),

- recognizing the same content or "spirit" of drawn images (e.g. perceiving a square in a painting as a human's head) or hand-written texts (cf. [Hofstadter and the Fluid Analogies Research Group, 1996, Chapter 10, in particular Figure 10.2 and 10.4, pp. 413 and 418, respectively]), even if the styles look different, or

- supporting or refuting judgements following a line of argumentation (e.g. "arguing in court for a case based on its common structure with another case" [Kokinov and French, 2003, pp. 113], and assimilating a new case in Islamic law guided by earlier cases [Sowa and Majumdar, 2003, pp. 18–19]).

Many other daily life situations can be mentioned, like learning (e.g. a foreign language), teaching (e.g. natural or programming languages), understanding metaphors and making them (cf. [Gust et al., 2006; Indurkhya, 1992]), or communicating with others by conveying and comprehending ideas or even emotions (e.g. by perceiving similarly experienced facial expressions). These are just a few of the many, many application scenarios that consolidate analogy-making as being a core aspect of cognition that helps in acquiring, communicating, and creating ideas (cf. [Gentner et al., 2001; Kokinov and French, 2003]). All of the above (and more) have rendered cross-domain reasoning based on analogy-making into an important aspect of intelligent thinking that seems to be essential in computational models of AGI.

A Core Cognitive Realm: Analogies attracted the attention of many researchers in cognitive science and AI since these disciplines started (cf. Indurkhya [1989]), but the way in which analogy is approached has changed during the last few decades (cf. [Clement, 2008; Gentner et al., 2001; Holyoak and Thagard, 1996; Schwering et al., 2009b]). For a relatively long period of time, analogy was considered merely as a special case of reasoning that is rarely applied, but that is giving rise to creative solutions and poetic writing. Important roles have long been ascribed to analogical thinking, in particular, in:

- problem solving (cf. Helman [1988]; Pólya [1954]),

- measures of intelligence (cf. Sternberg [1977]),

- the development of concepts (cf. Lakoff and Johnsen [2003]), and

- creativity and scientific discovery (cf. Clement [2008]; Holyoak and Thagard [1996]).

Then, science studies provided more evidence that analogy was used by preeminent scientists in formative research meetings in leading laboratories [Clement, 2008, pp. 9], and that analogical reasoning can be important in both learning scientific models as well as transferring this learned knowledge to new, unfamiliar problem domains (cf. [diSessa, 1988; Rumelhart and Norman, 1981]). Quotes and discussions by prominent scientists, especially in the philosophy of science, clarify the respect they all have for the role of analogies, and suggest that analogies may be a source of hypotheses in science (cf. Campbell [2013]; Einstein and Infeld [1966]; Gentner et al. [1997]; Hesse [1966]). They argue that scientists "not only find patterns of empirical observations in their work", but also "think in terms of theoretical explanatory models", which "constitute a different type of hypothesis than empirical laws" [Clement, 2008, pp. 14].

Analogies attracted the attention of several cognitive science groups over the last few decades, especially those working in the AI field (cf. [Bartha, 2013]). Extensive research on analogy has been conducted, focusing on the central importance of analogical reasoning in diverse areas (like perception, learning, memory, language, and thinking). As these areas were realized to depend on relational matching, "analogy-making moved from the periphery to the core of human cognition" [Schwering et al., 2009b]. Some prominent figures even consider analogy as "the leading fact in genius of every order" [James, 1950; Mitchell, 1993], as "one of the ultimate foundation-pillars of the intellectual life" [James, 1950], or as '*the*' most important aspect of cognition [Hofstadter, 1995; Hofstadter and Sander, 2013]. Specifically, Hofstadter's view is that "analogy-making should not be thought of as a special variety of *reasoning*", but rather "the very blue that fills the whole sky of cognition" [Hofstadter, 2001, pp. 499; emphasis original]. According to his view, analogy is nearly everything and is "a core of cognition".

2.1.3 The Structure-Mapping Theory (SMT)

The literature in cognitive science provides several theories and experiments on analogical reasoning (cf. Forbus et al. [1997]; Gentner [1983]; Helman [1988]; Holyoak and Thagard [1996]; Hummel and Holyoak [1997]). Whilst earlier models tended to understand the basic "constraints that govern human analogical thinking" [Hummel and Holyoak, 1997, pp. 458], the predominant objectives of recent theories have become to uncover psychological mechanisms of the sub-processes involved in analogy making, and to model the functioning of these sub-processes (cf. Bartha [2013] and

section 2.2.1). The experiments aim to support the theories by reporting on either (i) completing "presented analogies", in which parts of analogies are given to human participants and are asked to complete these parts, or (ii) finding "spontaneous analogies", where the participants initiate and form the entire analogy (cf. Clement [2008]).

As for theories, Gentner's "*structure-mapping theory*" (SMT) is an influential theoretical framework for analogy that introduces and describes psychological processes utilized in analogical reasoning (cf. [Forbus et al., 1997; Gentner, 1983, 1989]). In order to founding analogy as being a way of focusing on relational commonalties, independently of the objects in which those relations are embedded, the SMT explicitly emphasizes the importance of both "structural similarity" and "structural consistency" of objects in the source and target domains by stressing that:

1. good analogies are determined by mappings of relations and not attributes (cf. [French, 2002, pp. 202]), and

2. mappings of coherent systems of relations are preferred over mappings of individual relations (cf. [Hofstadter and the Fluid Analogies Research Group, 1996, pp. 276]).

This is revealed in SMT through the central principle of "*systematicity*" (cf. Clement and Gentner [1991]), which shifts the emphasis in analogy-making to the structural similarity between the source and target domains. One observes systematicity by keeping "relations belonging to a systematic relational structure in preference to isolated relationships" [Falkenhainer et al., 1989, pp. 6]. The principle of systematicity states that people "prefer to map connected *systems of relations* governed by higher-order[1] relations with inferential import, rather than isolated predicates" [Gentner, 1989, pp. 201; emphasis original].

Example of SMT's Systematicity Principle: The role that systematicity plays can be explained by considering a concrete example, such as a famous analogy between the solar system (as the source domain) and the Rutherford atom model; namely: "the atom is like our solar system" (cf. Gentner [1983]).[2] A conceptualization of the source domain

[1] Keep in mind that constructs like (first-order) functions or predicates connect items (of a representation) with certain relationships, and perhaps give results of some sorts (like in the case of function constructs). A construct can also be of an order higher than 1, so it has at least another construct of a lower order as an argument (or gives such a lower-order construct as output in the case of a higher-order function).

[2] Section 3.2 elaborates more on this and presents other concrete analogy domain examples. It also describes specific representations of higher-order constructs in more detail.

would involve entities (e.g. the sun and the planets), as well as properties and (higher-order) relations between those entities (e.g. "the sun is yellow", "the planets revolve around the sun", and "the sun attracts planets, which causes them to revolve around it"). Similarly with respect to the target domain, a conceptualization would involve entities (e.g. the nucleus and the electrons), surface properties, and relations (e.g. "the electrons revolve around the nucleus"). In this sample analogy situation, the intended inferences about the atom being like the solar system concern more the relational structure than the mere surface properties of the involved entities (cf. [Gentner, 1983, pp. 159]). By predicting the sun as analogous to the nucleus and the planets as analogous to the electrons, the nucleus plays a role (w.r.t. the atom model of Rutherford) like that of the sun (w.r.t. the solar system) not because both are massive or have similar colors, for example, but rather because one can observe that electrons revolve around the nucleus just as planets revolve around the sun. Higher-order relations that are present in the source, such as "the sun attracts planets, which causes them to revolve around it", can be mapped to the target (that is, "the nucleus attracts electrons, which causes them to revolve around it").

Indeed, cognitive scientific studies consolidate the principle of systematicity by showing that people do not simply fetch any isolated fact from the source (that does not exist in the target) and infer a counterpart in the target. They rather infer a counterpart to a fact that is connected (in the source) via a higher-order relation to other matching facts (cf. Clement and Gentner [1991]) —an implication that inferences are implicitly oriented to find a larger matching system than isolated facts (cf. [Gentner and Forbus, 2011, pp. 266]). Thus, structural consistency requires that not only should objects in the source have correspondences with objects in the target, but arguments of corresponding predicates should also correspond. Alignments are found that parallel (as much as possible) the structure of connectivity among the objects in the respective domains, with inferences projected from the source to the target based on alignments. On the basis of these alignments too, further inferences are projected from source to target.

Cheng and Holyoak's Proposal: In contrast with Gentner's proposal that people prefer "structural consistency" in analogical reasoning, Cheng and Holyoak propose —based also on experimental results— that "goal relevance" is responsible for determining, for example, what is selected in the mapping of an analogy. According to Cheng and Holyoak's pragmatic focus, people are more oriented in analogical reasoning towards attaining goals (cf. [Cheng and Holyoak, 1985]). Holyoak and Thagard integrate both ideas (i.e. the structural and the pragmatic) in their approach to analogical map-

ping, commonly known as the *multi-constraint theory* (cf. Holyoak and Thagard [1989, 1996]). Holyoak and Thagard also focus strongly on analyzing the mapping process as the most difficult process to account for, arguing that it is easier to find a possibly analogous case but not so easy to find good or meaningful analogy (cf. [Holyoak and Thagard, 1989]). They argue that evaluating the soundness of an analogy is a more complex and goal oriented process (cf. [Clement, 2008, §2.1.3]).

In any case, the SMT and its emphasis on the structural aspects of analogical mappings has been, and still is, more influential in contexts ranging from child development to folk physics (cf. French [2002]; Gentner et al. [1997]). Furthermore, the family of computational models that are based on SMT have been traditionally considered the most influential, in particular because numerous psychological experiments have confirmed the crucial role of relational mappings in producing sound and convincing analogies (cf. [Kokinov and French, 2003, pp. 116]). Assumptions that underly the computational implementations of SMT are given by Falkenhainer et al. in their "*structural mapping engine*" (SME) (cf. [Falkenhainer et al., 1989]), and will be elaborated on in section 2.3.

Sub-Processes in Analogy-Making: With regard to sub-processes, few interrelated phases traditionally characterize the cognitive process of analogical thinking, when one is exposed to a new (target) situation. According to Falkenhainer et al., for instance, the SMT views analogical processing as being decomposed into three consecutive stages or phases (cf. [Falkenhainer et al., 1989, pp. 3–4]), on which models of similarity-based retrieval can be built (cf. Forbus et al. [1995] and section 2.2):

Access: The access phase retrieves a target description from the long term memory (LTM), based on a given base (source) situation.

Mapping and Inference: This phase constructs correspondences between the source and the target, and includes candidate inferences acceptable by the analogy. The candidate inferences specify what additional knowledge in the source can potentially be transferred to the target.

Evaluation and Use: In this phase, a quality of the match is estimated, according to three kinds of assessment criteria: structural (e.g. number of similarities and differences), validity (e.g. inferences must be checked against current world knowledge), and relevance (e.g. whether or not the analogy is useful to the reasoner's current purposes).

Broadly speaking, the number, labels, and order of phases may differ from one theory presentation to another, or from one situation of analogy-making to another.[1] As another example, the following (also three) steps are suggested as the core of the analogy-making process in [Schwering et al., 2009a, pp. 2], which slightly differ from Falkenhainer et al.'s:

Retrieval: similar to the access phase in Falkenhainer et al. [1989], retrieval identifies a source domain to which the new (target) situation can be related,

Mapping: this step concentrates on (only) establishing a mapping between the source and target cases, based on their common structure or relations, and

Transfer: a translation of information between the two domains takes place in this step, which also (possibly) infers new elements in the target (based on the mapping).

Cognitively speaking, the "mapping" may be seen at the core of the process of establishing a feasible comparison between dissimilar domains in analogical reasoning. In this comparison, (i) the "retrieval" initiates a contrasting process (which is triggered depending on suggestions for the source by the familiar background knowledge of the reasoner), (ii) the "mapping" characterizes and formalizes this contrast (by aligning a system of elements in one domain to a corresponding system of elements in the other), then (iii) the "transfer" stresses and finalizes the process (opening the door to further possible enrichments of the reasoner's background knowledge, based on analogical inferences or on creative assumptions; also cf. Chapters 5, 7 and 8). Still, of course, steps other than the previous ones could as well be considered. For example, the cognitive thinker may need to

- refine the understanding of the source before applying results to a target problem (cf. [Clement, 2008, pp. 24]),
- re-represent one (or both) of the base and the target domains in terms of the other one (cf. section 3.1.3),
- evaluate the soundness of the mapping that the thinker initially focused on, or
- abstract the schemes of the involved cases (cf. Gentner and Forbus [2011]).

[1] Gentner et al. went even further and suggested that different sub-processes utilize different kinds of similarity (cf. [Gentner et al., 1993, pp. 527]). The SMT also distinguishes analogies from "literal similarity" statements or other distinct types of comparison.

The mostly relevant steps involved in artificially computing analogies simulate the above-mentioned sub-processes. They are depicted in Figure 2.1, and will be elaborated on in the next section, which changes the focus from cognitive science to AI by giving a crisp overview on formalizing and modeling analogical thinking.

2.2 Computational Models for Analogy-Making

Is analogical thinking amenable to computing? Unlike usually encountered challenges in searching for a working definition of "intelligence" (cf. section 1.2.1), there is a wide agreement on specifying the major elements underlying one of its core aspects —namely making, and thinking in terms of, analogies. Analogical reasoning may be considered a domain of non-formal reasoning (cf. Clement [2008]; Sowa and Majumdar [2003]), but existing models show that 'analogy-based thinking' is more amenable to specification and formalization than the multifaceted 'intelligent thinking'. Many elements of analogy-making can be characterized, formalized, and modeled. Analogy-making can even be effectively or efficiently computed. In addition, essential sub-processes that are thought to be involved in making analogies are widely agreed upon by the adherents.

Analogy-making, be it human-based or computational, is typically conceived of as precisely involving:

1. "*two domains*", called the "*source*" (or base) domain and the "*target*" domain: these can be represented and manipulated within models in many ways —in this thesis, conceptual entities will be represented using a first-order language (cf. sections 3.1, 3.2, and 4.2); and

2. (representational or structural) "*connections*" between the individual entities and their relations in these domains. This can measure the extent to which the domains are comparably similar: the connections are modeled by a matching that is typically taken to be a mapping from the known source domain into the novel target domain. Depending on the KR used to model analogy-making, an analogical mapping relates entities in the source with others in the target.[1] This is how analogy efficiently provides a computational way as "a basic mechanism for effectively

[1]There is a controversial discussion whether or not this mapping relations or assignments must define a function. Broadly speaking, it does not necessarily have to be a (partial) function unless all the assignments are one-to-one (like in SME; cf. section 2.3), but assignments in general can be allowed to be one-to-many or many-to-many (cf. section 3.1).

connecting a reasoner's *past* and *present* experience" [Hall, 1989, pp. 39; emphasis added].

2.2.1 Processes Involved in Computing Analogy

When the modeling of analogy-making as a cognitive process is considered, it is traditional to decompose it into multiple "*abstract processes*" that reflect the stages of computing analogical reasoning, inspired by characterizing cognitive phases of analogical thinking mentioned in section 2.1.3. In addition to its cognitive-plausibility, this decomposition helps in modeling the father process of analogy-making in AI by modularizing it (i.e. making the process consists of concrete modules to facilitate implementations). The following listing tries to systematically debrief guidelines to (ideally) building a general framework for analogy-making, by combining and abstracting proposals of several eminent works (such as e.g. [Hall, 1989], [French, 2002], [Kokinov and French, 2003], [Schwering et al., 2009a], and [Gentner and Forbus, 2011]).[1] The given schematization suggests that a computational framework for the main process of analogy-making may need to (ideally) employ (as many as possible of) the following sub-processes (cf. Figure 2.1):

Representation-Building: *Producing* or formalizing representations based on input. The sub-process of "representation-building" is mostly absent in cognitive models of analogy-making, and is typically achieved by supplying (hand-made) representations into the model. Certain models may produce high-level representations based essentially on unprocessed input, and attempt to build context-sensitive representations during the "mapping" sub-process (e.g. the Copycat model [Hofstadter, 1984; Hofstadter and the Fluid Analogies Research Group, 1996]).[2]

Re-Representation: In contrast to "representation-building" which comes at the beginning of analogy-making, "re-representation" allows the modification and adaptation of the original representations during analogy-making (but it may not be clear at which point during analogy-making). The idea is that, in most of the times, structural commonalities characterizing potentially analogous domains are

[1] A description of some involved processes in analogical reasoning from a cognitive scientific perspective is given in Forbus et al. [1997] and Gentner [1983], whilst Gentner and Forbus [2011] and Kokinov and French [2003] focus more on summarizing the widely known computational models that implement these processes.

[2] This step can be essential when a cognitive system needs to connect several capacities, such as linking vision with reasoning (e.g. perceiving some object first, then building representations and finding analogy, and finally drawing inferences or creating new concepts).

not obvious in advance, but become more visible as a result of the main process of analogy making itself (cf. Schwering et al. [2009a]). During the establishment of an analogy, it could be very essential to change the representation of one or both domains to allow the discovery of implicit common structure (cf. Clement [2008]; Indurkhya [1992] and the first analogy situation in section 3.2.1).

Retrieval: *Recognition* of a candidate, analogous base given a target description. The "retrieval" sub-process seems to be a rather difficult step, and it heavily depends on the reasoner's background knowledge.[1] However, it is an important step in finding and accessing an analogous base case from permanent memory (cf. [Forbus et al., 1997; Gentner, 1983]). Retrieval is usually guided by the shared scheme of (familiar) properties between an already-known base (source) and a newly encountered target —the so-called "superficial similarity". *Superficial similarity* has been extensively studied experimentally and found to play the major role in the retrieval of a base for analogy (cf. [Kokinov and French, 2003, pp. 114]), since a generally shared scheme makes it easier for a reasoner to retrieve a suitable source in an analogical reasoning process. Gentner et al. give empirical evidence (cf. Gentner et al. [1993]) that indicates the higher tendency of a reasoner to retrieve items from memory, based more on surface similarity (i.e. accessing objects and attributes) than on relational similarity (i.e. accessing common relational structures). In many models, retrieval is based on exhaustive search of "*long term memory*" (LTM) and on the assumption that old memory episodes have "context-independent, encapsulated representations".[2]

Mapping: *Elaboration* of an analogical mapping generated between corresponding entities in the structures of the source and target domains. The "mapping" sub-process takes as input two structured representations of the source and the target, then possibly produces (cf. [Gentner and Forbus, 2011, pp. 267]):

1. a set of "*correspondences*" that indicate 'what corresponds to what' by aligning individual entities and their relation in the source with counterparts in the target,
2. a set of (analogical) "*candidate inferences*" that follow from the alignment and imply what may be true in one description based on projecting structure

[1] This has been shown, e.g., in the experiments by Gick and Holyoak when participants attempt to solve Duncker's radiation problem (cf. [Gick and Holyoak, 1983, pp. 3]).

[2] An exception is AMBR (cf. [Kokinov and Petrov, 2001]), which relies on context-sensitive reconstruction of old "*episodes*", performed in interaction with the mapping process (cf. [Kokinov and French, 2003, pp. 114]).

from the other, and

3. a "*structural evaluation*" score that provides a numerical measure of how well the domains align (hence, the soundness of the analogy can be assessed).

Unlike the case in "retrieval", relational similarity is preferred over surface similarity in the "mapping", emphasizing the role of Gentner's systematicity principle (cf. section 2.1.3). Unquestionably, the "mapping" sub-process is the core defining sub-process (cf. [Gentner and Forbus, 2011, pp. 267]), and is therefore included in all computer models of analogy-making.

Transfer: The *transfer* of information from the source to the target domain based on the "mapping". Sometimes, "transfer" is considered an extension of the already-established mapping (and, thus, integrated within the "mapping" sub-process).

Evaluation: The *evaluation* of the mapping and inferences in some context of use, including justification, repair, or extension of the mapping. "Evaluation" should establish "the likelihood that the transferred knowledge will turn out to be applicable to the target domain" [Kokinov and French, 2003, pp. 116].

Learning: The *consolidation* of the outcome. This is implemented in only few models of analogy-making, despite the fact that "analogy-making is clearly a driving force behind much learning" [Kokinov and French, 2003, pp. 116].

Abstraction: The *generalization* that may accompany the establishing of a "mapping". Psychologically, the comparison of the domain descriptions of the source and target can lead to a generalization, but it is still an open question how and when this happens [Gentner and Forbus, 2011, pp. 272]. The "results of comparison may be stored as an abstraction, producing a schema or other rule-like structure" [Gentner and Forbus, 2011, pp. 267].

To the best of my knowledge, no existing model incorporates all the sub-processes altogether. Models rather focus on basic sub-processes (as also pointed out in Kokinov and French [2003]). The precise modeling steps constituting one concrete computational framework or another may slightly deviate from the previous schematization (e.g. Chalmers et al. [1992]), but it is highly unlikely that the main functions are completely distinguishable.

Labels of the Sub-Processes: The names of the sub-processes listed in the given scheme are mainly based on Kokinov and French's, Schwering et al.'s, and Gentner

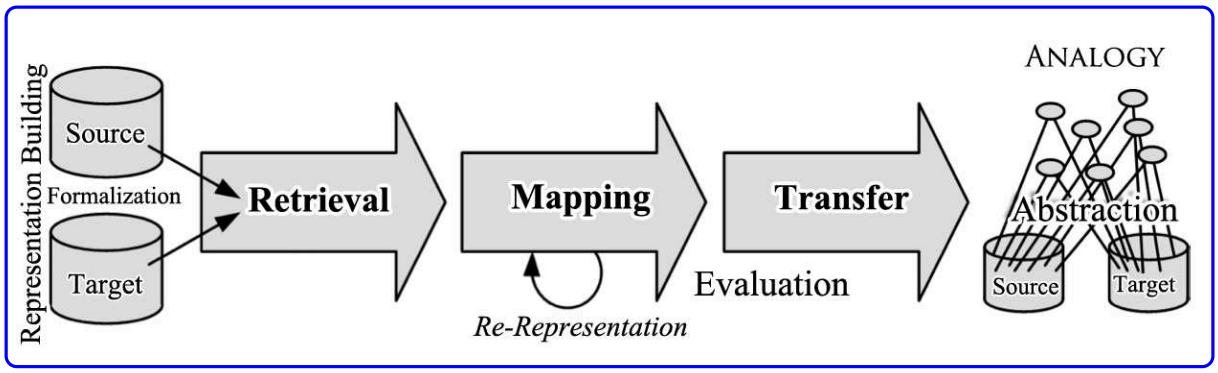


Figure 2.1: A depiction of the phases of analogy-making based on the debriefing given in section 2.2.1 and Schwering et al.'s (cf. [Schwering et al., 2009a, pp. 3]).

and Forbus's decompositions. However, (i) recognition, (ii) elaboration, and (iii) consolidation were initially presented in Hall [1989], then listed again in French [2002], to reflect the essence of (i) retrieval, (ii) mapping, and (iii) learning, respectively. French considered "transfer" and "evaluation" as one sub-process, unlike Kokinov and French who in addition (i) added the process of "representation-building", (ii) renamed Hall's "recognition" to "retrieval", and (iii) renamed Hall's "consolidation" to "learning". French also pointed out the additional suggestion of Chalmers et al. to include *dynamic* representation-building mechanisms and parallel sub-process interaction in such a basic framework. The sub-processes given in Gentner and Forbus's decomposition of analogy-making are: "retrieval", "mapping", "abstraction", and "representation". Gentner and Forbus's "abstraction" (or "generalization") sub-process is of a special interest to this thesis' cognitive-scientific focus, and will be extensively discussed later. Schwering et al. presented the framework for analogy-making, on which this thesis' contributions are based, namely HDTP, which is presented in section 3.1. The previously mentioned sub-processes will also be recalled when introducing and discussing HDTP (cf. section 3.1).

2.2.2 Modeling Approaches

Many scientific schools emerged in cognitive science and AI that concentrated on theoretically suggesting frameworks on how analogy-making functions in principle, and experimentally realizing the proposed functionality by developing computational models and systems. There are currently many models, which may be categorized depending on their chief characteristic features (cf. [Kokinov and French, 2003] and [Gentner and Forbus, 2011, Table 1; pp. 268]). In particular, according to the underlying architectural paradigm, the models are usually attempted to be grouped into three classes:

Connectionist Models: Sub-symbolic processing is generally characterized by (fuzzy)

constraints affecting continuous, distributed representation tokens. Systems that are based on the connectionist approach employ *overlapping patterns of activation* in a neural network in order to represent objects, relations, and episodes. Thus, analogical reasoning in such systems is mostly based on the *relational interconnections* between distributed tokens (e.g. neurons of an artificial neural network [Jain et al., 1996]) and the propagation of activation patterns among these tokens. The *distributed representations* provide an internal measure of similarity, which is what allows overcoming the problem of similar, but not identical, relations in a relatively straightforward manner. According to Kokinov and French, this latter inherent capability of modeling relational structures "is crucial to analogy-making and has proved hard for symbolic models to implement" [Kokinov and French, 2003, pp. 117]. A glimpse of cognitive adequacy is inherently maintained in connectionist models: they have drawn their inspiration from the computational properties of neural systems, so the proposed functionalities of human brains are assumed to be artificially simulated by networks of neuron-like processing units. They are also inherently capable of easily learning and adapting. However, distributed representations make it extremely difficult to trace what actually happens in connectionist models to achieve a solution. One of the good examples in this category of models is Holyoak and Thagard's ACME (cf. [Holyoak and Thagard, 1989]).

Symbolic Models: Symbolic processing is generally characterized by hard-coded, explicit rules operating on discrete, static tokens. In systems based on classical symbolic models, the mechanisms of representing, storing, and processing information employ *separate local representations* of objects, relations, propositions, episodes, etc. Thus, analogical reasoning in such systems is explicitly based on manipulating *symbol structures*, which is what makes these systems well-equipped to process and compare the complex structures required for computing analogy-making. Selected representatives of this type of models, such as SME (MAC/FAC) and HDTP will be further discussed (cf. section 2.3 and section 3.1, respectively).

Although it is possible to track and reconstruct the implementations in symbol-based models, they are traditionally criticized for lacking a needed glimpse of cognitive adequacy (e.g. they do not mimic the way the human brain seems to work nor the way humans adapt to unforeseen situations). Having already brought the subject, it is worth clarifying that the framework, on which the thesis' contributions are based is symbolic (cf. section 3.1), yet has successfully proved cognitive

plausibility in a variety of application scenarios.[1] The focus in this text is on emphasizing the modeling possibility of cognitive adequacy, no matter what the type of the model is. This emphasis is based on (and inspired by) the way the proposed solutions follow in order to model cognitive abilities, not the other way round.[2]

Hybrid Models: Systems that are based on hybrid models combine the symbolic and sub-symbolic approaches by combining symbolic representations with connectionist activations. They are based on the ideas that (i) "cognition is an emergent property of the collective behavior of many simple agents", and (ii) "high-level cognition emerges as a result of the continual interaction of relatively simple, low-level processing units, capable of doing only local computations" [Kokinov and French, 2003, pp. 115–117]. By combining principles from the two approaches into a hybrid approach, some of the typical drawbacks of one can be recovered by some of the typical advantages of the other. Hybrid models would in general be of higher cognitive adequacy levels.[3] Of this type of models, AMBR (cf. [Kokinov and Petrov, 2001]) and Copycat (cf. [Hofstadter, 1984]) are two famous representatives.

2.3 Concrete Symbol-Based Systems

A quick overview is intended in the following to cover main underpinnings, on which the modeling of two familiar symbol-based systems of analogy-making are based. The first model is commonly accepted as being historically the first advanced computational system that solves a special type of analogies; namely proportional analogies (cf. section 2.1.1), whereas the second has always been inspiring computational models of analogy-making systems that are based on the SMT (cf. section 2.1.3). In addition to continue the literature overview in this chapter, the purpose of this listing is to cover ideas that underpin further discussions about the computational modeling of analogy-

[1] Overviews of such scenarios are given in, e.g. Abdel-Fattah and Schneider [2013]; Guhe et al. [2011, 2010]; Martínez et al. [2011, 2012], but others will further be explicitly detailed in the forthcoming chapters. Section 4.2.2 elaborates in particular on this issue of cognitive plausibility when representing concepts.

[2] After all, I think that the best models should combine key features of symbolic and connectionist models (the latter models are not discussed in the current text, though).

[3] Despite the fact that some disadvantages may remain as such in hybrid models (like the problem in connectionist approaches to trace how a solution is achieved), it is still important, in my opinion, to utilize some low-level processes in achieving high-level cognition. Mitchell shows, for example, that analogy-making is related to low-level processes and that high-level perception is an emergent phenomenon arising from large numbers of low-level, parallel, non-deterministic activities [Mitchell, 1993]. However, this issue is not discussed here (but cf. Chalmers et al. [1992]).

making. Table 2.1 summarizes key features of the systems (as well as of HDTP, which is extensively discussed separately in section 3.1).

The ANALOGY System: Early computational paradigms in AI was aiming at getting machines to solve problems, by implementing any tasks that imitate intelligent thinking (cf. [Ringle, 1979, pp. 1–20]). At least three systems in the 1960's selected analogical reasoning as a proof-of-concept that computational intelligence is achievable (cf. Becker [1969]; Evans [1964]; Reitman et al. [1964]).

A historically popular system is Evans's "ANALOGY" program, which has been developed to solve geometric-analogy problems on the form of proportional analogies. To recall, a "*proportional analogy*" has the general form "$(A : B) :: (C : D)$" (that is, "'A' is to 'B' as 'C' is to 'D'"). In the type of the geometric proportional analogies considered by Evans's system, a problem consists of eight figures, each composed of one or more geometric objects (i.e. comprising elements from the same domain of geometric figures): the first two of them, namely A and B, define the source domain, whereas C and five answer alternatives for D define the target domain. The task is to find the way in which A has been changed to B, then apply the same way of change to figure C in order to select a resulting figure from the five answer alternatives, or indicate that no solution could be found (cf. [Hall, 1989, pp. 44]). Figure 2.2 gives an example of one such a problem: Figure 2.2a shows the three given geometric objects (A, B, and C), and Figure 2.2b shows five (possible D) answer alternatives.[1]

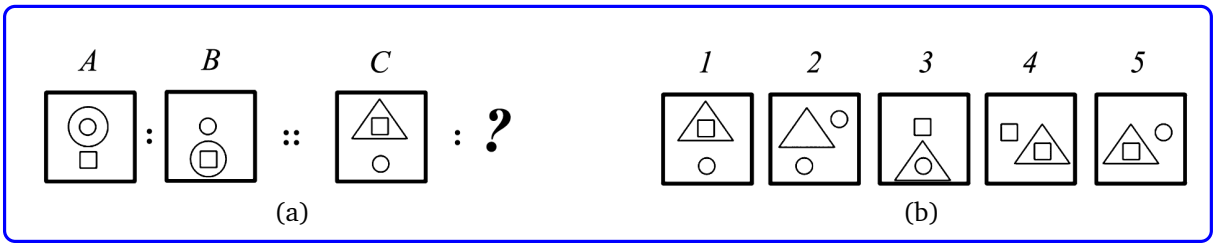


Figure 2.2: An example of a traditional geometric-analogy problem. As addressed by Evans: (a) depicts the three components of the problem (A, B, and C), whereas (b) depicts five answer alternatives (1, 2, 3, 4, and 5). (This is basically Case 5 in [Evans, 1968, pp. 330], for which the answer is alternative 3.)

The ANALOGY system does not take high-level descriptions of the problems, but rather low-level descriptions of each component of the geometric objects (e.g. dots, simple closed curves or polygons, and sets of closed curves or polygons). It then builds

[1] The problems for Evans's system are generally taken from typical IQ tests and college entrance exams such as the GMAT.

its own high-level representation describing the figures of A, B, C, and all the given answer alternatives, along with their properties and relationships.[1] The program represents the relationship between A and B as a set of possible *transformation rules* that describe how figure A is transformed into figure B (and also counts in how many steps such a transformation is obtained). Then, each rule is applied to C in order to get one of the alternative answers (or indicate no such answer could be found). During this "comparison process", the program finds numbers that somehow reflect the mapping complexity and uses them in comparing the answer alternatives.[2]

The ANALOGY system and its implementation are introduced in Evans [1964], yet a more detailed presentation appears in [Minsky, 1968, Chapter 5]. The system is symbol-based: this is obvious in the ways in which (i) the problem inputs are represented and coded, (ii) the storage is explicitly manipulated, and (iii) the rules are created and handled. The system employs the sub-processes of 're-representation building' and a form of 'mapping'. The comparison of figures is specific to the working of the ANALOGY program, and cannot be thus seen as an 'evaluation' of analogy-making (in the sense of section 2.2.1), because it has nothing to do with the applicability of the mapping to the target (cf. Table 2.1).

Evans' program is usually (or even indisputably) considered as "the most famous early attempt to model analogy-making on a computer" [Hofstadter and the Fluid Analogies Research Group, 1996, pp. 269]. However, at about the same time in the 1960's[3], Reitman et al. presented a less popular, rather more general, more cognitively inspired, information-processing model of thinking, called "Argus" (cf. Reitman et al. [1964]). Argus solves proportional analogies too, but they seem rather trivial by today's standards (cf. [French, 2002, pp. 201]).

The Argus model is based on the sequential processing organization, used in the General Problem Solver (GPS) of Newell and Simon [1963]. It aims to represent complex cognitive structures by semantic networks, and claimed to be applicable to a broad range of cognitive functions. But, as Reitman et al. stated, the running version of the Argus model was limited at the time to only solving proportional analogies of the form "$(A : B) :: (C : (W, X, Y,$ or $Z))$" (cf. [Reitman et al., 1964, pp. 274]). Argus' architec-

[1] In a sense, this can be considered as a "representation-building" sub-process of analogy-making (cf. section 2.2.1).

[2] According to the ANALOGY implementation in Evans [1964], but in my own terms, if $n(O_1, O_2)$ is the number of modification steps (e.g. addition or removal of elements) needed to transform object O_1 into another O_2, then the equality $n(A, B) = n(C, D_i)$ is an indication that "'A' is to 'B' as 'C' is to 'D_i'", for an answer alternative D_i. The system does not, however, process any semantics of the considered properties and relations (e.g. no interpretation is given of the geometrical meaning of a figure element being 'left' to, or 'inside', another figure element).

[3] Or even earlier; cf. [Hofstadter and the Fluid Analogies Research Group, 1996, pp. 270].

ture included "far-sighted principles", such as "the interactions between the concept network and the problem to be solved" [French, 2002, pp. 201]. Nevertheless, a significant aspect of the ANALOGY program —beside its recognition as the earliest symbol-based model of analogy-making— is its ability to automatically build representations of the source and target domains on its own. Kokinov and French indicate that this feature has unfortunately been dropped in most recent models (but the success of this kind of analogy-making programs crucially depends on the procedures that build the descriptions). As also pointed out by Hall in [Hall, 1989, pp. 43], Evans' system focuses on elaborating an analogical mapping between source and target descriptions to solve a restricted type of proportional analogies.[1]

The SME Family of Systems: Based on Gentner's SMT (cf. section 2.1.3), Falkenhainer et al. presented the "*structure-mapping*[2] *engine*" (SME) to simulate the theory and provide a "tool-kit" for constructing matching algorithms. Falkenhainer et al.'s purpose is not to give a single matcher, but a simulator for a class of matchers (cf. [Falkenhainer et al., 1989, pp. 8]). That is, SME states assumptions underlying the computational implementation of, not only one system, but rather a family of models consistent with SMT. The assumptions underlying the SME include the following ones:

1. the mapping mechanism is isolated from other sub-processes (such as representation, retrieval, and evaluation),
2. matches based on relations are preferred over those based only on mere properties,
3. relations must be (syntactically[3]) identical in both domains in order to be put into correspondence, and
4. the principle of systematicity (cf. section 2.1.3) is followed, so that systems of relations are favored over isolated relations.

According to Falkenhainer et al. [1989], SME uses typed higher-order predicate calculus[4] to represent knowledge facts, by means of constructs that Falkenhainer et al. call:

[1] Models of the 1960's other than Evans', such as Becker's JCM and Reitman et al.'s Argus, seem to be more cognitively inspired. They embed analogical comparison in more general problem solving frameworks, which also address some of the composing sub-processes mentioned earlier. But developments of these models have not been continued.

[2] Also, the "*structural mapping*" engine.

[3] Two relations are seen as analogous if they exactly match in terms of their number of arguments, as well as in terms of the types of these arguments.

[4] Since actual computations are known to always work on a propositional level, SME's "predicate calculus" can in fact be seen as a propositional-like version of predicate logic, or a propositional logic

"*entities*" (logical individuals), "*predicates*" (functions, attributes, and relations), and "*description groups*" (collections of primitive entities and facts about them). SME constructs all "*structurally consistent mappings*" between two given descriptions of source and target domains. These mappings consist of pairwise matches between statements and entities in the source and target, plus the set of analogical inferences sanctioned by the mapping. SME also provides a "*structural evaluation score*" of each mapping according to the constraints of systematicity and structural consistency (cf. [Falkenhainer et al., 1989, pp. 8]).

Based on the SME assumptions, Forbus et al. developed the MAC/FAC as a model of similarity-based retrieval coupled with the SME and later used as SME's front-end (cf. [Forbus et al., 1995; Gentner and Forbus, 1991]). In MAC/FAC, "*episodes*" are what encapsulate representations of past events, with an episode having a double character of encoding in the long-term memory (LTM). An episode gives both:

1. a detailed "*predicate-calculus representation*" (of all the properties and relations of the entities within this episode), and

2. a shorter vector representation (that summarizes the relative frequencies of predicates used in the detailed representation).

MAC/FAC is a two-stage analogical retrieval engine, of which the first stage sweeps through the LTM, retrieving potential source episodes that match the target (based on superficial search, and using the short vector representations of episodes); whereas the second stage selects the best episode that matches the target (using the detailed predicate-calculus representations; cf. French [2002]; Kokinov and French [2003]). Gentner's main focus in the SMT is the explanation of an analogy-making after a target has been retrieved from memory. Accordingly, the "structure-mapping" phase in the front-end of the SME (i.e. MAC/FAC) builds a mapping between the retrieved source and target based on their structures and their overall coherence.

The SME family of models are symbol-based systems, with systems in such a family employing at least the "mapping" sub-processes (MAC/FAC clearly performs "retrieval" as well; cf. Table 2.1). A detailed outline of the SME and an extended discussion of the MAC/FAC model are given by Forbus in Forbus [2001].

that has relations. SME's "predicate calculus" should not be seen as an alternative of, for example, first-order logic or whatsoever, because this "predicate calculus" of SME neither makes use of variables nor quantifiers that are crucial in this regard (though it clearly employs relations). (Cf. section 3.1.3 for a related discussion about SME's expressivity compared to HDTP's.)

Model:	Processes & Key Characteristics:
ANALOGY	Geometric proportional analogy-making – Employs "representation-building" and "mapping" – Hand-coded descriptions – Automatically builds representations for geometric proportional analogies from the descriptions of input figures (cf. [Evans, 1964, 1968])
SME (MAC/FAC)	Analogical "retrieval" provides input to SME – Employs "mapping" (based on SME) and lately "re-representation" – Dual encoding in LTM – Predicate calculus representations (and vectors for representing frequencies) – First-stage vector match to filter candidates; SME used as stage 2 matcher (cf. [Falkenhainer et al., 1989; Forbus, 2001; Forbus et al., 1995])
HDTP	General-purpose analogy making – Many-sorted, first-order KR – Employs "mapping" , "transfer", and "re-representation" – Uses anti-unification to construct "generalizations" (cf. [Schwering et al., 2009a] and section 3.1)

Table 2.1: Three computational models of analogy-making and their key characteristics.

Other Models: The literature of analogy-making models provides tens of such models. Hall gives extensive descriptions of many other (relatively early) symbolic models of analogy-making (cf. [Hall, 1989]). In addition, many SME-based models have been developed within the past forty years or so, modifying parts of the base SME model just mentioned (e.g. MAGI, IAM, I-SME, SEQL, CARL, etc. [cf. French, 2002, pp. 202–203]). Table 2.1 lists names of three selective, symbol-based analogy-making models: ANALOGY, SME (MAC/FAC), and HDTP. The former two are quickly presented in this chapter, whereas the latter, HDTP, is the main model for analogy-making on which later discussions about aspects of intelligence are based. Details about the HDTP framework are presented and explained in the next chapter, along with some concrete analogy examples.

3

A Logical Framework for Modeling Analogical Reasoning

Based on ideas and results from cognitive science, the SMT's principle of systematicity provides an essence for comparing (potentially analogous) situations in SME-based families of symbolic models for computing analogies. Even for a model that is not entirely symbol-based, such as the hybrid model Copycat (cf. [Hofstadter, 1984]), it agrees with SME-based models at least on such a crucial principle (cf. [Hofstadter and the Fluid Analogies Research Group, 1996, Chapter 6, pp. 275–299]). Therefore, it is obvious that what qualifies a mapping as being successful in capturing an analogy between source and target domains is not the mere existence of an alignment between (low-level) facts in the source and the target domains, but rather the power to preserve the coherent way in which the facts interact with each other. That is, an alignment that generalizes the roles played by individual entities and their relationships in their respective domains. To capture this view, a computational model for analogy-making is presented in this chapter, which also provides the basic framework for all next chapters. The presentation, therefore, carries more detailed and essential descriptions of the model.

3.1 Heuristic-Driven Theory Projection (HDTP)

Heuristic-Driven Theory Projection (HDTP) is a symbolic analogy-making model, which is based on first-order logic and reasoning techniques (cf. Schwering et al. [2009a]), and has a front-end implementation in Prolog (cf. Schmidt [2010]). The basis of all operations and processes in the formal framework of HDTP is the formalization of the source and the target domains as sets of many-sorted, fist-order formulae.

HDTP is a framework for computing analogical relations between two (conceptual) domains, where the two domains are represented using finite many-sorted first-order logic *axiomatizations*. This way of representation using first-order logic axiomatizations allows the framework to incorporate reasoning mechanisms of AI. It thus enables us to view each of the input domains as a formal theory defined by basic conceptual entities given as *axioms*. Axiomatizations constitute formulae that conceptualize a domain with a finite number of "*facts*" and "*rules*" in an expressive logic-based KR formalization (cf. section 3.1.2). Moreover, HDTP still retains the goal of only drawing inferences that are cognitively inspired.[1]

The task of HDTP is to compare the formulae of the source and target domain theories to analyze them and find common patterns (i.e. structural commonalities). These common patterns suggest alignments via finding a generalized axiom system that, in turn, describes a generalized theory of both input theories. Thus, based on alignments that result from finding commonalities, HDTP establishes an analogical relation via a *generalization* of the domains. Entities from the source can later be mapped and translated to enrich the target domain and help in drawing newer inferences. But in order to achieve a sought *generalization* of analogous pairs (of axioms or formulae) from the input domains, a crucial idea of HDTP is to employ the formalism of *anti-unification*.

3.1.1 First- and (Restricted) Higher-Order Anti-Unification

The anti-unification technique is one possible model of generalization that "involves finding the least general unifier of two expressions" [Gentner and Forbus, 2011, pp. 267]. A unifier of two expressions is a statement with variables, which will be identical to the two expressions when appropriate substitutions of values for the variables are used. Anti-unification is mathematically sound, and is particularly fundamental for the presentation of the HDTP framework, because it produces simple schemas of the common structures within given source and target domains (cf. Plotkin [1970, 1971]). Since the very first time it has outlined by Reynolds and introduced by Plotkin (cf. [Plotkin, 1970, 1971; Reynolds, 1969]), anti-unification has mostly been used in the context of inductive learning and proof generalization. The basic goal of the (first-order) anti-unification formalism was to generalize pairs of *terms* in a simple, yet meaningful way by producing for each term an *anti-instance*. Within each anti-instance, distinct sub-terms are replaced by variables, which themselves can restore the original terms by replacing the new variables by appropriate sub-terms. Three examples of Plotkin's first-

[1] There is an extended, yet related, discussion in section 4.2.2 regarding the cognitive plausibility of using this way of representation of domains.

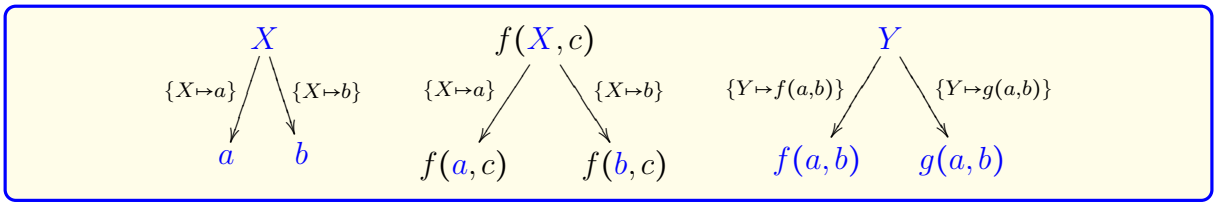


Figure 3.1: Examples of Plotkin's first-order anti-unification (cf. [Plotkin, 1970, 1971]), with terms in braces indicating substitutions of variables (X and Y) by instances (a or b, and $f(a,b)$ or $g(a,b)$, respectively).

order anti-unification are given in Figure 3.1 to demonstrate how generalizations can possibly be seen to induce an analogical relation.

Anti-unification is typically seen as a dual counterpart of unification: while in the latter the most general unifier (mgu) needs to be computed (by making terms equal via appropriate variable assignments), in anti-unification *more general* terms are constructed for given pairs of terms. Anti-unification aims to find a *most specific anti-unifier* (referred to as a least general generalization) that is minimal with respect to the instantiation ordering. From such a generalization of terms, an analogical mapping relation can be constructed by associating terms with a common generalization (cf. [Schwering et al., 2009a, pp. 6]).

Anti-unification has been shown by Plotkin to always succeed in finding a unique[1] "*least general generalization*" (lgg) for an involved pair of terms. But mere first-order structures of anti-unification do not have enough expressivity to capture the systematicity, required when domains are represented by first-order formulae. On the one hand, this motivates the use of higher-order anti-unifications instead of applying first-order anti-unifications. But, on the other hand, an unrestricted application of higher-order anti-unification opens the door for generalizations to become arbitrarily complex and may no longer reflect structural commonalities of the original terms: the structural commonalities are ignored in such case (cf. Schwering et al. [2009a]). Furthermore, and in fact even more crucially when using (unrestricted) higher-order anti-unification, there can be infinitely many anti-instances, and no longer is there the most specific anti-instance. This dilemma is approached in HDTP by extending Plotkin's classical first-order anti-unification to a restricted form of *higher-order anti-unification* (cf. Krumnack et al. [2007]).

[1] Up to renaming of variables, of course.

3.1.2 HDTP's Language: Conventions and Terminologies

Detailed descriptions of the language, which encodes knowledge of domains in HDTP, is briefly presented in the following. All notations and assumptions used in the presentation are mainly based on, and understood in the sense of, the settings defined in Gust et al. [2006]; Krumnack et al. [2007]; Schwering et al. [2009a]. In particular:

1. A many-sorted signature Σ provides the vocabulary strings that are used as building blocks for domain formalization. It describes a domain by comprising entities such as predicates and functions. (Constants are functions of arity 0.) The signature is identified by:

   (a) a set $Sort_\Sigma$ that forms a partially ordered set of sorts, which can be interpreted as high-level concepts of an ontology (e.g. $object$, $integer$, $real$, $time$, $massterm$, $bool$),

   (b) a set $Func_\Sigma$ that provides "function symbols" of the form

   $$f : s_1 \times \ldots \times s_n \to s, \text{ where } s_1, \ldots, s_n, s \in Sort_\Sigma,$$

   which are used to represent functions that map individuals to individuals, and

   (c) a set $Pred_\Sigma$ that provides "predicate symbols" of the form

   $$p : s_1 \times \ldots \times s_n, \text{ where } s_1, \ldots, s_n \in Sort_\Sigma,$$

   which are used to express relations between individuals.

2. Individual variables have sorts (i.e. variables are typed), and formulae can be well formed similar to how well-formed formulae are defined in classical first-order logic. However, in HDTP's jargon, (classical) first-order terms are extended by introducing *variables of arity* n, where $n \geq 0$, with $n = 0$ explicitly indicating classical first-order variables. In this setting, a term is either a first-order or a higher-order term, and variables can represent any possible term.

3. Logical operators (e.g. $\wedge$ and $\vee$) and quantifiers (e.g. $\forall$ and $\exists$) are used in constructing complex "facts" and "rules". A domain can thus be described by a finite set of formulae, which is called an "axiomatization". All formulae that can be inferred from an axiomatization of a domain constitute the domain theory. Different

equivalent axiomatizations for the same given domain are possible.[1]

4. A substitution σ on terms is a partial function that acts on the terms' variables, mapping the variables to terms. A substitution σ can formally be represented as $\sigma = \{x_1 \mapsto t_1, \ldots, x_n \mapsto t_n\}$, where sorts of each $\langle x_i, t_i \rangle$ pair match, $x_i \neq x_j$, and $i \neq j$ for all $i, j \in \{1, \ldots, n\}$. For two terms t and t', if t' is obtained from t by a substitution σ then the notation $t \xrightarrow{\sigma} t'$ is used to indicate that t' is an instance of t and t is an anti-instance of t'.

5. A generalization g of a pair of terms t and t' is denoted by $t \xleftarrow{\rho} g \xrightarrow{\sigma} t'$, where σ and ρ are substitutions. Nevertheless, since the classical notion of substitution is replaced by higher-order substitution, generalizations can also be defined for variables.

HDTP allows not only the anti-unification of terms but also of formulae, where the set of possible generalizations are extended in a controlled way by introducing a new notion of "*basic substitution*". The idea is to control resulting generalizations from becoming arbitrarily complex by restricting applicable substitutions in HDTP to only compositions of the following four basic substitution kinds (cf. [Krumnack et al., 2007]):

Renaming: A renaming is a substitution that replaces one variable[2] by another of the same argument structure. The notion ρ^X_Y denotes the replacement of a variable X by another Y of the same argument structure. That is: $X(t_1, \ldots, t_n) \xrightarrow{\rho^X_Y} Y(t_1, \ldots, t_n)$, where t_i is a term for $i \in \{1, \ldots, n\}$.

Fixation: A fixation is a substitution that replaces a variable by a function symbol of the same argument structure. The notion ϕ^X_f denotes the replacement of a variable X by a function symbol f of the same argument structure. That is: $X(t_1, \ldots, t_n) \xrightarrow{\phi^X_f} f(t_1, \ldots, t_n)$, where t_i is a term for $i \in \{1, \ldots, n\}$.

Argument Insertion: This kind of substitution replaces a subset of arguments of a variable by another variable. Specifically,

$$X(t_1, \ldots, t_n) \xrightarrow{\iota^{X,Y}_{Z,i}} Y(t_1, \ldots, t_i, Z(t_{i+1}, \ldots, t_{i+k}), t_{i+(k+1)}, \ldots, t_n)$$

denotes the replacement of k of the n arguments of a variable X by a variable Z (with k arguments) after the i^{th} argument of X, resulting in a variable Y with $m := n - (k - 1)$ arguments.

[1] Some formalizations are given in section 3.2.
[2] In the sense just described (cf. [Schwering et al., 2009a]).

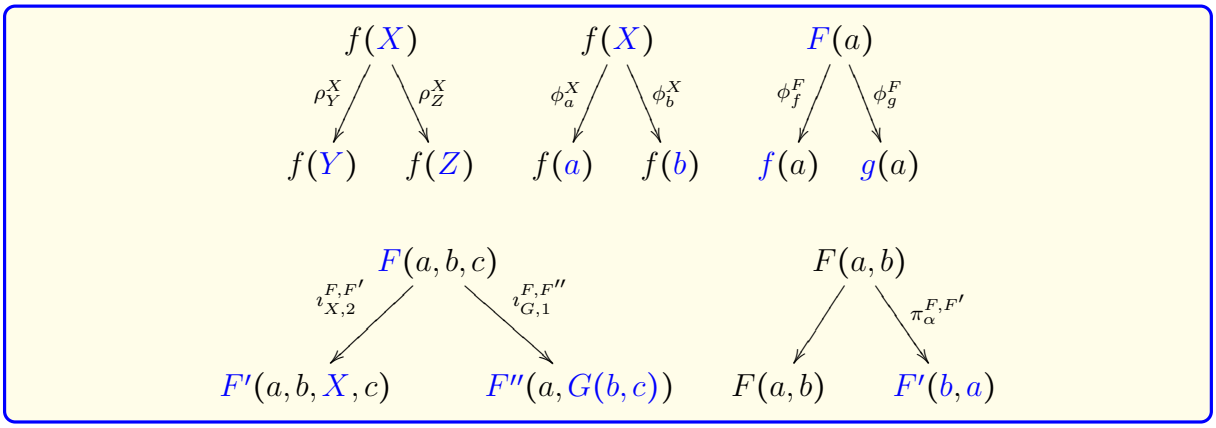


Figure 3.2: Examples of higher-order anti-unifications illustrating the four basic substitutions of the restricted higher-order anti-unification (cf. [Schwering et al., 2009a, pp. 260]).

Permutation: Permutation is a kind of substitution that re-arranges the arguments of a term. The arrangement uses a bijection $\alpha : \{1,\ldots,n\} \to \{1,\ldots,n\}$, and the permutation is denoted by: $X(t_1,\ldots,t_n) \xrightarrow{\pi^{X,Y}_{\alpha}} Y(t_{\alpha(1)},\ldots,t_{\alpha(n)})$.

Examples of restricted higher-order anti-unification substitutions are given in Figure 3.2.

It has been shown that the utilization of the given four forms of restricted higher-order substitutions provides the needed capability of detecting structural commonalities not accessible to first-order anti-unification (cf. [Krumnack et al., 2007]). Using restricted higher-order anti-unification, anti-unifiers always exist for a pair of terms (i.e. anti-unifiers are well-defined). Moreover, this utilization will never make a term less complex, which guarantees that only finitely many anti-instances (up to the renaming of variables) coexist for any given term (cf. [Schwering et al., 2009a, pp. 256]).

A drawback is, however, that the least general generalization is still no longer unique under restricted higher-order anti-unification. But having multiple possible generalizations is not necessarily counted as a bad criterion. It can be interpreted as delivering an advantage, particularly in the context of computing analogies, because humans too may simultaneously conceive more than one plausible mapping from the source to the target, with different *degrees of plausibility*. Figure 3.3 gives 3 higher-order anti-unifiers for the same pair of terms $f(g(a,b,c),d)$ and $f(d,h(a))$. All the three anti-unifiers are also least general generalizations (i.e. the 3 are most specific). No substitutions are shown in the figure, since there can be an infinite number of valid, basic substitution chains (e.g. by using renaming or permutation substitutions indefinitely).

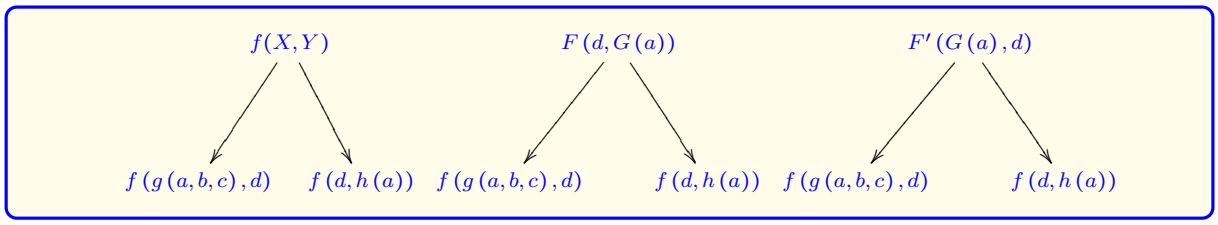


Figure 3.3: Higher-order anti-unification examples with several least general generalizations. All of $f(X,Y)$, $F(d,G(a))$, and $F'(G(a),d)$ are most specific anti-unifiers for the terms $f(g(a,b,c),d)$ and $f(d,h(a))$.

3.1.3 HDTP's Framework: Characteristics and Aspects

As it will become more evident in later chapters (e.g., cf. Chapters 7 and 8), the core anti-unification procedure can be further used for more than implementing analogical reasoning. HDTP allows for a more flexible merge of domains (not only generalizations), which can be used in merging domains for conceptual blending (cf. Chapter 4) as already explored in Guhe et al. [2011]; Martínez et al. [2012], for instance. Extensive, detailed presentations of HDTP's algorithms and implementations can be found in Schmidt [2010].[1] This section, however, informally presents certain core principles behind the computational implementation of HDTP's heuristics and algorithms. This serves to connect the previously discussed formal language and the later less-formal parts that mainly focus on modeling GI aspects (but uses HDTP as a grounding framework).

Heuristics, Ranking, and Coverage: The establishment of an analogy between two domains is carried out in HDTP via anti-unifying their sets of formulae (i.e. axiomatizations). Since this is not an anti-unification of only two formulae, it can be the case that one formula of the target domain is aligned with several of the source. Thus, HDTP gives ordering criteria to rank competing generalization alternatives, based on the complexity[2] of substitutions in the generalizations. HDTP selects a generalization with minimized complexity as a "*preferred generalization*". Broadly speaking, a preferred generalization is a least general generalization that has no less-complex least general generalization that play the same role (in anti-unifying the same pair of terms).

As the notion of "heuristics" in its name implies, HDTP uses a sequential heuristic-driven algorithm to compute preferred generalizations. Generic heuristics are imple-

[1] HDTP has a front-end implementation in Prolog, which is released as open source under the GNU General Public License. Also, HDTP's state of computational implementation, source codes, and stable publicly released versions can be downloaded from http://cogsci.uni-osnabrueck.de/ ai/analogies/.

[2] Complexity of substitutions can be defined as the processing effort, in terms of the needed steps, for the basic four substitutions given earlier (cf. [Definitions 7 and 8 Schwering et al., 2009a, pp. 256]).

mented to speed up the mapping, and thereby the construction of analogies. According to Schwering et al. [2009a], this is achieved by consecutively selecting formulae of the target domain, and searching for corresponding formulae in the source for anti-unification, in a manner that minimizes the complexity of alignments between the input domains. The generalized formulae form a generalized theory that links the source and target and represents their commonalities at an abstract level.

A resulting analogy is influenced by the ordering in which formulae are anti-unified. Substitutions that were required earlier to anti-unify two formulae might be applicable again to a later anti-unification of other formulae. Beside the plausible computational property to reuse substitutions without any cost of processing effort, a mapping that includes reused substitutions is more cognitively plausible than another that does not. The former is more likely to allow for achieving systematicity than the latter. Whence, the former is both computationally and cognitively preferred over the latter, since it leads to a *coherent mapping*.

HDTP implements a principle (by using heuristics) that maximizes the coverage of the involved domains (cf. [Schwering et al., 2009a]). Intuitively, this means that the sub-theory of the source (or the target) that can be generated by re-instantiating the generalization is maximized: the higher the coverage the better, because more support for the analogy is provided by the generalization. A further heuristics in HDTP is the minimization of substitution lengths in the analogical relation: the simpler the analogy the better (cf. [Gust et al., 2006]). There is a trade-off between high coverage and simplicity of substitutions: an appropriate analogy should intuitively be as simple as possible, but also as general and broad as necessary, in order to be non-trivial. This kind of trade-off is similar to the trade-off that is usually the topic of model selection in machine learning and statistics.

Two-Phase Analogy-Making System: HDTP finds an analogical relation by specifying a generalized theory of two domain theories.[1] Based on this generalization (and the involved substitutions), formulae in the source domain that have no correspondences in the target domain can be found, which enables the sub-process of analogical transfer. In fact, HDTP proceeds in two phases:

Mapping: In the *mapping phase*, the source and target domains are compared to find structural commonalities, and a generalized description is created that subsumes the matching parts of both domains. The mapping phase constructs a set of gen-

[1] Domain theories will also be referred to as "conceptual domains" or "concepts", starting Chapter 4 onwards.

eralized axiomatizations that anti-unifies two input sets of axiomatizations, each with a finite number of rules and facts.

Transfer: In the *transfer phase*, unmatched knowledge in the source domain is mapped to the target domain to establish new hypotheses.

Figure 3.4 depicts HDTP's overall approach to creating analogies, in which analogical transfer results in structure enrichment of the target side. This usually corresponds to the addition of new axioms to the target theory (but may also involve the addition of new first-order symbols).

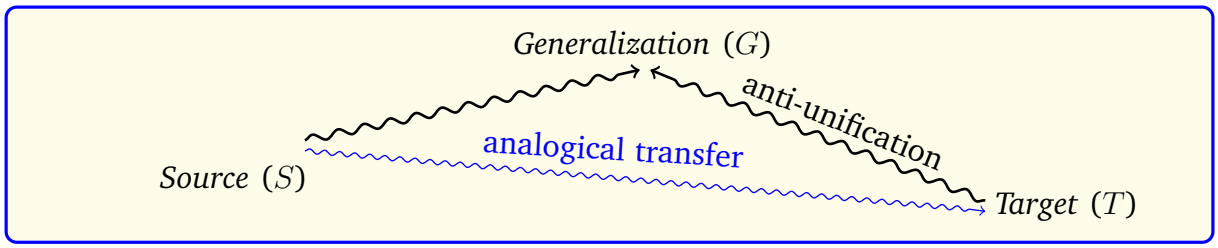


Figure 3.4: HDTP's overall approach to creating analogies. HDTP applies anti-unification to find analogical relations between the source and target domains, then proposes analogical transfer based on generalizations.

HDTP and SME: HDTP is similar in spirit to SME with respect to the mentioned mapping and transfer phases, and the utilization of symbolic formalisms for the representation of domains (though the formalisms and their expressivity differ). Also, preferring reused substitutions is more likely to allow for achieving the cognitively plausible principle of systematicity, principally aimed at by the SME-based family of models. Nevertheless, HDTP also differs significantly from SME at least with respect to:

1. the stronger expressive power that HDTP provides by the underlying domain theories (many-sorted first-order logic in HDTP vs. propositional-like logic in SME) enables HDTP not only to represent situations but general laws as well,

2. the requirements for labeling in HDTP are more flexible (SME requires labels to be identical for the alignment of attributes and relations, whereas HDTP is flexible in this regard),

3. the establishment of the analogy relation in HDTP is always available as a by-product of an explicit abstraction, which can be useful in many ways,

4. the massive usage of heuristics that differ from the ones used in SME,[1] and

5. the possibility for HDTP to account for semantic information (also cf. [Krumnack et al., 2013a, 2010]).

Re-Representation: As briefly mentioned in section 2.2.1, the establishment of an analogy may essentially need to change the representation of the potentially analogous domains to allow making implicit analogous structures explicit (cf. Clement [2008]; Indurkhya [1992]; Schwering et al. [2009a]). Indurkhya argues that the process of "re-description" or changing representation is realized to underly creative metaphors and analogies (cf. [Indurkhya, 1992, pp. 409]), but the sub-process of re-representation is, unfortunately, maltreated in most models of analogy-making. The HDTP framework entails a mechanism for re-representation in a quite natural way, since a logical representation of a domain does not only provide the explicitly given axioms, but also makes available their logically deducible formulae. HDTP's idea is to incorporate the derived formulae into the mapping process whenever the original axiomatizations do not lead to a satisfying analogical relation (cf. the first example in section 3.2.1 and [Schwering et al., 2009a, pp. 258]).

3.2 Application in Analogy Domain Examples

This section discusses three scenarios of analogy-making to show how HDTP can represent and find mappings between the involved domains (and transfer knowledge between them). Full-fledged axiomatizations that formalize representations for HDTP are already presented and extensively discussed in several places (cf. Guhe et al. [2010]; Krumnack et al. [2013a, 2007]; Martínez et al. [2012] and, in particular, [Figure 2, Figure 7, and the Appendix of Schwering et al., 2009a]). These representations do not only show the power of the HDTP framework in finding analogical mappings and executing analogical transfer by means of generalizing the axiomatizations of given conceptual domains, but they also explain HDTP's potential capability of re-representing such conceptual domains when an analogy cannot directly be found from the originally given axiomatizations. In the following, yet a new analogy-making scenario is thoroughly

[1] Note that SME-based models also use heuristics. The SME framework is not heuristics-free, though this is not prominently focused on or mentioned explicitly (e.g. local and global matches do not just use plain searching algorithms, but employ some heuristics as well; cf. Falkenhainer et al. [1989]; Forbus et al. [1995] and section 2.3). However, the usage of many heuristics in HDTP is more fundamental and prominent.

presented after concrete, particularly interesting ideas are discussed about two of the analogy-making situations given in Schwering et al. [2009a]. In typesetting the axiomatizations, certain conventions are followed: (i) domain entities (such as predicates and functions) are shown in `typewriter` font, (ii) `CONSTANT` entities are capitalized, and (iii) *sorts* are italicized.

3.2.1 Two Classical Analogy Situations

The Rutherford Analogy: A well-known situation is the analogy between the solar system and the Rutherford atom model, typically stated as: "the atom is like our solar system" [Gentner, 1983, pp. 159]. The source domain simplifies a description of the solar system, where a planet revolves around the sun because the differences in mass result in different gravitational forces. (They do not, however, collide with each other.) The target domain describes the Rutherford atom model, where the Coulomb force results in lightweight electrons being attracted by the nucleus. The electrons and the nucleus keep a distance greater than zero, which is an abstraction of the results form Rutherford's gold foil experiment (cf. Rutherford [1911]). Formalizations of this scenario are given in Table 3.1. The analogy situation illustrates several aspects of HDTP by comparing (simplified and partial) knowledge representations about the solar system, on the one hand, with representations (from a different field of knowledge) about the Rutherford atom model, on the other hand.

A generalized theory Th_G of the two input domain theories (that are given in Table 3.1) is shown in Table 3.2. Entities that originally have similar string symbols keep their names unchanged when anti-unified in the generalization, such as the `revolves_around` predicate. Whilst, matched entities with different strings in the input domains are anti-unified in the generalization using new symbols. In Table 3.2 for example, the function symbol F is an anti-instance of both `centrifugal` from the source and `coulomb` from the target. Thus, the fact $\forall(t):F(o_1,o_2,t)>0$ in Th_G generalizes the idea of having a continuous nonzero interrelationship (i.e. attraction force) between two objects, where $\langle o_1,o_2\rangle \in \{\langle \texttt{SUN},\texttt{PLANET}\rangle, \langle \texttt{NUCLEUS},\texttt{ELECTRON}\rangle\}$.

The formalizations in Table 3.1 were chosen in a way that do facilitate the direct discovery of an analogy by matching the axioms of the two domains. This treatment is similar to the treatment of the same specific analogy situation by structure-mapping in Gentner [1983], for example. Unlike Gentner's treatment, however, if formalizations of the solar system and the Rutherford atom model are given in a different way, such as those in Table 3.3, the discovery of the analogy becomes indirect. For example, there is no di-

Source: SOLAR SYSTEM	**Target:** RUTHERFORD ATOM MODEL
sorts $real$, $object$, $time$	**sorts** $real$, $object$, $time$
entities SUN, PLANET: $object$	**entities** NUCLEUS, ELECTRON: $object$
functions mass: $object \to real \times \{kg\}$ dist: $object \times object \times time \to real \times \{m\}$ gravity, centrifugal: $object \times object \times time \to real \times \{N\}$	**functions** mass: $object \to real \times \{kg\}$ dist: $object \times object \times time \to real \times \{m\}$ coulomb: $object \times object \times time \to real \times \{N\}$
predicates revolves_around: $object \times object$	**predicates** revolves_around: $object \times object$
facts mass(SUN)>mass(PLANET) mass(PLANET)> 0 $\forall(t{:}time)$:gravity(SUN,PLANET,t)> 0 $\forall(t{:}time)$:dist(SUN,PLANET,t)> 0	**facts** mass(NUCLEUS)>mass(ELECTRON) mass(ELECTRON)> 0 $\forall(t{:}time)$:coulomb(NUCLEUS,ELECTRON,t)> 0 $\forall(t{:}time)$:dist(NUCLEUS,ELECTRON,t)> 0
laws $\forall(t{:}time)(o_1, o_2{:}object)$:dist$(o_1, o_2, t) > 0 \wedge$ gravity$(o_1, o_2, t) > 0 \to$ centrifugal$(o_1, o_2, t) =$ -gravity(o_1, o_2, t) $\forall(t{:}time)(o_1, o_2{:}object)$:$0 <$ mass$(o_1) <$ mass$(o_2) \wedge$ dist$(o_1, o_2, t) > 0 \wedge$ centrifugal$(o_1, o_2, t) < 0 \to$ revolves_around(o_1, o_2)	**laws**

Table 3.1: Axiomatizations of the solar system domain (base) and the Rutherford atom model domain (target). (Reproduced from [Schwering et al., 2009a, Figure 2, pp. 254].)

Generalization:
Th_G

sorts
$real, object, time$

entities
$X, Y : object$

functions
$\texttt{mass}$: $object \to real \times \{kg\}$
$\texttt{dist}$: $object \times object \times time \to real \times \{m\}$
$F : object \times object \times time \to real \times \{N\}$
$\texttt{centrifugal}$: $object \times object \times time \to real \times \{N\}$

predicates
$\texttt{revolves\_around}$: $object \times object$

facts
$\texttt{mass}(X) > \texttt{mass}(Y)$
$\texttt{mass}(Y) > 0$
$\forall(t{:}time){:}F(X,Y,t) > 0$
$\forall(t{:}time){:}\texttt{dist}(X,Y,t) > 0$

laws
$\alpha{:}\forall(t{:}time)(o_1,o_2{:}object){:}\texttt{dist}(o_1,o_2,t) > 0 \wedge F(o_1,o_2,t) > 0 \to \texttt{centrifugal}(o_1,o_2,t) = -F(o_1,o_2,t)$
$\beta{:}\forall(t{:}time)(o_1,o_2{:}object){:}0 < \texttt{mass}(o_1) < \texttt{mass}(o_2) \wedge \texttt{dist}(o_1,o_2,t) > 0 \wedge \texttt{centrifugal}(o_1,o_2,t) < 0 \to \texttt{revolves\_around}(o_1,o_2)$

Table 3.2: A generalized theory Th_G of the solar system and the Rutherford atom model domains (cf. Table 3.1). α and β are generalizations from the transfer. (Reproduced from [Schwering et al., 2009a, Figure 6, pp. 257].)

rect way to achieve a formula that generalizes the idea of having a continuous nonzero attraction force between objects (i.e. the role $\forall(t){:}F(o_1,o_2,t) > 0$ plays within Th_G). But using Table 3.3's axioms, one can logically deduce $\forall(t){:}\texttt{gravity(SUN,PLANET,}t) > 0$ on the source side, and $\forall(t){:}\texttt{coulomb(NUCLEUS,ELECTRON,}t) > 0$ on the target side, which can be anti-unified. HDTP's entailed mechanism for re-representation allows to generalize not only explicitly given axioms, but their logically deducible formulae too. Thus, by allowing logical deduction before anti-unification, the matching can still be achievable.[1]

The Heat-Flow/Water-Flow Domains: The *"heat/water"*-flow is another classic scenario, in which "two connected vessels filled with different quantities of water are analogically related to two massive bodies of different temperature that are connected via some metal bar". This famous analogy situation was originally given in [Falkenhainer et al., 1989, Figure 1, pp. 3] and is being graphically depicted in Figure 3.5.

According to axiomatizations given in [Appendix of Schwering et al., 2009a, pp. 264–

[1]But note that this is not always possible, because of the nature of the first-order logic used for the representation. Cases exist in which no useful analogy can be computed if one only considers the given domain axiomatizations (cf. [Schwering et al., 2009a, pp. 259]).

Source: SOLAR SYSTEM	**Target:** RUTHERFORD ATOM MODEL
mass(SUN)>mass(PLANET)	mass(NUCLEUS)>mass(ELECTRON)
mass(PLANET)>0	charge(ELECTRON)<0
$\forall(t)$:dist(SUN,PLANET,t)>0	$\forall(t)$:dist(ELECTRON,NUCLEUS,t)>0
$\forall(x)\forall(y)\forall(t)$:mass($x$)$>0\wedge$mass($y$)$>0\rightarrow$ gravity(x,y,t)>0	charge(NUCLEUS)>0
$\forall(x)\forall(y)\forall(t)$:gravity($x,y,t$)$>0\rightarrow$ attracts(x,y,t)>0	$\forall(x)\forall(y)\forall(t)$:charge($x$)$>0\wedge$charge($y$)$<0\rightarrow$ coulomb(x,y,t)>0
$\forall(x)\forall(y)\forall(t)$:attracts($x,y,t$)$\wedge$ dist(x,y,t)>0 $\wedge$mass(x)>mass(y) $\rightarrow$ revolves_around(y,x)	$\forall(x)\forall(y)\forall(t)$:coulomb($x,y,t$)$>0\rightarrow$ attracts(x,y,t)
background knowledge $\forall(x)\forall(y)\forall(t)$:dist($x,y,t$)=dist($y,x,t$) $\forall(x)\forall(y)\forall(z)$: $x>y\wedge y>z\rightarrow x>z$	

Table 3.3: Another formalization of the Rutherford analogy situation. (Reproduced from [Schwering et al., 2009a, Figure 7, pp. 259].)

265], a direct comparison of the terms height(in(WATER,BEAKER),t_{start}) in the source and temp(in(COFFEE,CUP),t_{start}) in the target might lead to the assumption that the individual entity WATER should be mapped to COFFEE, and BEAKER to CUP. But this is a fundamental misunderstanding of the analogy the way it is given in Schwering et al. [2009a]. By considering the other parts of the axiomatizations, another preferred generalization can map height(in(WATER,VIAL),t_{start}) to temp(ICE-CUBE,t_{start}) which means that the height of the water currently being in the vial maps to the temperature of the ice cube. This clarifies the importance of the role that a preferred generalization plays, as well as the importance of the deeper commonalities captured by the "argument insertion" kind of basic substitutions (cf. section 3.1.2).

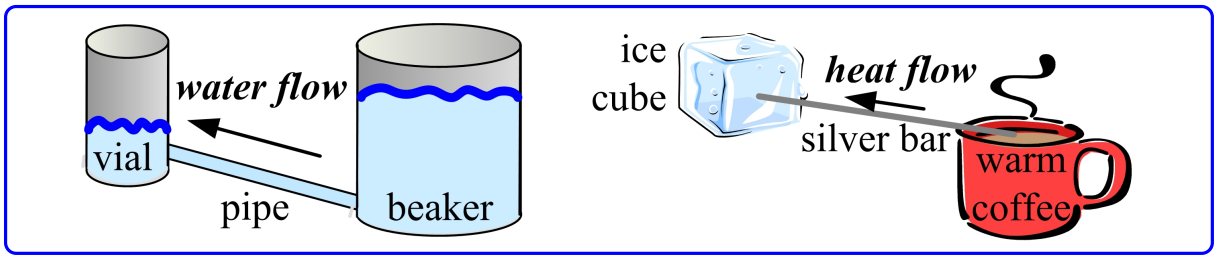


Figure 3.5: The *"heat/water"*-flow analogy situation. (Re-produced according to [Schwering et al., 2009a, Figure 8, pp. 260]; also cf. [Hofstadter and the Fluid Analogies Research Group, 1996, Figure 5.1, pp. 276].)

3.2.2 The Flower/Brain Metaphor

The Scenario: Looking at Figure 3.6 for the first time, humans would perceive it as being categorized into two parts: one to the left and another one to the right.[1] Based on their knowledge, available from previously encountered mundane situations, it is highly likely that the individual entities composing the left part form a more coherent domain (with more deeply related entities) to humans than those composing the right part: common knowledge tells us that a cloud is known to have some obvious relation with rain drops that, in turn, have some obvious relation with nourishing the flower by increasing the (intensified) water it contains. The two parts of Figure 3.6 can, thus, define two separate domains (i.e. the left and the right parts). By comparing these domains, an analogy is triggered to establish a mapping between the entities appearing in the two parts, and a metaphorical relation is brought to one's mind. It makes sense to assume the left part as the source domain in such scenario, since it contains richer background knowledge and multiple interrelations between the entities it comprises. The cloud, the rain drops, the plant, and the flower pot can be matched with the book, the ink drops, the brain, and the head, respectively. This also can be clearly seen by comparing individuals in the two domains and their interrelationships.

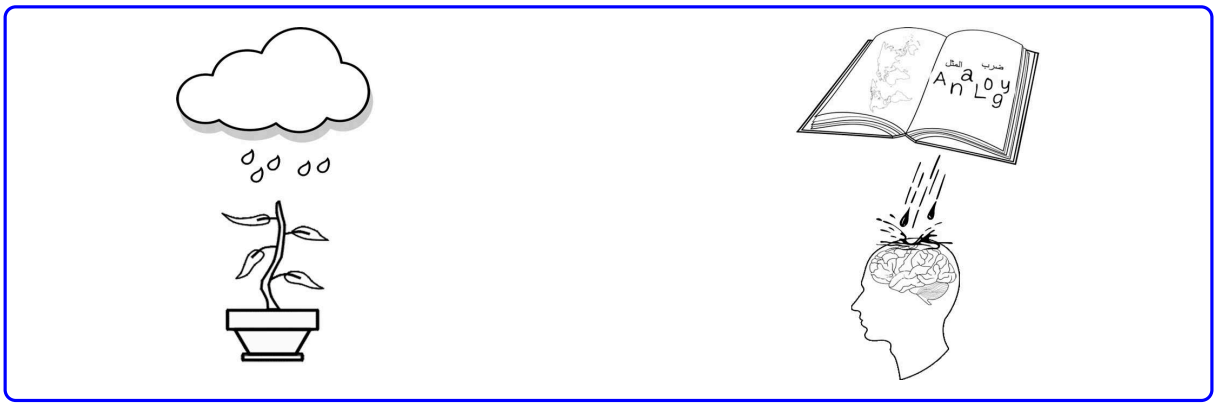


Figure 3.6: A matching of the components in the left part of the image (source) can be found by contrasting them with those in the right part (target). The governing relations among the components can also be mapped, enriching an understanding of what the target may conceptualize (irrespective of what it could have conceptualized if initially given alone).

A deeper understanding of the 'system of relations' that holds among the individual entities within the source domain would allow people to efficiently transfer knowledge

[1] This perceiving competency may alone need to utilize complex capacities, such as vision, categorization, spatial reasoning, as well as others.

from the source to the target, perhaps even in a metaphorical manner. They may even start to see the ink drops playing the role of information gathered by reading the book (though none of these views is explicitly mentioned). For example, one may infer an effect of (metaphorically) "growing one's knowledge" inside the "brain" by "reading" in the same manner that rain helps a flower to naturally "grow" by nourishing. One may further infer that "information" nourishes "knowledge" like "water" nourishes the "plants".[1] To humans, this definitely reflects at least one more important aspect of cognition and intelligence; namely finding a metaphorical relation (through analogy) and applying it (cf. Gust et al. [2006]): note that "information" in the previous sentence is the metaphorical use of the "ink" that is actually present in the figure. The following explanations show that cognitively inspired models of analogy-making are very helpful in endowing cognitive agents with the capacities needed to perform this task, which by far seems to be a human-directed task (cf. [Gust et al., 2006; Indurkhya, 1992]).

Axiomatizations, Mapping, and Transfer of Knowledge: Assume that the KB of a cognitive agent already contains required representations (in a many-sorted, first-order logic language). The HDTP framework can be used to find the aforementioned system of relations that connects the book, the ink drops, and the brain, by considering the base system of relations that connects the cloud, the rain drops, and the flower. The example is explained in terms of the suggested domain axiomatizations given in Table 3.4, beside using visual representations and illustrating graphs akin to semantic networks (cf. Figures 3.7 and 3.8) to simplify the explanations.

Figure 3.7 is an attempt to visually explain what the axiomatizations of Table 3.4 formally state. The source domain in this example is referred to as "the *flower* domain", and the target as "the *brain* domain". Knowledge that is initially available within the "brain" (target) domain would not (normally[2]) imply an obvious connection between the sub-domain represented by the "book" entity and the sub-domain containing the "brain" entity (each is separately encircled in Figure 3.7). The axiomatizations in Table 3.4 list entities, facts, laws, and background knowledge that represent the knowledge of the two domains. For example, in the flower domain axiomatization, the idea that the rain drops feed the flower is represented by `nourish(RAIN,FLOWER)`. Also, the given law rep-

[1] I showed Figure 3.6 (with both its left and right sub-images) to four personal contacts, then asked about what they see, and which sub-image looks more natural. In their responses, all of them used metaphorical terms to interpret "growing" (or "feeding") the flower, and indicated that "drops of information are falling from the book into the brain". They also reported that the right sub-image seems more artificial (yet more creative).

[2] When the source (flower) domain is completely absent, for instance.

Source: FLOWER	**Target:** BRAIN
sorts $object$, $massterm$, $real$, $bool$	**sorts** $object$, $massterm$, $real$, $bool$
entities CLOUD, FLOWER, PLANT, POT, VESSEL: $object$ WATERVAP, RAIN: $massterm$	**entities** BOOK, BRAIN, ORGAN, HEAD, SKULL: $object$ INK, INFO: $massterm$
functions solid: $object \to bool$ amount: $object \to real$ intensify: $massterm \to object$	**functions** solid: $object \to bool$ amount: $object \to real$ intensify: $massterm \to object$
predicates isa, supply, contain, protect, role: $object \times object$ synthesize, nourish: $massterm \times object$ ersatz: $massterm \times massterm$ incrLvl: $object \times object \times real$	**predicates** isa, supply, contain, protect: $object \times object$ synthesize: $massterm \times object$
facts isa(FLOWER,PLANT) solid(VESSEL)= TRUE contain(POT,FLOWER) synthesize(WATERVAP,CLOUD) supply(CLOUD,intensify(RAIN)) ersatz(WATERVAP,RAIN) role(POT,VESSEL) nourish(RAIN,FLOWER)	**facts** isa(BRAIN,ORGAN) solid(SKULL)= TRUE contain(HEAD,BRAIN) synthesize(INK,BOOK) supply(BOOK,intensify(INFO))
laws $\forall(o_1{:}massterm)\forall(o_2{:}object)$:nourish$(o_1,o_2) \to$ $\exists(o{:}object)$:synthesize$(o_1,o) \wedge$ incrLvl$(o,o_2$,amount(intensify$(o_1)))$	**laws**
background knowledge $dewPoint : real$ humid: $real \to bool$ $\forall(x,y{:}object)$:solid$(x) \wedge$ contain$(x,y) \to$ protect(x,y) $\forall(d{:}real)$:(humid$(d) \wedge d \leq dewPoint) \to$ supply(CLOUD,WATERVAP) $\forall(d{:}real)$:(humid$(d) \wedge d > dewPoint) \to$ supply(CLOUD,RAIN)	

Table 3.4: Suggested axiomatizations of the flower/brain analogy situation. (The used function and predicate names should be obvious from the given explanations. The predicate incrLvl is a shorthand of "increase level", so that incrLvl$(\square_1,\square_2,\square_3)$ reflects an increase in the level of $\square_1$ within $\square_2$ by an amount $\square_3$. Furthermore, $dewPoint$ corresponds to the temperature at which water vapor condenses into water.)

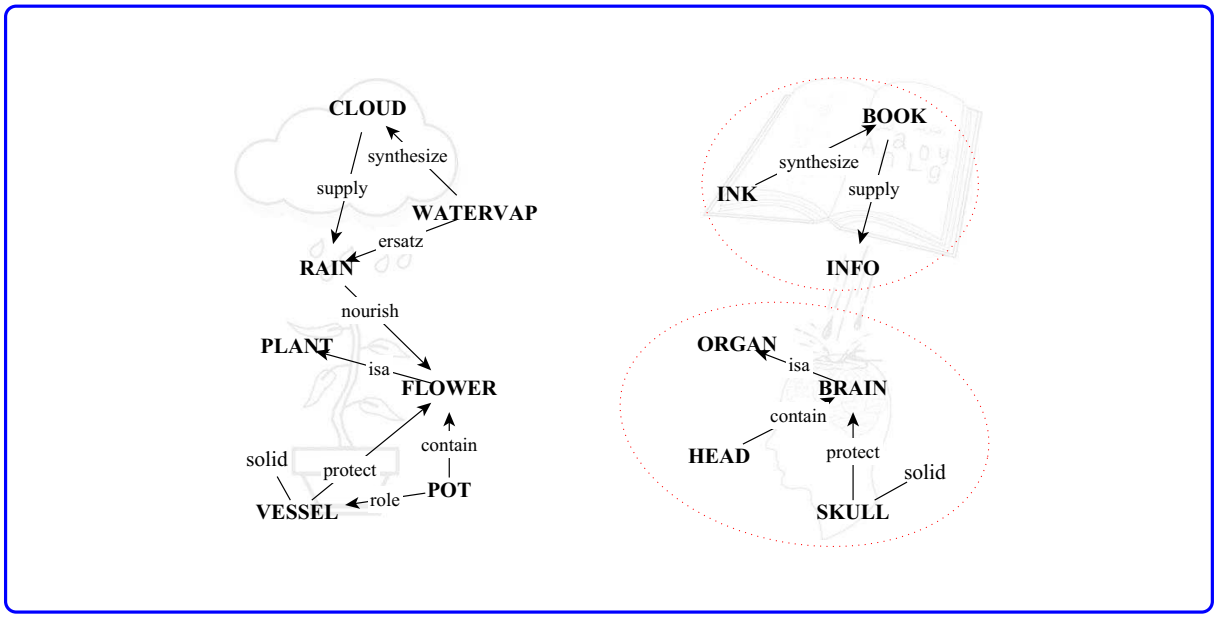


Figure 3.7: A possible (graphical) illustration for representing the source and target domains in the flower/brain analogy (cf. Table 3.4).

resents the idea that: "if an object o_2 nourishes another o_1, then an increase[1] happens to o's level in o_1, where o is an object that results from intensifying o_2. The increase of o's value (in o_1) equals the amount of the intensification caused by nourishing".

HDTP can generalize the "flower" and "brain" domains, and a generalization is shown in Table 3.5. Individual entities, and relations between them, within the "flower" domain are mapped to entities and relations that play similar roles within the "brain" domain. There can in principle be several possible ways to map the individuals and their relations. In this particular example, and depending on the part of knowledge being represented, a preferred generalization by HDTP would result in the mapping depicted in Figure 3.8 (also see the generalization's formalization in Table 3.5). Among other things, (i) the `CLOUD` under consideration is mapped to `BOOK`, (ii) `intensify(RAIN)` is mapped[2] to `intensify(INFO)`, and (iii) `FLOWER` is mapped to `BRAIN`.

An analogical transfer is triggered, which enriches the "brain" domain with new entities and relations, implying that the sub-domain represented by the "book" entity and the sub-domain containing the "brain" entity become now linked (they were initially not linked, in the sense of Figure 3.7). After the mapping is established, a copy of the

[1]The predicate `incrLvl(`$\square_1$`,`$\square_2$`,`$\square_3$`)` in Table 3.4 shorthands the "increase in the level of $\square_1$ within $\square_2$ by an amount $\square_3$".

[2]This mapping captures, somehow in a metaphorical sense, the underlying idea that "knowledge" (the intensified metaphorical version of `INFO`) feeds `BRAIN`, like "water" feeds the `FLOWER`. Note, however, that neither "water" nor "knowledge" are defined in the given representation.

Generalization:
Th_G

sorts
$object, massterm, real, bool$

entities
$E_1, E_4, E_5, E_6, E_7, X_1 : object$
$E_2, E_3 : massterm$

functions
`solid`: $object \rightarrow bool$
`amount`: $object \rightarrow real$
`intensify`: $massterm \rightarrow object$

predicates
`isa, supply, contain, protect`, $P_2 : object \times object$
`synthesize`, P_3: $massterm \times object$
P_1: $massterm \times massterm$
$P_4 : object \times object \times real$

facts
`isa`(E_4, E_5)
`solid`(E_7)`= TRUE`
`contain`(E_6, E_4)
`synthesize`(E_2, E_1)
`supply`(E_1, X_1)
$P_1(E_2, E_3)$
$P_2(E_6, E_7)$
$P_3(E_3, E_4)$

laws
$\forall(o_1{:}massterm)\forall(o_2{:}object){:}P_3(o_1, o_2) \rightarrow \exists(o{:}object){:}$`synthesize`$(o_1, o) \wedge P_4(o, o_2,$ `amount`$($`intensify`$(o_1)))$

Table 3.5: A generalized theory of the flower/brain analogy situation, on which the analogical mapping and transfer are based. (The anti-unifications can be easily obtained by consulting Table 3.4.)

fact: `nourish(RAIN, FLOWER)`, for instance, is transferred from the "flower" domain to enrich the "brain" domain with the fact: `nourish(INFO, BRAIN)`. Using Table 3.5, both facts can be obtained as instances by applying the following substitutions in the (generalized) fact $P_3(E_3, E_4)$ (which appears in Table 3.5's generalized theory Th_G):

1. substituting P_3 with `nourish` (in both domains),
2. substituting E_3 with `RAIN` and `INFO`, in the flower and the brain domains, respectively, and
3. substituting E_4 with `FLOWER` and `BRAIN`, in the flower and the brain domains, respectively.

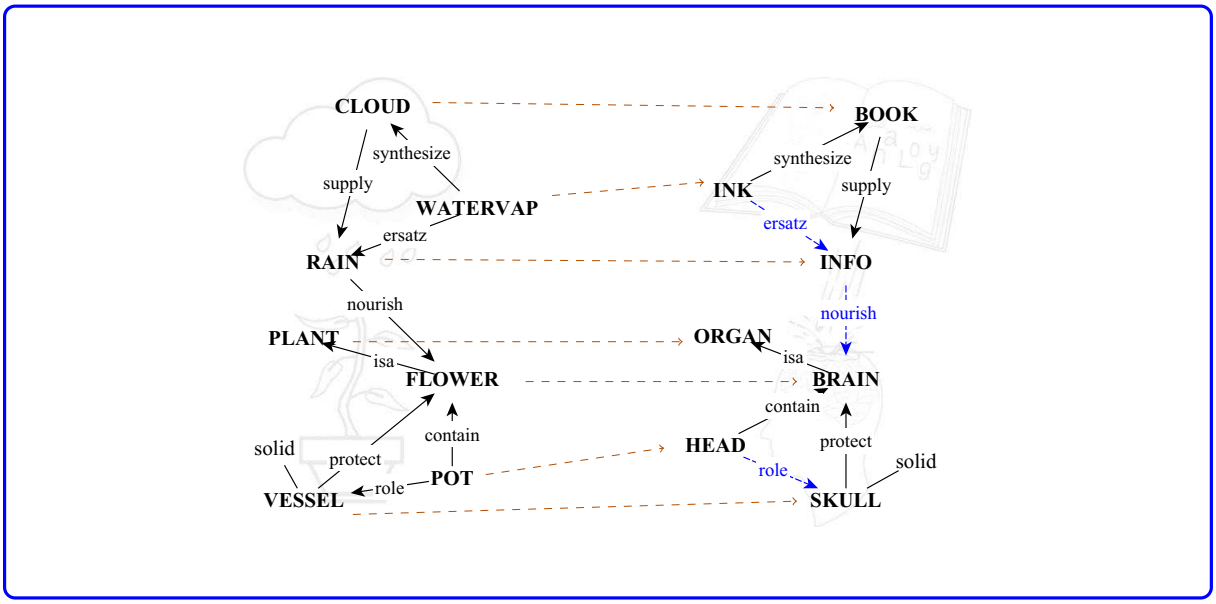


Figure 3.8: After establishing a mapping from the flower domain to the brain domain, the system of relations in the latter becomes richer.

4

Cross-Domain Reasoning via Conceptual Blending

A simple puzzle given by the famous mathematician (and magician), Smullyan, is the *Rate-Time Puzzle*, which is stated as follows (cf. [Smullyan, 1978, pp. 12]):

> "A train leaves from Boston to New York. An hour later, a train leaves from New York to Boston. The two trains are going at exactly the same speed. Which train will be nearer to Boston when they meet?"

Regardless of whether an intelligent, cognitive reasoner can find the correct answer[1] to this puzzle, it is highly likely that the reasoner will somehow be *stimulated* to combine clues from two imagined or conceptualized domains, firstly, then form a third coherent domain, before reaching such an answer.

Cases like this, which stimulate us to merge knowledge, happen quite often in many reasoning situations that involve thinking of conceptualized domains. After a cognitive reasoner is somehow stimulated to integrate knowledge pieces (such as the spatial, temporal, or directional knowledge about the trains and cities in the given puzzle) from two or more different conceptualizations, a conceptual mixture of cross-domain elements results from a process, widely known as "conceptual blending". In such situations, the pieces of knowledge are initially dispersed or belong to different conceptualizations, and their mixing or "blending" is achieved by establishing a coherently combined conceptualization that fuses salient entities of the former ones into the latter.

Conceptual blending (henceforth CB) provides a powerful mechanism for facilitating the creation of new conceptions by constrained integration of conceptual knowledge from diverse situations. Whence, it is also referred to as "conceptual integration"

[1] Obviously the two trains will be at the same distance from Boston when they "meet" (cf. [Smullyan, 1978, pp. 19]).

(cf. Fauconnier and Turner [1998, 2002]; Turner and Fauconnier [1998]). No commonly agreed upon definition of CB exists. Nevertheless, it can roughly be described as the creation of concepts by a principle driven combination of existing ones. "In general, there are many ways to combine the same concepts but in any case the result maintains parts of their initial structure. Furthermore, the blending of two concepts is not a simple union, but may give rise to emergent structure" (cf. [Martínez et al., 2011, pp. 210]).

Fauconnier and Turner introduced the early ideas of a framework for performing CB in the 1990's. CB was originally intended to explain specific cognitive phenomena, such as metaphor and metonymy [Fauconnier and Turner, 1998]. Then, it has developed to explain a variety of other cognitive phenomena that are ubiquitous to everyday thought. About a decade later, Fauconnier and Turner proposed the newer version of the CB framework of reasoning as a theory of cognition that explains "the nature and origin of cognitively modern human beings" (cf. Fauconnier and Turner [2002]). Fauconnier and Turner acknowledge Koestler's forerunner idea of "bisociation of matrices" (cf. [Koestler, 1964]) as an earlier "symptom" that is akin to CB and is "shared by all examples of remarkable creative invention" [Fauconnier and Turner, 2002, pp. 37]. Although Fauconnier and Turner's proposal of CB as a theory of cognition is a bold claim, and did stimulate some critics[1], CB is still continuing to develop and show its importance and applicability in a wide range of areas.

This chapter gives a foundational background that serves to introduce, motivate, and try to characterize facets of modeling the CB process. Critics will also be crisply mentioned, in particular because pivotal components of CB are still ill-defined for computationally oriented modeling. Nevertheless, throughout the rest of the thesis, it is intended to strongly advocate the CB process as a fundamental higher-level cognitive mechanism, by demonstrating and explaining concrete, uncontroversial key roles it plays in a variety of cognition situations that show clear signs of human intelligence.

4.1 General Assumptions and Basic Elements

This section introduces the basic components of, and lists general assumptions about, the CB framework, using a model-independent view. The presentations in this section are thus intended to be as broad as possible but, at the same time, as concise as possible, because they are used later to describe the functioning of the framework.

The first important assumption about CB that should be made clear is that I consider

[1] Some critics are briefly mentioned in [Pereira, 2007, §3.2.2, for example] and section 4.4.

CB an explanatory cognitive mechanism. Unlike Fauconnier and Turner's proposal of the CB framework as a theory of cognition, I view the framework as a basis for simulating aspects of higher-level cognition. The framework itself does not tell us to always blend our knowledge conceptualizations in order to reason, nor is it a method showing how one becomes creative. It has rather the power to explain how concrete types of reasoning and creativity could be achieved (consequently, simulated in computational models of reasoning or creativity; also cf. Abdel-Fattah and Schneider [2013]).

The next assumption concerns the idea that all approaches for blending assume that knowledge is organized in some form of "spaces", or "domains". Therefore, the underlying knowledge base of computational models for CB should be able to provide a way to represent these spaces as conceptual entities (cf. section 1.3) that can be grouped and taken as input knowledge structures to the blending process. Such a process will be described after the elements, needed for modeling a general CB framework, are outlined (cf. section 4.2).

4.1.1 Conceptual Spaces and Frames

Conceptual Spaces: The CB process is described as involving two knowledge structures, referred to in Fauconnier [1994] as "mental spaces" and in Fauconnier and Turner [2002] as "conceptual spaces" (also cf. [Goguen, 2006; Magnini and Strapparava, 1990]). Fauconnier defines conceptual spaces as "small conceptual packets constructed as we think and talk, for purposes of local understanding and action" [Fauconnier and Turner, 2002, pp. 102]. They can be seen as partial representational structures for understanding a perceived (or imagined) situation (especially in linguistics).

A conceptual space needs to be somehow connected to long-term knowledge, since it is considered a partial assembly containing interconnected elements that are typically structured by "frames" and "cognitive models" (cf. Fauconnier [1994, 1997]; Fauconnier and Turner [1998, 2002]). Fauconnier and Turner hypothesize that "elements" in the spaces correspond to activated neuronal assemblies, with interconnections corresponding to a "kind of neurobiological binding, such as co-activation" [Fauconnier and Turner, 2002, pp. 102]. Conceptual spaces can be activated in many different ways and for many different purposes, where the interconnected elements of a space "can be modified as thought and discourse unfold" [Fauconnier and Turner, 2002, pp. 41]. Conceptual spaces are built up "dynamically in working memory, but they can also become *entrenched* in long-term memory" [Fauconnier and Turner, 2002, pp. 103; emphasis added]. Thus, mental spaces "can be used generally to model dynamic mappings in

thought and language" [Fauconnier and Turner, 2002, pp. 41], because they are "built up in any discourse according to guidelines provided by the linguistic expressions" [Fauconnier, 1994, pp. 16].

For models concerned with the processing of CB, mental spaces could be represented in many ways, such as semantic network graphs, cases (in case-based reasoning), or an activation pattern of a neural-network at a given moment (cf. [Pereira, 2007, pp. 56]). In relevance to this thesis, concepts consisting of grouped conceptual entities (cf. section 1.3) are assumed as a means for representing mental spaces. Using this way to represent concepts, conceptual entities and relationships between them can substitute the notion of "elements" given in the above definition of a conceptual space for processing simulated computations (cf. [Fauconnier, 1994; Fauconnier and Turner, 2002]).

Frames: On a linguistically motivated account, Fillmore defines a frame as a collection of categories whose structure is rooted in motivating context experiences. Fillmore also emphasizes the role these experiences play in building word meanings (cf. Coulson [2006]; Fillmore [1982] and Chapter 7). This is paralleled in AI by Minsky's proposal of frames as data structures that represent commonly encountered, stereotyped situations (cf. Minsky [1974]).

Entities (e.g. knowledge pieces or beliefs) and relationships between them are the essence of representing conceptual spaces. A conceptual space contains "a partial representation of the entities and relations of a particular scenario", where frames play in this representation the role of a pattern representing "the relationships" that exist between the entities (cf. [Coulson, 2006, pp. 21] and section 7.2.2). Thus, when the entities and relations can be organized in patterns, as packages that we already know about, the organized packages are called *"frames"*, and the conceptual space is, thus, framed. "A single mental space can be built up out of knowledge from many separate domains", where several sources of knowledge help in building the spaces, such as patterns of experience and assertions by other people (cf. [Fauconnier and Turner, 2002, pp. 102–103]). In this sense, frames are "*entrenched* mental spaces that we can activate all at once" [Fauconnier and Turner, 2002, pp. 103; emphasis added].

Frames provide a kind of an *abstract prototype* of (interrelated) entities, actions, or reasonings that —unlike frames and scripts from early AI— can be dynamic (i.e. it changes with time, person, or context) and compositional (i.e. it may have many layers of abstraction). The principal frame that underly a given conceptual space is called the "organizing frame". For example, entrenched sources of knowledge for a conceptual space like 'BUS' could include frames of "transport means" and "container", whereas the

latter frame may not be the organizing frame in most situations. Frames are important in guiding the blending construction process to "recognizable wholes" of repeated patterns, but this does not result in a blend having one single frame. In fact, an acceptable blend should mostly inherit a mix of structures from the (organizing) frames of the input conceptual spaces that are forming it (cf. [Pereira, 2007, §3.2.1; pp. 57-58]).

Representing Spaces and Frames: Illustrations of conceptual spaces and frames in this thesis employ compact visual representations of graphs akin to semantic networks, in which nodes that identify conceptual entities correspond to the structured, interconnected elements of conceptual spaces. For example, a depiction of two mental spaces for Smullyan's Rate-Time Puzzle is given in Figure 4.1. In each one of the spaces, there are entities that correspond to:

1. two cities (NY and BOSTON representing *New York* and *Boston*, respectively),
2. a train (moving from one city to the other, where the direction of traveling is indicated by a pointing arrow),
3. the time (in hours) at which the train starts to travel (with the starting time being denoted by t_0 for the train leaving away from *Boston*, and by $t_1 = t_0 + 1$ for the one leaving away from *New York* an hour later),
4. the distance between the train and the city it departed from, and
5. the current location of the train.

This kind of visual illustrations is customarily used for simplifying the representations of what are being referred to as mental, or conceptual spaces. Due to the vague nature of spaces and frames in Fauconnier and Turner's treatment of CB[1], this way helps in delivering the basic entities in involved spaces (perhaps also their interrelationships) without the need to give detailed or more specific representations, which will definitely have to be based on one particular KR formalism or another. Similar examples that give two very simple conceptual spaces, representing "BOAT" and "HOUSE" domains, can be found in [Goguen, 2006, Figure 1] (or its variant in [Goguen, 1999, Figure 1]; also see Figure 4.2 below).[2] Classic examples, such as Goguen's, which are often cited in various classic work on CB, will be referred to as "*established examples*". Richer representations

[1] Refer to section 4.4.2 for more clarifications, and to section 7.2 for a proposed treatment.

[2] There is a little difference between the representations of "BOAT" and "HOUSE" given in [Goguen, 2006, Figure 1] and those given in [Goguen, 1999, Figure 1]. Figure 4.2 is based on the former (and so do the figures that follow in this chapter).

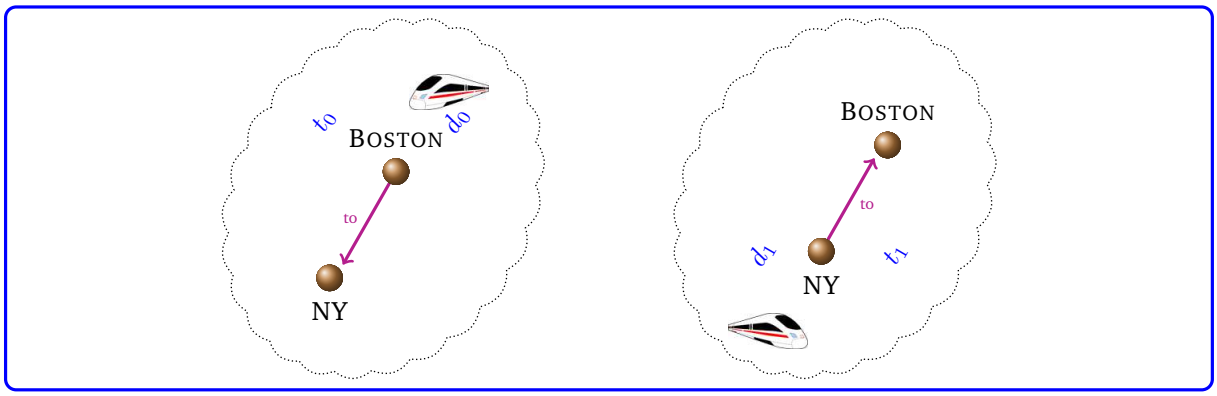


Figure 4.1: A simple illustration of the two conceptual spaces involved in Smullyan's Rate-Time Puzzle. In each space, entities exist that correspond to the two cities, a train, a direction of traveling, the distance between the train and the city it departed from, and the time (in hours) at which the train starts to travel from one city to the other. The starting time is t_0 for the train leaving Boston (left), and t_1 for that leaving New York (NY) an hour later (right).

of established examples of two conceptual spaces for "COMPUTER" and "VIRUS" are depicted in [Pereira, 2007, Figure 5].

4.1.2 Cross-Space Mappings

In order for a blending process to take place, it is (inherently) essential that the reasoner detects a correspondence between some elements in one space with others in another space. Interconnected elements (or frames thereof) in a given conceptual space can be linked during the CB process to corresponding elements (or frames, respectively) in other conceptual spaces.[1]

Figure 4.2 depicts a mapping between Goguen's established conceptual spaces that represent the "BOAT" and "HOUSE" domains. A more detailed visual illustration of a possible mapping between the mental spaces "COMPUTER" and "VIRUS" is shown in [Pereira, 2007, Figure 6] for explaining a cross-space mapping in the blending of the two spaces, forming a "COMPUTER VIRUS" blend.

It should not come as a surprise that this linking process plays a similar role to what an analogical mapping plays in finding corresponding entities in the structures of a given source and target domains in analogical reasoning (cf. section 2.2.1). Moreover, and also similar to an analogical mapping, this cross-space, or cross-domain, mapping

[1]Note that this 'linking' can be found in may ways, such as identity, structure alignment, or analogy. The reasoner's background knowledge and current context affect how the reasoner views elements in one space as being linked to elements in another. This also opens the door for a modeled version of CB that helps in computing forms of creative thinking (cf. Chapters 5 and 7).

does not have to be unique, since elements or frames of elements of one space could correspond to zero or more counterparts of another space. Therefore, and as will become clearer in the following part of the thesis, an analogy engine can be utilized to facilitate the process of blending in models seeking to attain the CB framework computationally (cf. Chapters 7, and 8).

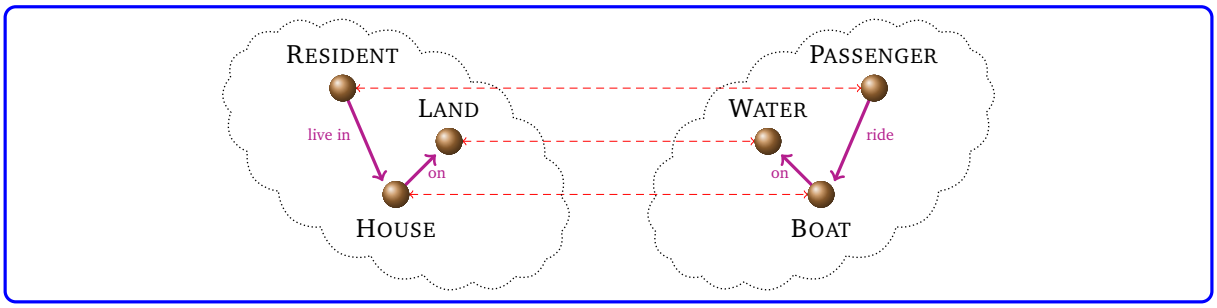


Figure 4.2: A mapping between Goguen's simple mental spaces that represent the "HOUSE" and "BOAT" domains (cf. [Goguen, 2006, Figures 1 and 4]).

4.2 A Cognitively Plausible View of Conceptual Spaces for the CB Framework

The CB framework is not yet well-formalized enough to precisely define notions such as "conceptual spaces", "mental spaces", "mental domains", and sometimes even "mental models"[1], or to sharply distinguish them from each other. The view undertaken in this text is that they may all be used interchangeably within the CB framework to only serve the same purpose of representing "knowledge domains" or "concepts" in a proposed modeling context (cf. section 1.3). (Though, of course, they can in general be used to capture different ideas in other texts.)

Throughout the rest of the thesis, "conceptual spaces", or simply "concepts", replace the use of "mental spaces" and the related notions. However, and before any further continuation of the main CB discussion itself, characterizations and representation assumptions are proposed in the following for the view hereby undertaken of "concepts" or "conceptual spaces" (in AGI modeling contexts).

[1]The mental models theory is proposed by Johnson-Laird, where models are considered "the natural way in which the human mind constructs reality, conceives alternatives to it, and searches out the consequences of assumptions" [Johnson-Laird, 1995, pp. 999]. The text does not further discuss this issue, but a view about model construction in general intelligence is proposed in Abdel-Fattah and Schneider [2013]; Schneider et al. [2013] (also cf. Clement [2008]).

4.2.1 Characterizations of Concepts

It should be kept in mind that the essence of the notions about "concepts" in cognitive science share at least the following underlying assumptions (cf. [Fauconnier and Turner, 2002; Fodor, 1998; Lamberts and Shanks, 1997; Mareschal et al., 2010; Murphy, 2004]):

1. For modeling aspects of intelligence in cognitive agents, the proposed perspective is based on Murphy's view that concepts are "the glue that holds our mental world together" (cf. [Murphy, 2004] and section 1.3.1).

2. Concepts are considered as the basic elements of thought that abstract conceptions (i.e. ideations), objects, or actions.[1]

3. Concepts can embody not all but "much of our knowledge of the world" (cf. [Murphy, 2004]), since we cannot always "comprehend every microscopic detail of entities in the world" (cf. [Clement, 2008]). Therefore, in modeling aspects of cognition, concepts are considered incomplete naive theories that are often idealized, always simplified, and may in minor cases be contradictory (the contradictions must, however, be maintained; see below and Murphy and Medin [1985]). Concepts can be developed at different levels of detail (cf. section 1.3.2), where a single concept may be used to account for many events, making it an efficient way to store knowledge.

4. Concepts are more useful when they represent the important interrelationships in a system, as opposed to being a collection of isolated facts. Cognitive science has shown that cognitive beings perceive their environments as *structures* and *changes in the relationships* among the structures (cf. [Clement, 2008; Fodor, 1998]).

5. Very few concepts are static, whereas most of the concepts can be changed with respect to many factors, such as time, person, and context.

The previous assumptions form a subset of the most predominant traits, common to existing, debatable views that characterize and define what concepts are in cognitive science. They serve guiding a general, computational modeling of the CB framework.

Of at least an equal relevance should be the more debatable views of how concepts seem to be represented for cognition. Basics of both view types (i.e. "characterizing" and

[1] A broad sense of distinction should be kept in mind between conceptions and concepts. With the exception of this subsection of "concepts" in cognitive science, I usually use "concepts" (on the modeling side) to indicate representations that can be used as surrogates (cf. section 1.3) that account for aspects of structure or knowledge of "conceptions" (i.e. ideation on the cognition side).

"representing" concepts in cognitive science) can help in AI models that computationally aim at simulating aspects of general intelligence. The view about representing concepts is introduced in the following as a further support for the discussions. The reader can refer to [Lamberts and Shanks, 1997; Mareschal et al., 2010; Murphy, 2004; Wrobel, 1994, for example] for extensive and comprehensive discussions of both views.

4.2.2 Representations of Concepts

It has long been assumed that concepts can be represented in the mind by the Aristotelian tradition of giving necessary and sufficient conditions based on properties. Modern approaches in cognitive semantics that focused on the notion of "concepts" as a cognitive phenomenon[1] have undertaken sharp turns by not neglecting natural fuzziness of conceptual processing, and presenting representation views that consider more plausible assumptions regarding "gradations of typicality" and the existence of "borderline cases" [Murphy, 2004, pp. 64].

The following are three of the most current general views, typically considered as theories of concept representations:

The Prototype Approach: Rosch's experiments resulted in a theory of human concepts that greatly differs from the classic tradition based on strictly explicit properties (cf. [Goguen, 2005; Rosch, 1975]). The "*prototype view*" of concept representation is based on innovative results showing that concepts exhibit prototype effects, which indicate membership degrees (based on frequency), correlating with similarity to a *central member*, or basic-level concept (cf. [Rosch, 1975]). This view explains why TABLE, for instance, has been (empirically) found more prototypical as a piece of FURNITURE than CLOCK, and why ORANGE is more prototypical of FRUIT than OLIVE (cf. [Murphy, 2004, Table 2.1; pp. 33]).

Rosch's theory views a concept as being represented based on a *summarized prototype*, which is assumed to have the average characteristics (or purposes) of the concept (e.g. a prototype of BIRD would be characterized by having wings, having feathers, and flying). There are *basic level concepts* that tend to (i) occur in the middle of concept hierarchies, (ii) have the most associated knowledge and shortest names, and (iii) be the easiest to learn. One of the shortcomings of the prototype view is its inability to account for concepts with dimensions that do not have set feature values, since concepts may not be static entities defined with a

[1] More precisely, cognitive semanticists treat meaning "as a cognitive phenomenon invoked to construe conceptual content" [Coulson, 2006, pp. 17].

fixed set of properties. For example, a dimension, say 'size', may either be represented as 'small', 'medium', or 'large'; or as some continuous measurement of 'size' (cf. [Murphy, 2004]). The prototype view is commonly applied in AI systems that represent concepts as attribute value sets [Pereira, 2007, pp. 48].

The Exemplar Approach: In the "*exemplar view*", proposed by Medin and Schaffer, representations of concepts are assumed neither to encompass an entire concept definition, nor to arrange a list of features according to their typicality. Instead, a concept is more or less the remembered set of the most salient exemplars (plus some fuzzily remembered ones) that could be consulted when one wants to make decisions regarding this particular concept. According to this view, there is no summary that stands for all members of the concept and, in a sense, there is no real concept. For example, one's concept of DOG is the set of all *dogs* that one remembers, and if one wants to figure out whether a newly perceived object is a DOG, one has to consult sets of exemplars and decide.[1] The exemplar view makes a good sense from a cognitive scientific point of view, since it can account for some psychological behaviors (cf. [Murphy, 2004]). However, at least it contradicts basic intuition, on the one hand, and gives no room for abstraction, on the other. Moreover, applying this theory (as it is) to represent "concepts" in computational systems would (unrealistically) require a potentially infinite memory. In addition, and given that the theory does not say anything about defining characteristics[2], it would not be possible to relate to a concept in some usage modes. For instance, how would the metaphorical usages of 'doggy' or 'snaky' be defined? The exemplar view is commonly applied in case-based reasoning systems that employ episodic memory to compare old and new problems [Pereira, 2007, pp. 48].

The Theory Approach: The "*theory view*" was introduced in Murphy and Medin [1985] to view the whole knowledge as a kind of a (consistent) theory. It is also referred to by the "*knowledge view*". This view is built (in some sense) upon the previous two views, but it emphasizes that concepts are "the part and parcel" of one's "general knowledge" of the world. Murphy and Medin's approach to represent concepts argues that "concepts are part of our general knowledge about the world": they

[1] This consultation assumes that the perceived object is not merely fairly similar to some of the exemplars in the set of dogs, but the object should be mostly similar to more of the exemplars in the set of dogs than of those in other sets of exemplars. The consultation is assumed to be done as extremely quick as the speed of thought —something very difficult to attain in current computational models, if at all.

[2] That the exemplar theory does not say anything about defining characteristics can be considered an advantage, because problems for the classical views are no longer problems for this view (cf. [Murphy, 2004, pp. 50]).

are not learned in isolation but as "part of our overall understanding of the world" [Murphy, 2004, pp. 60]. Knowledge about animals, for example, integrates with knowledge about biology, behavior, and climate, as well as other domains. This integration relation works both ways, so that concepts both are influenced by whatever knowledge and experience we already have, and cause a change in our overall understanding. This gives a pressure for concepts to be "consistent" with whatever else one knows (cf. [Murphy, 2004; Murphy and Medin, 1985]), and motivates the use of the "*theory*" label in some sense. One may consider the representation of a specific concept as being based on smaller parts (or sub-theories) that describe the concept with facts about related conceptual entities (or concepts), as well as causal connections between them all.

In the context of this thesis, the previous views are seen as a guide-lining background that utilizes existing theories in suggesting ways in which conceptual domains can be represented. Hopefully, this helps in facilitating the computational modeling of general intelligence aspects that agree with cognitive science-based results.

Analysis: On the one hand, and as concluded in [Murphy, 2004, pp. 64–65], it is extremely important to point out that no "single form of conceptual representation will account for everything". Even according to the prototype theory itself, exemplar knowledge must exist side by side with prototype knowledge: e.g. if one encounters an object for the first time, then one can form a prototype based only on that single exemplar. Also, neither the prototype view denies that people learn and remember exemplars, nor the exemplar view prevents people from relying on summary representations rather than specific exemplars in making judgments about a concept (especially when the concept representation has grown to be mature enough). Moreover, the theory view advocates must "admit that there is an empirical learning component to concepts", whereas the exemplar theorists must agree that a level of general knowledge (that is separate from exemplar knowledge) does in reality affect concepts and their usage. Thus, several views need to be combined with one another to form a comprehensive, complete theory about representation of concepts.

On the other hand, the theory view is commonly applied in logic-based AI systems or systems that use semantic networks[1], which renders this view more connected to previous (and forthcoming) ideas considering the modeling of concepts for CB and systems showing GI aspects. The view further supports, and fits better to, the overall text and its underlying framework, namely HDTP. Recall that the KR language of the

[1]For example, Copycat (cf. [Hofstadter, 1984; Mitchell, 1993]).

HDTP framework is a many-sorted, first-order logical one, and can thus account for beliefs, facts, and laws, in terms of axiomatizations, and also account for the overall knowledge in terms of theories (cf. sections 3.1 and 3.2). Therefore, the theory view will be used here for motivating the representation of conceptual domains.

However, this adoption raises some deficiencies, which are typical when applying the view itself. First of all, the view is referred to by the "*theory*" view, though it must be accepted under the natural constraint that "people's naive theories are incomplete and in some cases contradictory, given our incomplete knowledge and understanding of the things around us" [Murphy, 2004, pp. 61]. Striving to achieve a cognitively plausible way of representing concepts (in artificial models) should not imply that humans think only in terms of (scientific) theories. So, one must not confuse the view about theories (in the sense of a cognitive agent having an incomplete mental theory that mostly depends on experience) with consistent logical theories (in the sense of official, rigid, scientific theories).[1] Secondly, note that if a model applies the ideas of Murphy and Medin's approach to modeling similar representations for concepts in an agent's KB, it would require that the whole KB is viewed as a network of (sub) concepts, with coherent links and rules that should remain consistent. One should make a decision, for example, about the limit to which concepts are allowed to decompose into sub-theories (and what entities would we like to represent at all, in the first place). This raises a challenge similar (if not identical) to the decision challenge regarding how deep the level of details in KR should be (cf. section 1.3.2).[2] A related challenge would be to consider how dynamic changes of concepts (and their comprising conceptual entities) may be represented. Representing dynamic concepts in terms of theories would raise sorts of challenges akin to those typically encountered in non-monotonic reasoning, such as the problem of (tractably) maintaining consistency.

4.3 The Conceptual Blending Framework

CB is being more widely considered an important part of cognition now than ever. There is currently more to CB than just facilitating the creation of new concepts and ideas by constrained combination of available knowledge. It does provide explanations of central features of human-comparable intelligence like the ability to understand, learn, reason, and creatively construct new concepts and theories. Hence the inevitability of

[1] But, of course, one must work hard to rectify arising inconsistencies as the knowledge and experience of the agent develop in a computational modeling that adopts this view.

[2] Solving this challenge is an important goal, since this would hopefully contribute to overcoming the related challenge of formalizing the CB framework for computational models (cf. section 4.4.2).

incorporating CB in AGI models, though there are still very few formal or algorithmic accounts on CB.

4.3.1 The Network Model: Constructing Blend Spaces

Broadly speaking, the process of CB operates by mixing two input concepts (e.g. in the sense explained in section 4.2) to form a third one that basically depends on mapping identifications between the former two. The third, generated concept, the "*blend space*", maintains partial structure from the inputs and adds an emergent structure of its own.

In the process of generating (or computing) a blend in the CB framework, three steps usually take place. The ordering of the steps is not necessarily sharp, and can be changed. Moreover, the blending process may even need several iterations of these steps. They can be broadly characterized as follows (cf. Fauconnier and Turner [2002]; Pereira [2007]):

Composition: In the composition step, a new space is constructed by pairing selective constituents from the input spaces, then projecting them into the constructed space. The space, constructed during this step, may only serve as an initial trial to fill-in details in forming a useful conceptual blend. A potentially useful (or acceptable) blend will also be referred to as a feasible blend. It should neither be unique nor meaningful at this point of the CB process, and will therefore be referred to as a candidate —a "*blend candidate*". The composition step is also called the *fusion* step, where the pairing of two different elements, one from each of the input spaces, fuses these elements into one in the blend candidate.

Emergence: This step is a source of emergent contents, that fills some gaps in the potential blend. It helps in completing existing contents that were fused in the newly constructed blend during the composition step (cf. Coulson [2006]). In the emergence step, which is also called the *completion* step for an obvious reason, a pattern in the formed blend (that is, the blend candidate from the composition) is filled with projecting structures that match long-term memory information (e.g. based on knowledge of background frames).

Elaboration: The actual functioning of the blend comes in the third step, the elaboration step, in which a performance of cognitive work within the blend is simulated according to its own emergent logic (i.e. according to how the reasoner views the projected structures' functionality). This simulation is called *running* the blend.

Figure 4.3 illustrates the prototypical four-space model of CB as a network, in which two concepts, denoted by SPACE$_1$ and SPACE$_2$, represent the input conceptual spaces. Common parts of these inputs are matched by "identification", where the matched parts can be seen as constituting a GENERIC space. The BLEND space has an emergent structure that arises from the blending process and consists of some matched and possibly some of the unmatched parts of the input spaces in the network of concepts.

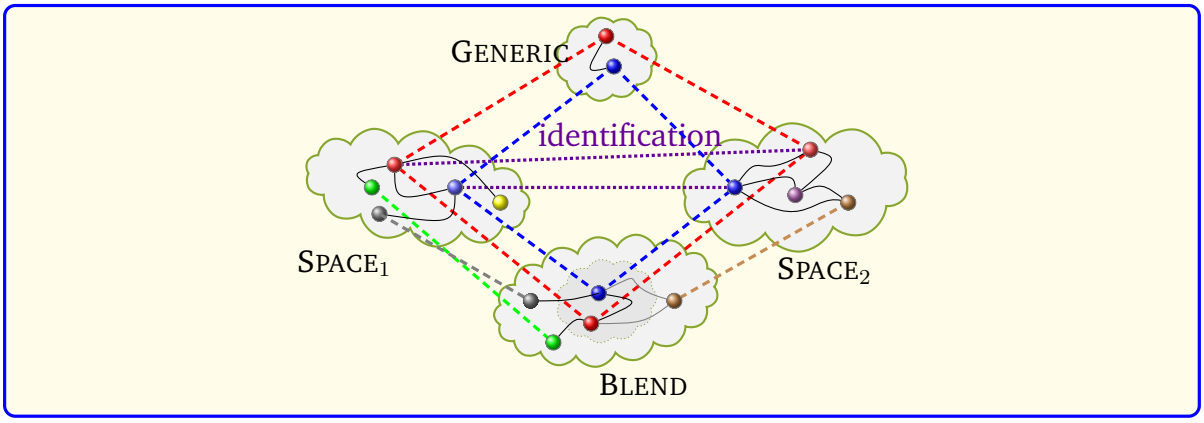


Figure 4.3: The four-space, network model of CB: common parts of SPACE$_1$ and SPACE$_2$ are identified, defining a GENERIC space and a BLEND. The connections within the concepts reflect the internal structures of relations among the conceptual entities.

Example: To further explain the CB process and the underlying network model, consider again Smullyan's Rate-Time Puzzle and the illustrations given in Figure 4.1 of its (isolated) input conceptual spaces. Figure 4.4 depicts the blending process for reasoning about an answer to the puzzle using the network model of CB:

1. In the blending process, commonalities between the input spaces of Figure 4.1 can be first found by means of cross-space mappings and generalized into the "GENERIC" conceptual space shown in Figure 4.4. The GENERIC space "maps onto each of the inputs and contains what the inputs have in common" [Fauconnier and Turner, 2002, pp. 41]. It holds commonalities between SPACE$_1$ and SPACE$_2$, such as the two cities, as well as the spatial, temporal, and directional knowledge about the moving train.

2. Composition allows elements from the inputs to fuse into the candidate blend, such as the cities, the distances between them and the trains departed from them, and a (general) time point at which the space runs. This also makes available new relations (between the fused elements) that might have not existed in the

(isolated) inputs: e.g. only in the blend we have two trains, and only in the blend one would think of a decreasing distance between these trains.

3. Completion can fill-in missing details by bringing additional structure to the BLEND, if necessary. For example, in the BLEND space, the reasoner integrates the familiar frame of "two trains moving opposite to each other", by virtue of which the BLEND runs (i.e. elaboration). This allows to apply intuitive movement laws in opposite directions, which eventually make the decreasing distance between the moving trains reaches zero, so that they meet at at some location. The reasoner now focusses on a specific time point, $t_m > t_1$, and a specific location, at which the two trains meet somewhere between the two cities. This affects the reasoner's conceptualization of the distance between the trains and the cities, allowing the reasoner to focus on the distances from the meeting location to BOSTON (as stimulated by the puzzle statement).

4. The blending (of train locations, in particular) allows the reasoner to discover that both distances (between BOSTON and each one of the trains) must be the same regardless of where the (blended) meeting location lies or what time exactly t_m is (because $t = t_m$ in this space, and the trains share the same location).

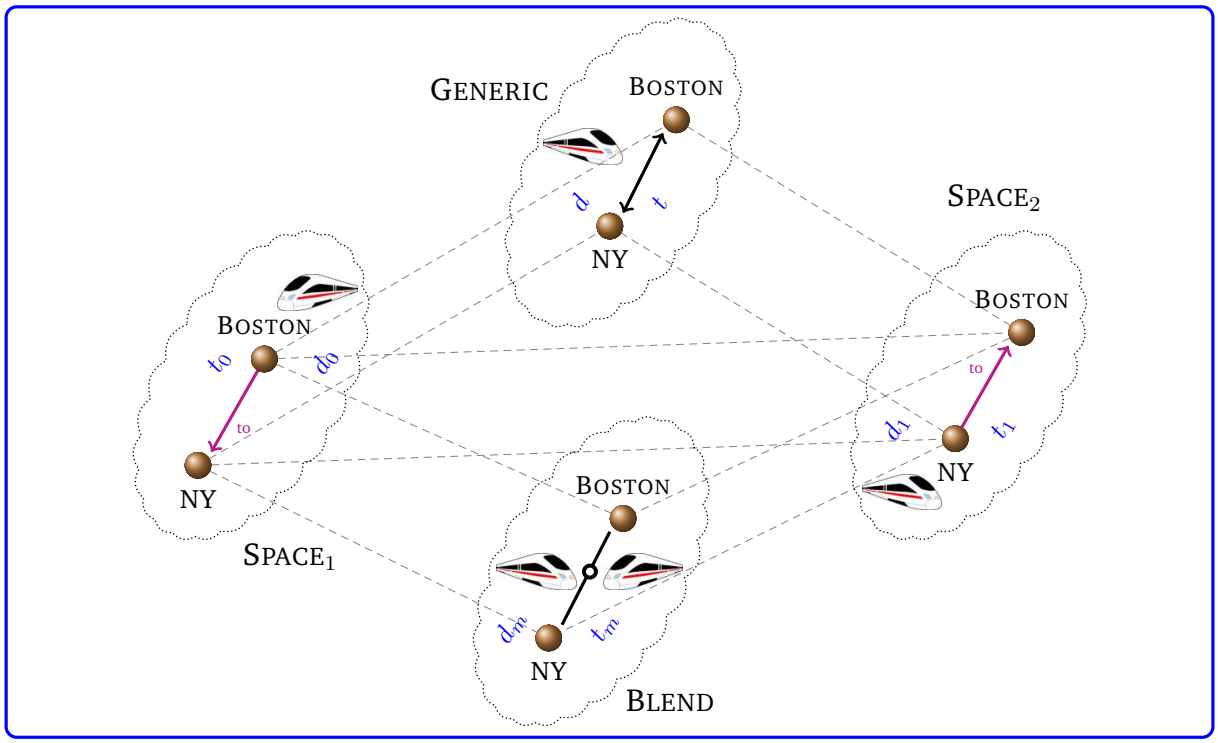


Figure 4.4: Solving Smullyan's Rate-Time Puzzle by the network model of CB.

Fauconnier and Turner laid out the central principles of the network model by schematizing a similar reasoning for a different classic puzzle. In [Fauconnier and Turner, 2002, pp. 40–44], Koestler's *Riddle of the Buddhist Monk* is used for revealing how the blending processes can be elaborated using the network model of CB (cf. Fauconnier and Turner [1998, 2002]; Koestler [1964]).

4.3.2 An Overview of Former Accounts

Within the realm of modeling higher-level cognitive mechanisms related to achieving general intelligence, work on formalizing computational models for CB is very little. Moreover, most of those formalizing models have neither focused on how to mainly compute blending in general frameworks, nor on utilizing CB in modeling GI aspects for cognitive agents. It can even be viewed the other way round, since they rather utilized the way their proposed systems behave in computing a form of CB restricted to their systems. Nonetheless, these contributions are credited for being prominent in modeling a computation of CB. Amongst them, the following ones can be mentioned (cf. [Goguen, 1999; Lee and Barnden, 2001; Pereira, 2007; Veale and O'Donoghue, 2000]):

1. Goguen's approach to describe CB by an algebraic semiotic formalism provides a theory and a computational treatment of sign systems (cf. Goguen [1999]). Blends in this approach are described as structure-preserving mappings of systems of signs (i.e. sign systems are used as input spaces).

2. Veale and O'Donoghue's approach provides a computational model relying on a metaphor interpretation system that is an extension of their Sapper system (cf. Veale and O'Donoghue [2000]).

3. Lee and Barnden's approach analyzes reasoning of counterfactual conditionals from a CB perspective based on their ATT-Meta system (cf. Barnden et al. [2002]; Lee [2010]; Lee and Barnden [2001]).

4. Pereira's Divago system implements CB into a model of computational creativity that uses a parallel search engine based on genetic algorithm (GA) (cf. Pereira [2007]).

One of these formal accounts on CB is the classical work of Goguen and his colleagues, which is especially influential to the approach undertaken in this thesis. Goguen formulated CB at an abstract level in category theory (cf. [Diaconescu, 2008, Chapter 2]), describing blending of concepts using algebraic semiotic formalisms (cf. [Goguen,

1999, 2006; Goguen and Harrell, 2004, 2010; Goguen and Malcolm, 1996, among others]). Goguen built on insights from CS that discrete structures can be described by "*algebraic theories*" with extra structure. He proposed to treat theories as being computational treatments of systems of "signs", giving his formalizations credit as the first attempts to characterize specific CB components. The theories included axioms, sorts, and constructors, where sorts can be used to enrich the representations of Fauconnier's conceptual spaces with types (cf. [Goguen, 2006, §1]), while constructor functions build complex signs from simpler ones (cf. Goguen [1999, 2005]). These sign systems can therefore handle features for special types of representations. For example, in a representation of the words in a sentence, or in representing the visual constituents in a diagram, the components can be treated as complex signs that have parts, in which relations between the parts can only be put together in certain ways to constitute the representations (cf. Goguen [1999, 2005]).

Combining notions from semiotics and category theory, Goguen applied algebraic semiotics to formalize CB by describing blends as being semiotic morphisms of sign systems (simply stated, these are structure-preserving mappings between signs). Goguen's version of CB can be described by the diagram in Figure 4.5, with C_1 and C_2 representing two input conceptualizations (i.e. two given sign systems being treated as input spaces). In this diagram, two input concepts C_1 and C_2 are related by correlations that are induced by a generalization G. Like in the network model, a generalization G is first looked for, then a blend B is constructed in such a way as to preserve the correlations between C_1 and C_2 established by the generalization G. This may involve taking the morphisms to B to be *partial*, in that not all the structure from C_1 and C_2 is mapped to B. One should in any case be assured that the blend respects the relationship between C_1 and C_2 implicitly established by the generalization. So the diagram will commute and in fact will be a pushout (cf. [Goguen, 1999]).[1]

Goguen's approach implicitly accounted for basic processes of CB. It however left out the specification of many details of how important issues can be realized, such as the three traditional steps of the construction process in the network model of CB (cf. section 4.3.1), and Fauconnier and Turner's *"optimality principles"* (cf. section 4.4.1).

A standard example, discussed in Goguen [2006], is that of the possible conceptual blends of the concepts HOUSE and BOAT (cf. Figure 4.2) into concepts such as BOATHOUSE and HOUSEBOAT (cf. Figure 4.6), as well as other less-familiar (and less-obvious) blends (see e.g. [Goguen, 1999, Figure 7] and [Argamon et al., 2010, Fig-

[1] Refer to the briefing of Goguen's work given in [Pereira, 2007, pp. 63–64] for more elaborations about signs and pushouts. See [Diaconescu, 2008, Chapter 2] for a more detailed, rather condensed, presentation on category theory.

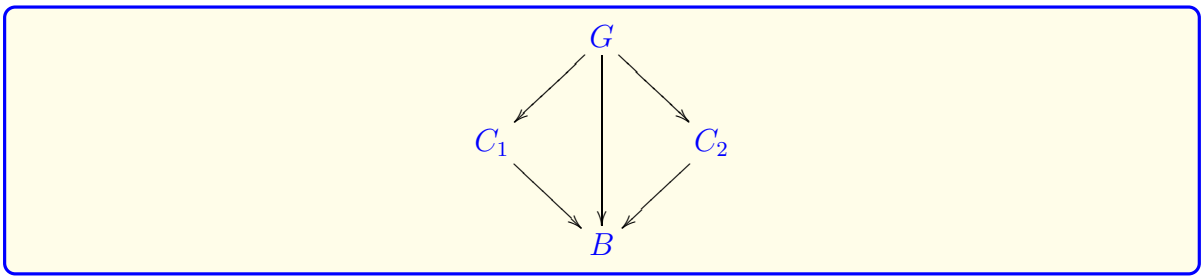


Figure 4.5: Goguen's version of CB (cf. Goguen [1999, 2006]).

ure 12.3]). According to Goguen and Harrell [2010] (cf. Figures 4.2 and 4.6), part of the conceptual space of HOUSE reflects that a RESIDENT lives in the HOUSE, and part of BOAT reflects that a PASSENGER rides on the BOAT. These parts can be aligned (as already illustrated in Figure 4.2). Conceptual blends are then created by combining features from the two spaces, while respecting the constructed alignments between them. As argued in Fauconnier and Turner [2008], newly created blend spaces will coexist with the original spaces. For example, the concepts of HOUSE and BOAT are still needed in Goguen's example, even when we are aware that a BOATHOUSE has some relation to both HOUSEs and BOATs. In [Argamon et al., 2010, Figure 12.3], a quite unusual BOATHOUSE blend is given, in which the BOAT ends up living in the HOUSE. The idea is that RESIDENT is mapped to BOAT with no type checking, which gives a type of metaphor referred to in literary theory as "personification": an object being considered a person (cf. [Goguen and Harrell, 2010, pp. 303]). Examples like the previous one link to earlier work on computational aspects of blending, which has been carried out in cognitive linguistics (cf. for example Veale and O'Donoghue [2000]).

Another more recent algorithmic account is given in Pereira [2007], where the CB mechanism has been implemented in a system that uses a parallel search engine based on genetic algorithm (GA). This consists of six modules: (i) a *Knowledge Base* in which a set of concepts is defined, (ii) a *Mapper* which builds structural alignments between concepts, (iii) a *Blender* which takes in a structural alignment and produces a set of projections that implicitly define the set of all possible blends, (iv) a *Factory* which is a reasoning mechanism based on a GA, (v) a *Constraints* module (which is based on an implementation of the *"optimality principles"*; cf. section 4.4.1), and (vi) a logic and rule-based *Elaboration* module. In Pereira's account, a pair of concepts is selected from the KB and passed to the *Mapper* which builds a structural alignment between them. This is then passed to the *Blender*, which passes the set of all possible blends to the *Factory*. The *Factory* then uses a GA parallel search engine to search for the blend that best complies with the evaluation given by the *Constraints* module. When the GA reaches a satisfactory solution, or after a specified number of iterations, the *Factory*

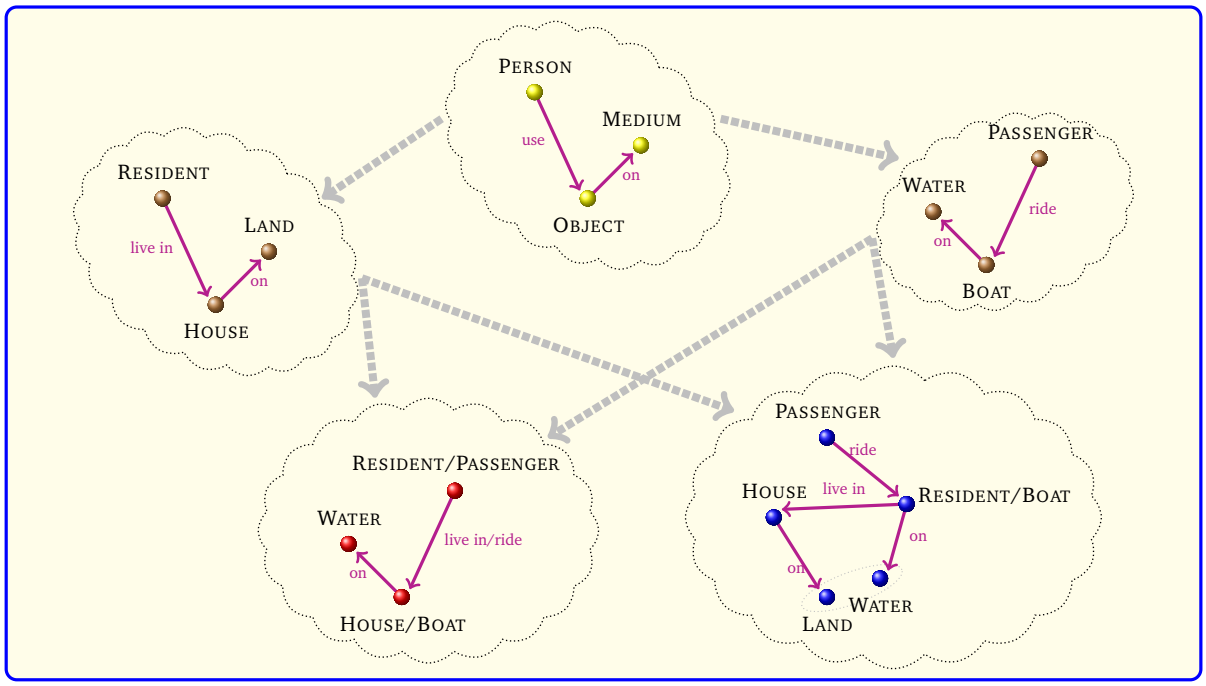


Figure 4.6: Goguen's standard example for blending the HOUSE and BOAT domains. The domains, two blends (HOUSEBOAT and BOATHOUSE) and a generalization are shown (cf. [Goguen, 1999, Figure 6]).

stops the GA and returns the best solution.

Other work on CB can be found in several application domains, such as in creative production of new conceptions (cf. Abdel-Fattah et al. [2012b]; Guhe et al. [2011]; Pereira [2007] and Chapter 5), investigating meanings of novel noun compositions (cf. Abdel-Fattah [2012]; Abdel-Fattah and Krumnack [2013]; Coulson [2006]; Veale and O'Donoghue [2000] and Chapter 7), reasoning about counterfactual conditionals (cf. Abdel-Fattah et al. [2013a,b]; Lee and Barnden [2001] and Chapter 8), and understanding certain mathematical developments (cf. Alexander [2011]; Hersh [2011]; Lakoff and Núñez [2000]; Martínez et al. [2012]). It is expected that formalizations of CB and its many peculiarities produce significant development in AI, when basic parts of its framework are (partially) formalized. Unfortunately, some components of the CB framework still pose hard challenges to this end, because they lack a clear specificity. A computational realization of a formal framework of CB is still a big challenge in AI.

4.4 Challenges and Weaknesses of CB

Based on the previous discussions and its relatively short literature, CB is still a young, ongoing research paradigm. As such, it must be the case that CB faces valid critics.

Constructive criticisms hopefully point the evolution of the research to the right direction, by answering the questions the critics are mostly concerned with. More elaborate formalizations of the CB components, in particular, are expected to produce significant development in modeling aspects of general intelligence and cognition.

As an important part of cognition, CB already proved its importance in expressing and explaining some cognitive phenomena, such as metaphor-making, counterfactual reasoning, as well as its usefulness in analogical reasoning and creating new theories. Nevertheless, there is no general computational account of blending as a framework model, that has been proven powerful enough to cover all the examples in the literature.[1] Only few accounts have been given to formalize CB or its aspectual components, yet they are not broad enough to suit generic computational accounts of CB (cf. Alexander [2011]; Goguen [2006]; Martínez et al. [2012]; Pereira [2007]; Veale and O'Donoghue [2000] and section 4.3.2). Nonetheless, the overall CB framework itself suffers from the evident lack of formality. Even Fauconnier and Turner's insightful and well-discussed ideas in their intellectually profound book, entirely dedicated to discussing the framework, are still insufficient to cover various sorts of challenges to characterize and formalize (or even well-define) the framework.

4.4.1 Principles of Optimality for Conceptual Integration

Developing characterizations or formalizations for computational models of CB faces many challenges. One is raised by the well-known "*optimality principles*" of CB, which are broad "principles and pressures that guide the formation of integration networks" (cf. [Fauconnier and Turner, 2002, pp. 90]). They provide the network model of CB with general guidelines, which are assumed (in theory) to help in generating feasible blends and distinguishing good blends from highly unacceptable ones. However, they also provide (in practice) an extremely challenging stumbling block in characterizing, formalizing, or modeling the CB framework.

Vital Relations and Constitutive Principles: It is obvious that the selective projection in the blending process brings considerable complexity when searching the (huge) space of potential blends for a new blend construction. There is, in general, no constraint preventing one specific element or another of one particular input space or another from being included in (or excluded from) a resulting, potential blend or another. Moreover,

[1] Nor is it promised to give such an account in this thesis. The thesis rather focusses on proposing, at a conceptual level, how the mechanism can be better utilized in many directions to emphasize the importance of overcoming modeling challenges.

elements from input spaces can be projected into a resulting blend space in at least two ways: (i) by fusing them with their counterparts (e.g. NY in the blend of Figure 4.4 is a fusion of the two corresponding NY entities of the inputs), or (ii) by accompanying them with other elements in the blend (e.g. one can think about two meeting trains in the blend of Figure 4.4). Projection of the latter type is referred to as "composition", in contrast with the former type: "fusion".

Under blending, the relatively large amount of conceptual knowledge, available in (representations of) input spaces, are compressed in a blend within a network of conceptual spaces. The establishment of connection links between mental spaces and blended spaces aims at maintaining "global insight, human-scale understanding, and new meaning" [Fauconnier and Turner, 2002, pp. 92]. Fauconnier and Turner refer to links between the input spaces by "*outer-space links*". These links can be compressed into relations inside the blend, which are referred to by "*inner-space relations*". Blending plays "imaginative tricks" with these links, so that outer-space links between the inputs have compressed inner-space counterparts in blends. For instance, conceptual relations like "cause-effect" or "time" are scaled down to tighter "cause-effect" or briefer "time" in blends, and incompatible physical spaces may possibly be compressed into the same physical space in blends (cf. [Fauconnier and Turner, 2002, pp. 93]).

Fauconnier and Turner hypothesize that certain such conceptual relations "show up again and again" in compressing the input spaces under the blending process, and call these all-important, conceptual relations "*vital relations*". They identify, and deeply discuss, 15 different types of these repeatedly shown-up, conceptual relations in [Fauconnier and Turner, 2002, pp. 93–102], and propose to maximize and intensify the vital relations as a guiding constraint for identifying good blends. Fauconnier and Turner present a handful of examples to explain the roles of relations in human thinking, but fail to (i) generalize the major aspects characterizing the given examples, (ii) abstract a description of what they call "maximization and intensification of vital relations", (iii) distinguish the level their 15 relations are presented at from the level these relations are widely understood[1], or (iv) clarify why they chose these 15 relations in particular. Moreover, and although the optimality principles suggest strategies for (somehow) characterizing and optimizing emergent structures of blends, none of these strategies is constitutive enough for the blends. In other words, the strategies are governing (but not constitutive) principles, in the sense that they do not tell us what should blends exactly constitute or how blends precisely are constructed from compressing or fusing

[1]For example, Fauconnier and Turner consider "identity", "analogy", "disanalogy", "property", "similarity" as different vital relations with meanings different from what may come directly into one's mind about such vital "relations".

entities from the (outer-space links of) input spaces (cf. [Fauconnier and Turner, 2002, pp. 310–311]). Clearly, tens of distinguishable, yet operational, ways could be suggested to interpret Fauconnier and Turner's unsatisfactorily defined notions, depending on purpose-oriented examples (cf. Pereira's Divago system in Pereira [2007]).

Governing Principles: The proposed "principles and pressures" that are assumed to govern blend construction are introduced and extensively discussed in several parts of [Fauconnier and Turner, 2002, e.g. Chapters 6, 7, 8 and 16]. The discussion give a glimpse of how perplexing it is for models to characterize such a fundamental feature of the (not yet well-defined) CB framework. Moreover, "the relative weight of a guiding principle can depend on purpose" [Fauconnier and Turner, 2002, pp. 330]. Therefore, the optimality principles are sometimes competing or working against each others: the individual influence of one principle varies according to the situation, and may grow or decrease whilst another's influence decreases or grows, respectively. For instance, an optimality principle (called "integration") requires that blends should constitute a tightly integrated scene that can be manipulated as one unit. Whilst, another (called "pattern completion") requires elements in potential blends to be completed by using already existing, integrated patterns and frames as additional inputs. The "relevance" principle also necessitates that important outer-space links between inputs have corresponding compressions in blends (cf. [Fauconnier and Turner, 2002, pages 328, 329, and 333, respectively]).

To the best of my knowledge, there is only one contribution in the literature that subjectively studies these principles, by measuring some blend examples of two specific domains (HORSE and BIRD; cf. Pereira and Cardoso [2003]). The work by Pereira and Cardoso is exceptional in the sense of proposing for a first time formalizations of the optimality principles, for the sake of building an AI-based system. On the other hand, the proposed set of (constraint-restricted) measures in Pereira and Cardoso [2003] tries only to "reflect as much as possible the rationale behind each principle", based specifically on reporting experiments using the Divago system (cf. [Pereira and Cardoso, 2003, §4]). Objective studies of the principles that generally define and formalize them, or even specify these principles in more detail, are so far not present (but are still necessarily needed).

4.4.2 What Concepts are "not" Blends?

The most troublesome (and most obvious) weakness of CB is its apparent "vagueness and lack of formality across its many aspects" [Pereira, 2007, pp. 66]. It is unclear what

exactly conceptual spaces are, despite the many works that thoroughly study them, and although they have been used extensively in explaining functionalities of CB (as well as in other areas of linguistics). Given the (bold) claims by Fauconnier and Turner of CB as a theory of cognition, and given the (undeniable) fundamental roles that CB plays as an aspect of cognition, it is still unclear to which extent mental spaces are cognitively plausible or computationally feasible. "This problem of definition of mental spaces becomes harder when discussing domains and frames" [Pereira, 2007, pp. 66], which are at least as unclearly characterized as mental spaces (for the purpose of computational modeling, in particular). If the CB framework employs both frames and mental spaces, with blending being defined only in terms of the latter (i.e. mental spaces), then can we tell the difference between frames and mental spaces in the CB process? I think this is very hard, if at all possible, because a singled out frame cannot be prevented from being thought of as a mental space, and a mental space with tightly connected entities can be considered as a single frame (cf. section 7.2.2). A distinction needs to be made.[1] The CB framework according to Fauconnier and Turner's presentation may be irrefutable, which renders it susceptible to the falsifiability problem. That is, as long as CB is not falsifiable, it cannot be a theory either. As also implied in Pereira [2007], this obscurity further affects the vagueness of defining other aspects of CB, such as (defining or computing) the optimality principles.

People may sometimes trace concepts back to original components, so that concepts are seen as (evolutionary) blends of (developed) combinations of basic constituents, and so on.[2] For example, youngster biological creatures (e.g. human children) may be seen as a result of continuously blending inherited chromosomes that carry characteristics of their parents (or even grandparents) from a microscopic, DNA-based level up. Thus, it is neither common to clarify whether or not input spaces can themselves be considered blends, nor to clarify what concepts cannot be considered blends in the first place. That is, it is not clear to what degree we are allowed to decompose spaces, or at which level we should stop to continue blending blends.

[1] This is one of the most important reasons why a decision in this thesis has been taken to differentiate between whole "concepts" and their composing, interrelated "conceptual entities". The former notion generally indicates "conceptual spaces" or "mental spaces" that represent knowledge conceptions (or ideas) in modeling aspects of cognition, whereas the latter refers to the constituents (e.g. beliefs, frames, relations, etc.) of such spaces (cf. section 1.3 and Chapter 7).

[2] In fact, this still makes perfectly sense according to the current characterization of CB. Fauconnier and Turner discussed the idea of multiple blends, where they emphasize that blending can operate over any number of mental spaces as inputs, because the product of blending can become the input to a new operation of blending [Fauconnier and Turner, 2002, Chapter 14].

4.4.3 Strengths Vs. Weaknesses: Searching for a Missing Link

Despite the evident strengths of the cross-domain, multifaceted mechanism of blending in explaining many aspects of human-level intelligence, the previous discussions also clarify evident weaknesses to satisfactorily compute blends within the CB framework in its current form of presentation. The inability to formally characterize or define fundamental facets of the framework can block further advancements in modeling the blending process. I think that adequate characterization and formalization of CB and its aspects will overcome such challenges. For example, and although it did not consider many important formalization challenges, Goguen's approach succeeded in coherently formalizing notable aspects of CB which allows to suggest and implement algorithms to compute blends. In my opinion, this was mainly because the domain of discourse has been well-specified before building further works on it (even though it was restricted to systems describable as sign systems).

For achieving a more generally applicable theory of the CB framework, a missing link needs to be established between (i) mental processes responsible for our cognitive ability in characterizing feasible, meaningful blends (as well as rejecting nonsensical ones), and (ii) formally characterizable processes that could reasonably simulate these mental processes. Such a missing link would render Fauconnier and Turner's interesting insights amenable to computational modeling. I think such a link is related to our cognitive ability of making coherent combinations of our knowledge and conceptualizations.[1] In addition, defining formal semantics for CB would help in restricting the discussions to meaningful parts of the world that reflect a current context or situation. One would perhaps need a mathematical framework for representing the CB framework to achieve that, but this would be harder done than said.

[1] Formalizing components that appear in the CB framework could be connected to a formalization of coherence in thought. See [Joseph, 2011; Murphy and Medin, 1985; Thagard, 1989, 2002, for example] for more about 'coherence in thought'. More discussions are also given in Chapter 6.

Part II

Applicability within Cognitively Inspired AGI Systems

5

Roles of Multifaceted Mechanisms in Logic-Based Computational Creativity

A central, distinguishing capacity of humans is being creative in, for example, utilizing acquired knowledge and skills for solving newly encountered problems or inventing novel conceptualizations. Therefore, systems that intend to model aspects of GI have to be endowed with the capability to somehow show a mode of creativity. Over many centuries painters, writers, poets, and other workers in the creative arts have frequently discussed creativity (cf. Runco and Pritzker [1999]). However, creativity is usually not considered a major issue in current research mainstream focused on modeling aspects of cognition and intelligence (though it clearly does deserve to be considered as such).

Creativity (be it natural or artificial) is approached in the current chapter as a specific means to distinguishing human-comparable intelligence from other forms of intelligence, since it is a clear-cut aspect of human-level general intelligence. The idea addressed in this chapter is that some creativity challenges are possible to be modeled by taking multifaceted cognitive mechanisms into account. Creative production of conceptions is proposed to be achieved through computation of generalizations of concepts and theories, reducing it to, particularly, analogy-making and concept blending (cf. Chapters 2 and 4). The latter mechanisms can best be modeled using (non-classical) logical approaches (also cf. Chapters 6, 7, and 8). So, the chapter also argues for the usage of logic-based approaches in modeling manifestations of creativity in order to step further towards the goal of building computational models of artificial general intelligence and creativity.

5.1 Introduction

Broad interest in creativity is as old in time as the human history of cognition.[1] But the study of creativity from a scientific perspective "has been neglected until the second half of the twentieth century", because (as also pointed out in Sternberg and Lubart [1999]) scientific schools in the early twentieth century (e.g. behaviorism; cf section 1.1) "devoted practically no resources at all to the study of creativity" (cf. [Pereira, 2007, pp. 7]). Guilford's contribution to studying creativity as an important attribute of intelligence is typically cited as the foremost turning point in this regard (cf. [Guilford, 1950, 1967, in particular]).[2] Since then, modern definitions of creativity as an intelligent aspect are moving away from the mere focus on aesthetics, constructions, or discovery to considering creativity as "a social phenomenon that is facilitated by some social factors, and inhibited by others" [Runco and Pritzker, 1999, pp. 511], or even to confronting recent issues about our relationships with intelligent machines (cf. McCormack and d'Inverno [2012]).

During the last decades many cognitive abilities of humans have been modeled with computational approaches trying to formally describe such abilities, to develop algorithmic solutions for concrete implementations, and to build robust systems that are of practical use in application domains. Whereas in the beginnings of AI as a cognitive science discipline the focus was mainly based on higher cognitive abilities, like reasoning, solving puzzles, playing chess, or proving mathematical and logical statements (cf. Newell and Simon [1956, 1963]), this has been changed during the last decades. In recent years, many researchers in AI focus more on lower cognitive abilities, such as perception tasks modeled by techniques of computer vision, motor abilities in robotic applications, text understanding tasks requiring the whole breadth of human-like world knowledge, etc. In experimentally studying higher cognitive abilities, some cognitive scientists are concerned with low-level factors that could affect the functioning of these abilities. In chess-playing, for instance, studies investigate how better chess-master players win more often.[3]

[1] Ancient examples include the artificial persons resembling the Greek Gods (cf. [McCorduck, 2004, pp. 4]), the novel constructions of the Great Pyramid in Egypt and the Great Wall in China, the discussions in Plato's Ion about the society's need for creative people (cf. [Runco and Pritzker, 1999, pp. 512]), etc.

[2] Guilford is well-known as the pioneer who proposed that intelligence could not be characterized in a single numerical parameter (e.g. IQ), and proposed three necessary dimensions for accurate description of intelligence; namely the operations, content, and products dimensions. Guilford also introduced the operation of "divergent production" (that is, cognition that leads in various directions) in his "Structure Of Intellect (SOI)" model (cf. Guilford [1950, 1967]), which has 180 different kinds of intellectual processes and skills (cf. [Runco and Pritzker, 1999, pp. 577]).

[3] Holding, for example, reported that the excellent chess-player needs to be aware of a general plan of

Due to the undeniable success of these endeavors, the following question can be raised: what is a cognitive ability that makes human cognition unique in comparison to animal cognition on the one hand and artificial cognition on the other? At the beginning of AI most researchers would probably have said "*higher cognitive abilities*" (as may be indicated by the above examples), because only humans are able to reason in abstract domains. In current (classical) AI research, many researchers would, on the contrary, (perhaps) say that all in all still "*lower cognitive abilities*" like performing motor actions in a real-world environment, perceiving natural (context-dependent) scenes, the ability to integrate multi-modal types of sensory input, or the social capabilities of humans are the basis for all cognition as a whole, and therefore also the key features for human-level intelligence. Finally, a researcher interested in aspects of "GI" and (higher-level) cognition would probably stress the combination and integration of both aspects of cognition: a successful model of "AGI" should be able to integrate higher and lower types of cognition in one architecture.[1]

Beside these possibilities, there is nevertheless an important cognitive ability that seems to be usable as a rather clear feature to distinguish human intelligence from all other forms of animal or artificial intelligence: "*creativity*". Although we can ascribe creativity to many human actions, we would hardly (or, at most, seldom) say that a certain animal shows creative behavior or a machine solves a problem creatively. Even in the case of IBM's chess-playing computer (DeepBlue; cf. Hsu [2002a]) and question-answering system (Watson; cf. Ferrucci et al. [2010]) —where the latter is probably the most advanced massive knowledge-based system that exists so far, most people would not ascribe general creative abilities to either. At most we may say that certain particular solution methods of the systems seem to be creative (in the sense of being new or non-classical), because they are extremely hard to achieve for humans.[2]

This chapter conceptually discusses some aspects of creativity, as well as the possibility to explain creativity with cognitive principles and to subsequently model creativity with logic-based means. The underlying main idea is not to model creativity directly with classical logic, but to reduce many forms of creativity to cognitive mechanisms

action for each game, and that verbal memory is an integral part of (blindfold) play (cf. Holding [1985]). The Dutch psychologist de Groot also presents psychological enquiry into the minds of chess-players (cf. de Groot [2008a]).

[1] See for example the discussions in [Hofstadter and the Fluid Analogies Research Group, 1996, Chapter 4], where an argument is given on the inability of models, that separate conceptual processes from perceptual processes, to lead to satisfactory understanding of the human mind.

[2] Looking at the issue philosophically, Searle argues that Watson —regardless of whatever impressive capabilities it really has, cannot truly think. Drawing on his "Chinese Room" thought experiment (cf. Searle [1980]), philosopher Searle argues that Watson does not even know it won on a competition (cf. Searle [2011]).

like analogy-making and concept blending (cf. Chapters 2 and 4). Such mechanisms in turn can be modeled with (non-)classical logical formalisms that mainly depend on a utilization of cross-domain intelligence. In section 5.2, some forms and manifestations of creativity are sketched. Section 5.3 discusses the possibility to describe creative acts by cognitive mechanisms, such as analogy-making and concept blending. It is explained that this cannot only be done for examples of creativity from highly structured domains but for a broad variety of different domains. Section 5.4 revisits basic principles of the logical HDTP framework (cf. section 3.1), and proposes HDTP for analogy-making and concept blending, in order to model creativity. Section 5.5 concludes the chapter's discussions.

5.2 Forms of Creativity

Creativity describes a general cognitive capacity that is in different degrees involved in any process of generating an *invention* or *innovation*. Both concepts, invention and innovation, describe properties of concrete products, services, or ideas. The distinction between them in the following is based on Burki and Cavalluci [2011].

From a more engineering- and business-oriented perspective, an invention is usually considered as the manifestation of the creative mental act, resulting in a new artifact (prototype), a new type of service, a new concept, or even the mental concretization of a conception. An innovation requires standardly the acceptance of the invention by the market, where market is not exclusively restricted to business aspects. In this chapter, creativity is considered as a cognitive ability, but one has to also refer to inventions, innovations, new concepts, new findings, etc., in order to exemplify creativity in a concrete setting.

Creativity appears in various forms and characteristics. Creativity can be found in science, in art, in business processes, and in daily life. Creative acts can occur in highly structured and clearly defined domains (like in mathematics), in less structured domains (like business processes), or even in relatively unstructured domains (like a marketing department of a company having, for instance, the task to design a new advertisement for a certain product).

Different types of creativity are summarized in Table 5.1. Taking into account the various domains in which creativity can occur, it seems hard to specify a domain, in which creativity does not play a role. Rather certain aspects of creativity can appear in nearly all environments and situations. This is one reason why the specification of common properties and features of creativity is a non-trivial task. For example, some attempts

Domain	Areas	Examples for creative acts
Science	Mathematics	Argand's geometric interpretation of complex numbers (cf. Argand [1813]; Martínez et al. [2012])
	Linguistics	Chomsky's recursive analysis of natural language syntax (cf. Chomsky [1957])
	Physics	Einstein's theories of special and general relativity
Art	Music	Invention of twelve-tone music by Arnold Schönberg
	Poetry	The invention of a novel (as a genre of poetry)
	Visual arts	Usage of iconographic and symbolic elements in paintings (Jan van Eyck)
Other	Daily life	Fixing a household problem
	Business	Nested doll principle for product design

Table 5.1: Some domains, areas, and examples of manifestations of creativity are mentioned. Clearly, the table is not considered to give a complete overview of domains in which creative inventions of humans can occur.

have been made to specify certain phases in the creative process (cf. Wallas [1926]). Unfortunately, such phases, as for example a "preparation phase", are quite general and hard to specify in detail. It is doubtful whether any interesting consequences for a computational model can be derived from such properties.

5.3 Creativity and Cognition

Studying creativity for computationally modeling aspects of human-level intelligence was not considered as important and attractive to researchers in AI and cognitive science as it currently is. As aspects of cognition and general intelligence are becoming more and more connected to both disciplines (i.e. AI and cognitive science), it is also becoming more clear that creativity deserves to be considered as a major issue in mainstream research focused on modeling these aspects in AGI. But one must acknowledge that the topic is extremely intense in width (e.g. ways of defining creativity and functioning types of creativity), depth (e.g. previous works done), and intertwining with many other topics, especially in cognitive science.

A cognitive scientific study of creativity may not provide a detailed explanation of all various ways, in which creativity can be approached by interested cognitive scientists. Nonetheless, the general processes to noticeably demonstrate a glimpse of creative behavior in computational models of creativity and general intelligence may well be found in each of these domains. In fact, it is still a big challenge for the interested cognitive scientists and AI researchers to exhaustively define the notion of 'creativity'. During only

the past few years, a multitude of relatively new (interdisciplinary) theories and ideas on creativity were introduced, which multiplied the number of numerous studies that have already been given during the past few decades.[1] Moreover, creativity and intelligence require knowledge, but they can be distinguished from each other, and much the same can be said about creativity and problem solving (cf. [Runco and Pritzker, 1999, pp. 511]).

For many people, there may seem to be an opposition between creativity and logical frameworks. Certain creative insights, inventions, and findings do seem to be creative precisely because the inventor did not apply a deterministic, strictly regimented form of formal reasoning (the prototypical example being classical logical reasoning), but rather because the inventor departed from the strict corset of logic. Therefore, often a natural clash and opposition between logical modeling and creativity seem to typically be perceived. The ideas in this chapter stress that this claim should be rejected. On the contrary, it is advocated here that the natural way to start is to model creativity with logical means (at least in highly structured domains like science, business applications, or classical problem solving tasks). The reason for this is based on the hypothesis that creativity is to a large extent based on certain cognitive mechanisms like analogy-making and concept blending. But now, due to the fact that analogy-making and concept blending essentially employ the cross-domain identification and association of structural commonalities, in turn a natural way to model the mechanisms are logic-based frameworks.

Although creativity seems to be an omnipresent aspect of human cognition (compare Table 5.1), not much is known about its psychological foundation, the neurobiological basis, or the cognitive mechanisms underlying creative acts. One reason might be that examples for creativity cover rather diverse domains, where completely different mechanisms could play important roles. Nevertheless, the next discussions hypothesize that, in many interesting cases, classical examples for creativity can be reduced to two important cognitive mechanisms, namely analogy-making on the one hand and concept blending on the other. Some examples are mentioned first in order to make this hypothesis more plausible:

1. Conceptually, the usage of analogy-making is rather clear in cases where one is using a general principle in a new domain. For instance, consider the use of the nested doll principle in design processes (cf. Figure 5.1), in which creativity can

[1] Such theories and ideas are not needed here, with the exception of scratching the surface of Colton et al.'s computational creativity theory in section 5.5 (cf. Colton et al. [2011]; Pease and Colton [2011]). The interested reader may still refer to McCormack and d'Inverno [2012]; Runco and Pritzker [1999] or the list of references given in [Boden, 1996, Chapter 9] for more about such theories. There is a somewhat longer, related discussion in Abdel-Fattah and Schneider [2013].

Figure 5.1: Two design examples (one from the engineering domain and one from product design) that are based on the same principle, namely the nested doll principle: Objects are contained in similar other objects in order to satisfy certain constraints. The left image is taken from http://en.wikipedia.org/wiki/Planetary_gears which is licensed under the terms of the GNU Free Documentation License. The image to the right is a set of "Pyrex Nesting Bowls", found at http://www.whitetrashnyc.com/products/set-of-pyrex-nesting-bowls/1780.

be considered as a transfer of a structure from one domain (e.g. the structure of a planetary gearing, namely gears that revolve about a central gear) to another domain (e.g. the design of nesting bowls containing each other). This transfer of structural properties is best described as an analogy.

2. In science, analogies and blend spaces do appear quite regularly. For example, in Guhe et al. [2011] it is shown how analogies can be used to learn a rudimentary number concept and how concept blending can be used to compute new mathematical structures. Furthermore, in Martínez et al. [2011, 2012] it is shown that concept blending can lead to a geometric interpretation of complex numbers, inspired by the historically important findings of Argand mentioned in Table 5.1.

3. Also, the interpretation of certain visual inputs can easily be described by analogy-making (visual metaphor, as shown in Schwering et al. [2009c]). Figure 5.2 gives an example, depicting an advertisement. In order to understand this advertisement a mapping between tongue and fish as well as a transfer of properties of fish need to be performed.

The number of examples, which show that analogy-making and concept blending can be used to explain manifestations of creativity, are numerous. If it is true that several characteristics of creativity can be modeled by analogies and concept blending, a computational approach towards creativity can naturally be based on an algorithmic theory of analogy and concept blending. Due to the fact that analogy-making is the

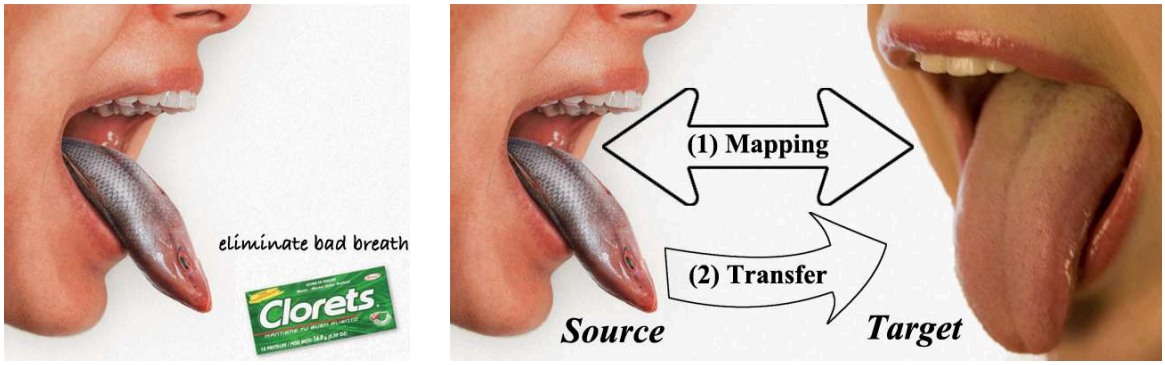


Figure 5.2: A creative advertisement on the left side depicts an association between a tongue and a fish. In order to understand this advertisement (as a marketing tool for the widely-known "*Eliminate bad breath*" campaign of Clorets®' hard candy) the establishment of a mapping between tongue and fish is necessary. Then, hard candy can be used as a means against breadth odor. In Schwering et al. [2009c], a formal modeling of a similar advertisement is specified.

identification of structural commonalities and concept blending is the (partial) merger of cross-domain structures, the natural way for an algorithmic approach is to use logic as the methodological basis. Whereas for concept blending, a symbolic approach for modeling may be quite undisputed (but cf. sections 4.2 and 4.4), the situation in analogy-making is more complicated. Concerning the modeling of analogies, also several neurally inspired and hybrid models have been proposed. Nevertheless, when having a closer look, it turns out that the most important subsymbolic aspects of such models are activation spreading properties or synchronization issues in a (localist) network, whereas the basic computational units of the network still are quite often symbolic entities (cf. section 2.2.2).[1] Additionally, logic-based models of analogy-making have a wider application domain in comparison to neurally inspired or hybrid models. Therefore, in total, it seems a natural choice to apply logical means in modeling these two cognitive mechanisms.

5.4 Towards a Logic-Based Framework

A Tool for Making Analogies (HDTP): In what follows, *Heuristic-Driven Theory Projection*, HDTP (cf. Schwering et al. [2009a]), is used as the underlying modeling framework. HDTP is already presented and extensively discussed in Chapter 3, but the current chapter discusses HDTP's potential to be employed as a powerful analogy-making en-

[1] For example, cf. Hummel and Holyoak [2003] and Kokinov and Petrov [2001] for two of the best known neurally inspired analogy models.

gine in modeling facets of creativity.

To recall, HDTP is a mathematically sound framework for analogy making, together with the corresponding implementation of an analogy engine for computing analogical relations between two logical theories, representing two domains. Domains are represented in HDTP as sets of axioms formulated in a many-sorted, first-order logic language. HDTP also provides an explicit generalization of these domains as a by-product of establishing an analogy, where a generalization can be a base for concept creation by abstraction. HDTP applies restricted higher-order anti-unification (cf. Krumnack et al. [2007] and section 3.1.1) to find generalizations of formulae and to subsequently propose analogical relations between source and target domains. Proposed analogical relations can later be used as a basis for an analogy-based transfer of knowledge between the two domains.

Figure 3.4 (cf. page 61) depicts HDTP's overall approach to creating analogies, in which analogical transfer results in structure enrichment of the target side. There are application cases in which two conceptual spaces (here, the input source and target theories) need not to be (partially) mapped onto each other, but rather partially merged in a new conceptual space. In such cases, HDTP uses the computed generalization, the given source and target theories, and the analogical relation between source and target in order to compute a newly constructed conceptual space (cf. explanations in the next paragraph).

The Other Facet of the Same Tool (CB): The integration of concepts by utilizing the network model of CB (cf. section 4.3) has been proposed as a powerful mechanism that facilitates the creation of new concepts by a constrained integration of available knowledge. CB also operates on conceptual domains by merging (at least two) input knowledge domains to form new domains, which crucially depend on (and are constrained by) structural commonalities between the original input domains. New domains are the blends, with each blend candidate maintaining partial structure from both input domains and presumably adding emergent structures of its own. Figure 5.3 is a summarized reproduction of the ornamented illustration given in Figure 4.3, of the prototypical conceptual integration network model (cf. page 86). Notwithstanding, Figure 5.3 is based on the way, explained below, in which the HDTP framework functions toward creatively constructing concept blends from two given inputs.

To constructively help in creating new concepts, the HDTP framework's view of blending is proposed to function as follows (cf. Figure 5.3):

1. Two concepts, such as *Source* and *Target* (denoted by S and T, respectively, in

Figure 5.3), can be used to represent two input spaces (the mental spaces).

2. Common parts of the input spaces are matched by identifying their structural commonalities using HDTP. The matched parts of S and T may be seen as constituting a generic space, or the *Generalization* (denoted by G in Figure 5.3). Such a generalization G is computed by anti-unifying S and T.

3. As previously illustrated (in Figure 3.4), an analogical relation (denoted in Figure 5.3 by m), which can be found by HDTP between S and T, facilitates the transfer of some knowledge from S to T based on the generalization G.

4. A blend space candidate, *Blend* (denoted by B in in Figure 5.3), has an emergent structure that arises from the blending process that imports additional knowledge from the input concepts. Based on the generalization G, a blend candidate, B, consists of some matched and possibly some of the unmatched parts of the input spaces. The dashed arrows in Figure 5.3 from S to B and from T to B reflect the importation of new (non-conflicting) conceptual entities form source and target to the blend candidate B.

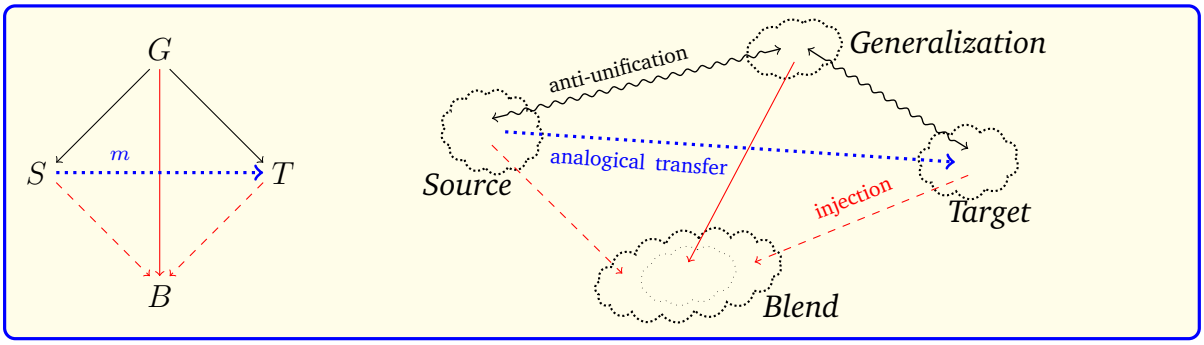


Figure 5.3: HDTP's view of concept blending: S and T represent source and target input theories, where their identified common parts can define generalizations (i.e. generic spaces) and blend candidates. A generalization G is computed by anti-unifying S and T, with m representing the analogical relation between S and T (cf. also Figure 3.4). The dashed arrows $S \to B$ and $T \to B$ describe the injections of facts and rules from source and target to the blend candidate B. Due to the fact that the input theories may contain inconsistent information, the injections are partial in general.

As discussed in Chapter 4, conceptual blending has already shown its importance as a substantial part of cognition and a means of constructing new conceptions. It has been extensively used in the literature in attempts at expressing and explaining cognitive phenomena, such as the invention of new concepts, the meaning of natural language metaphors, as well as its usefulness in expansion, reorganization, and creation

of mathematical thoughts and theories (Abdel-Fattah et al. [2012a]; Alexander [2011]; Fauconnier and Turner [2002]; Goguen [2006]; Guhe et al. [2011]). Nevertheless, CB itself noticeably still suffers from the lack of a formally precise model integrating its many aspects (cf. section 4.4).

5.4.1 Roles of Analogy and Blending in Modeling Creativity

What makes the process of cross-domain reasoning special to creativity is the "interesting characteristic" of analogies that "by seeming to move away from a problem the subject can actually come closer to a solution" [Clement, 2008, pp. 61]. An argument would start by emphasizing that "it is impossible for people to create new ideas out of the air", but rather discover or create new ideas based on their "knowledge, experience, or expertise working with problems in a given domain", because available evidence indicates that multifaceted, principle-based knowledge structures certainly "represent a necessary foundation for creative achievement" [Runco and Pritzker, 1999, pp. 72–73]. Clement emphasizes the usefulness of analogical reasoning, indicating that people infer knowledge from an analogy that could be useful in at least 3 possible ways (cf. [Clement, 2008, pp. 61]):

1. in predicting an answer to a specific problem,
2. in providing a suggested method of attack, and
3. in providing a principle that applies to the target.

But exactly how people do use extant knowledge to create something new takes us to the role of blending in creative thinking. In various domains, many interested scholars have been perplexed over time in proposing answers to the latter question (cf. [Boden, 2003; Clement, 2008; Finke et al., 1992; Guilford, 1950; McCormack and d'Inverno, 2012; Newell et al., 1963; Sternberg and Lubart, 1999; Wallas, 1926; Weisberg, 1993, to mention a few]). Their proposals range from the notion of "*divergent thinking* to the idea that creative thought might be based on remote *associations*" [Runco and Pritzker, 1999, pp. 73; emphasis added], which succeed in capturing some truth about certain fundamental aspects of the creative process. According to this chapter's view, associations may generally be achieved through cross-domain, analogical reasoning. Conceptual blending also captures creative thoughts based on remote associations, and achieves divergent thinking (which does play a particular role in idea generation). Consequently, it seems plausible to claim that creative thoughts should mostly involve structured knowl-

edge entities (based on experience), analogical reasoning, and blending of conceptual entities. However, it seems also hard to provide a grounding proof in the meantime.

As can be entailed from the above discussions, the ideas of CB are very much related to the properties of a creative process, since a creative process can result in new insights as a result of a ladder-ascending procedure that steps through "background knowledge", and subsequently increasingly refines the insights to spell-out an innovation (cf. section 5.2). Undoubtedly, in realizing a computational modeling of cognitively inspired, creative agents, it must also be the case that the agents have (enough) background knowledge before a creative process can take place. Mere knowledge most likely is still not sufficient, and even analogy-making alone helps us only to build up new knowledge by starting from old knowledge and make the unfamiliar familiar. But, for instance, simply having knowledge about Maxwell's equations, the principles of semi-conductors, and the principles of graph theory almost surely by itself is not enough in order to devise the ideas of very-large-scale integration (i.e. the creation of integrated circuits by combining thousands of transistors into one single chip), unless an additional cognitive process is utilized. The claim in this chapter is that here is exactly where multifaceted cognitive mechanisms, namely conceptual blending (in addition to analogy), come into play.

Roles of the Framework: Now, as for HDTP, it provides a potential framework for a CB-based computation of novel concepts, given a source and target domain axiomatizations:

1. Assume two input theories S and T are given.
2. The computation of an analogical relation between S and T by HDTP outputs (besides other things) a shared generalization G of S and T by the anti-unification process. This generalized theory G functions in the further process as the generic space in CB mentioned above.
3. The construction of the blend space is computed by
   (a) first, collecting the associated facts and rules from S and T generated by the analogical relation between S and T, and
   (b) second, by projecting unmatched facts and rules from both domains into the blend space.
4. The latter (second) step can result in clashes and inconsistencies. Furthermore, the coverage of the blend space concerning S and T can be more or less maximal.

Taking additionally into account that for every given S and T HDTP can compute different analogical relations, there can be many possible blend spaces for a given input.

It is worth mentioning that HDTP has successfully been used to compute concept blends in complex domains like mathematics. For example, Guhe et al. use HDTP to model Lakoff and Núñez's mathematical grounding metaphors, which are intended to explain how children can learn a rudimentary concept of numbers based on simple real-world actions in their environment (cf. Lakoff and Núñez [2000]). The mathematical grounding metaphors and the emergence of an abstract number concept can be explained by analogy-making and concept blending (cf. Guhe et al. [2011]). In addition, the invention of a geometrical interpretation of complex numbers (i.e. the complex plane) was computationally modeled by concept blending in Martínez et al. [2011] and Martínez et al. [2012]. This example shows that even for rather formal and complex theories the creative generation of a new concept can be computed using a logic-based approach. Other examples related to creative construction of concepts based on the HDTP framework are explained later (e.g., cf. Chapter 7).

5.5 Conclusive Remarks and Related Ideas

A particular issue for AGI systems is that creative problem solving abilities and the finding of novel solutions in unknown situations need to be more crucially considered in current systems than ever before. The idea of creativity is being delivered here as a yet more foundational step towards building a general form of AI. From a cognitive perspective, this chapter stresses that creativity can often be reduced to multifaceted cognitive mechanisms such as analogy-making and concept blending, which in turn can neatly be modeled using logic-based approaches. The apparent tension between creative abilities of agents and a logical basis for their modeling, therefore, disappears.

A Related Theory of (Computational) Creativity: To a certain extent, the chapter's view of creativity is compatible with recent proposals for computational creativity. In particular, one of the newest models for computational creativity is Colton et al.'s FACE model (cf. Colton et al. [2011]; Pease and Colton [2011]), which employs four "aspects" to capture and describe eight kinds of "generative acts" in artificial, computational systems that can be considered creative. According to the FACE model, each of its four "aspects" can be considered at two "levels": the ground and the process levels (g and p, respectively). A "generative act" can thus occur at either levels, and a creative act is

defined as a non-empty tuple of generative acts. Newly produced artifacts (e.g. poems or compositions) by creative systems that are based on the FACE model need to invent ways (i.e. processes) to generate and assess these artifacts (cf. [Colton et al., 2011, pp. 91]). The descriptive model's four aspects (in a strongly simplified sense) are: (i) "concept"; (ii) "expression" (of a concept); (iii) "aesthetic measure"; and (iv) "framing information". With a somewhat liberal interpretation of these aspects, the HDTP approach can be subsumed by the FACE descriptive model. That is, the FACE model can be instantiated with HDTP in the following way:

Concept: An executable program (e.g. an algorithmic realization of analogy-making and concept blending).

Expression (of a Concept): An instance of the input/output when the program is executed (e.g. pairs consisting of input theories and analogical relations or blend spaces).

Aesthetic Measure: An evaluation function that takes as input a concept and an expression and outputs a real number (e.g. ranking heuristics of potential candidates for analogical relations and conceptual blends; cf. section 3.1.3).

Framing Information: A contextual embedding (e.g. the retrieval problem in analogy-making and concept blending).

As a further support to the above claim, it should be noted that the ideas given in this chapter are not the first ones to investigate the computational modeling of creativity as a cognitive capacity. The chapter can be seen to stress the idea and propose a way to achieve such a modeling in logic-based frameworks. Going back already to work by Newell et al. (cf. Newell et al. [1963]), researchers in AI and related fields over the decades repeatedly have addressed different issues and aspects of creative thought. The results of these investigations range from contributions on the more conceptual side (as, for instance, Boden's theory of P- and H-creativity; cf. Boden [2003]), to concrete implementations of allegedly "creative systems" (as, for instance, "The Painting Fool"; cf. Colton [2011]). And also in the computational analogy-making domain, there already is relevant work on the relation between creativity and analogy, most prominently exemplified by Hofstadter's contributions related to the Copycat system (cf. Hofstadter [1984]). Still, on the one hand, work on issues of creativity within human-style intelligent systems this far has not gained wide attention in an AGI context. On the other hand, even within the more general setting of computational creativity research, only

very few approaches try to integrate models of different cognitive capacities into a system aiming for general creativity capacities, instead of limiting the focus to modeling one specific kind of creative act or another.

The chapter sketches a necessity to tackle the hard problem of creativity in AGI systems. Although the described HDTP framework has been applied to show that the computation of interesting blend spaces can be achieved in certain rather complex (but highly specific) domains, no generalizations of such specific examples exist so far. This remains a task for future work (besides a further formally sound and complete characterization of concept blending on a syntactic and semantic level; cf. section 4.4).

6

Rationality-Guided Aspects of General Intelligence

For more than five decades, AI has always been a promising field of research on modeling aspects of human-like intelligence. The recent success of projects like IBM's Watson (cf. Ferrucci et al. [2010]), for instance, increases the hopes in achieving not only language intelligence but also inference mechanisms at a human level and paves the way for solving more baffling tasks. However, AI has turned into a vague, unspecific term, in particular because of the tremendous number of applications that belong, in fact, to seemingly orthogonal directions. Philosophers, psychologists, anthropologists, computer scientists, linguists or even science fiction writers have disparate ideas as to what AI is (or should be). The challenge becomes more obvious when AI is looked at from a cognitive scientific perspective, where the focus is mainly on explaining processes of general cognitive mechanisms, not only on how one or another intelligence task can be solved by a computer.

The AGI research paradigm takes AI back to its original goals of confronting the more difficult issues of human-level intelligence as a whole (cf. Chapter 1). It is therefore becoming more necessary to give AGI a more prominent place within cognitive science, by elaborating on several indispensable cognitive criteria, as well as modeling them. Chapter 5 contributes to this issue by discussing the roles of cognitive mechanisms when considering "creativity" as one of such indispensable criteria of GI, and this chapter approaches cognition in AGI systems by particularly promoting "rationality" as an equivalently important, indispensable criterion. In particular, the current chapter: (i) focusses on some divergent, sometimes seemingly irrational, behaviors of humans, (ii) analyzes such behaviors, and (iii) proposes the utilization of cognitive mechanisms to overcome some of their challenges in modeling AGI systems. The text allocates ideas

from AGI within cognitive science, then gives a conceptual account on some principles in normative rationality-guided approaches. Arguments are given to constitute outward evidence that normative models of human-like rationality are vital in AGI systems, where the treatment of deviations from traditional rationality models is also necessary. After explaining one suggested approach at a general level, the chapter explains how two cognitively inspired systems, namely NARS and HDTP, have the potential to handle (ir)rationality. The chapter concludes by some remarks and future speculations.

6.1 AGI and Rationality

Why AGI? It is clarified in section 1.2 that AGI has currently a stronger relation to cognitive science than what conventional AI had in the past. From a cognitive scientific perspective, the kind of intelligence characterizing classical AI problems is not yet exhaustive enough, where solutions to most of the problems are not cognitively inspired: neither do they consider essential cognitive mechanisms (or general intelligence results) nor do they show the biological plausibility of the solutions. Current AGI research explores all available paths, including theoretical and experimental computer science, cognitive science, neuroscience, and innovative interdisciplinary methodologies (cf. Baum et al. [2010]). On the cognitive science side, AGI offers its adherents the possibility to think in less anthropocentric terms, so as to better treat "intelligence" and "cognition" as general notions that are not limited to individual human beings, but instead can be abstracted into a general systems theory. AGI treats intelligence as a general-purpose ability, and takes a holistic attitude towards intelligent systems. In current AGI research, there are approaches following different paths, including those

- inspired by the structure of human brain or the behavior of human mind,
- driven by practical demands in problem solving, or
- guided by *rational principles* in information processing.

The latter approach is of a special interest, because it has at least three essential advantages:

1. One advantage of the rationality-guided approach from an AGI perspective is that it is less bound to exactly reproducing human faculties on a functional level.
2. Another advantage of such an approach on a scientific meta-level is that it gives AI the possibility of being established in a way similar to other disciplines, where it

can give a theoretical explanation to intelligence as a process that can be realized both in biological systems and computational devices.

3. The third advantage of the rationality-guided approach is that it is not limited to a specific domain or problem.

Other related features of AGI (cf. [Baum et al., 2010; Goertzel and Pennachin, 2010, for instance]) further support the claim that AGI, in general, ties into cognitive science more closely than mainstream AI.

What is Rationality? The term *rationality* is used in a variety of ways in various disciplines. In cognitive science, rationality usually refers to a way a cognitive agent deliberatively (and attentively) behaves in, according to a specific normative theory. It is discussed in many contexts, such as problem-solving systems, where such systems obtain knowledge and problems from their environment, and solve problems according to this knowledge. The prototypical instance of cognitive agents that can show rational behavior is humans, who so far are also the ultimate exemplar of generally intelligent agents.

Surprisingly, little attention has been paid so far in AI towards a theory of human-comparable rationality (as an important GI aspect). A reason might be that the concept of rationality was too broad in order to be of interest to AI, where for a long time usually relatively specific cognitive abilities were modeled and heuristics were suggested. Moreover, an artificial cognitive agent is usually intended to reproduce rational behavior, not to act in seemingly irrational ways. Consequently, even generally interested AI researchers are not particularly interested in well-known results of some *classical rationality puzzles*. Still, a move towards integrating AGI in cognitive science cannot ignore rationality issues; neither the remarkable abilities nor the originalities that human subjects show in rationality tasks.

When modeling GI aspects, it is reasonable to initially take the remarkable abilities of humans into account with respect to rational behavior, but also their apparent deficiencies that show up in certain tasks. Two main challenges of the rationality-guided AGI research can immediately be seen. First, given the richness and complexity of the human mind, it is extremely challenging to find a *small* number of clearly specified principles and laws to explain *all* the relevant phenomena. Second, common particularities in human behavior need to be explained and justified for human thinking to really be taken as mostly rational —that is, the well-known "irrational" behaviors. Two possible answers to this can also be seen immediately:

1. one may confess that intelligent systems are often irrational, since they fail to follow classical normative theories, or

2. one may argue that intelligent systems are rational, though this rationality has not yet been summarized into traditional normative theories.

The second position is going to be advocated, believing that intelligent systems (like humans) are rational, and this rationality can be summarized into a positive, and finally even normative theory, though the traditional theories failed (and new principles are needed).

6.2 Traditional Models of Rationality

Different models of rationality use significantly different methodologies. Clustering such models according to the underlying formalism usually results in at least the following classic groups:

1. logic-based models, for which a belief is considered rational as long as it has been derived through a logically valid reasoning process, given the background knowledge (cf. Evans [2002]),

2. probability-based models, for which a belief is considered rational if its expectation value is maximized with respect to given probability distributions of other beliefs in the background knowledge (cf. Griffiths et al. [2008]),

3. game theory-based models, for which a belief is considered rational if the expected payoff of maintaining it is maximized relative to other possible beliefs (cf. Osborne and Rubinstein [1994]), and

4. heuristic-based models, which utilize the use of heuristics to give accounts on rationality (cf. Gigerenzer [2010]).

Several of these models have been proposed for establishing a *normative theory of rationality*, normally by judging a belief as rational if it has been obtained by a formally correct application of the respective reasoning mechanism, given some background knowledge (cf. e.g. Gust et al. [2011]; Wang [2011]). Therefore, such theories of rationality are not only intended to model "rational behavior" of humans, but to postdictively decide whether a particular belief, action, or behavior is rational or not. Nonetheless, although a conceptual clarification of rational belief and rational behavior is without

any doubts desirable, it is strongly questionable whether the large number of different (and quite often orthogonal) frameworks makes this task easier, or if the creation of a more unified approach would not be recommendable. From this thesis' perspective, multifaceted cognitive mechanisms seem to offer a basis for such an endeavor. This is explained later in this chapter, after some challenges are mentioned.

6.2.1 Some Rationality Challenges and Puzzles

Although the models mentioned above have been proven to be quite successful in modeling certain aspects of intelligence, all four types of models have been challenged.

1. In the famous Wason selection task (cf. Wason and Shapiro [1971]) human subjects fail at a seemingly simple logical task (cf. Table 6.1a).

2. Similarly, Tversky and Kahneman's Linda problem (cf. Tversky and Kahneman [1983]) illustrates a striking violation of the rules of probability theory in a seemingly simple reasoning problem (cf. Table 6.1b).

3. Heuristic approaches to judgment and reasoning try to stay closer to the observed behavior and its deviation from rational standards (cf. Gigerenzer [2010]), but they fail in having the formal transparency and clarity of logic-based or probability-based frameworks with regard to giving a rational explanation of behavior.

4. Game-based frameworks can be questioned due to the various forms of optimality concepts in game-theory that can support different "rational behaviors" for one and the same situation.

In order to make such challenges of rationality theories more precise, some aspects of the famous Wason selection task and the Linda problem are discussed in more detail.

Wason Selection Task: This task shows that a large majority of participants are seemingly unable to evaluate the truth of a simple logical rule of the form *"if* p *then* q*"* (cf. Wason and Shapiro [1971]). In the version of the task, presented in Table 6.1a, this rule is represented by the sentence (Wason-Cards):

> "If on one side of the card there is a D,
> then on the other there is the number 3".
>
> (Wason-Cards)

Every card which has a D on one side has a 3 on the other side (and knowledge that each card has a letter on one side and a number on the other side), together with four cards showing respectively D, K, 3, 7, hardly any individuals make the correct choice of cards to turn over (D and 7) in order to determine the truth of the sentence. This problem is called “selection task” and the conditional sentence is called “the rule”.

(a) Wason Selection Task (cf. Wason and Shapiro [1971]):

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. After they are given background information, the task for the participants is to rank statements about Linda according to their probability. The particular statements are: (Feminist): Linda is active in the feminist movement. (Teller): Linda is a bank teller. (Teller&Feminist): Linda is a bank teller and is active in the feminist movement. More than four fifth of the participants ranked the conjunctive statement (Teller&-Feminist) as more probable than the statement (Teller) or (Feminist).

(b) Linda Problem (cf. Tversky and Kahneman [1983]):

Table 6.1: (a) A short description of the Wason selection task. (b) An abbreviated version of the Linda problem setting.

In order to verify or to falsify this rule by assigning a truth-value to it, participants need to turn D and 7. That is, according to classical logic, participants need to check the direct rule application and the contrapositive implication (*modus tollens* of the rule). This is not, however, what the large majority of participants suggest when evaluating the truth of the rule (cf. Wason and Shapiro [1971]). What is interesting in this regard is the fact that a slight modification of the content of the rule to a setting more familiar from daily life, while keeping the structure of the problem isomorphic, makes participants perform significantly better (as e.g. shown in Cosmides and Tooby [1993]).

Linda Problem: Regarding the Linda problem, it seems to be the case that human subjects have problems to prevent the so-called *conjunction fallacy* (cf. Tversky and Kahneman [1983]). Subjects are told a story specifying a particular profile about someone called Linda. Then, some statements about Linda are shown and participants are asked to order them according to their probability (cf. Table 6.1b). About 85% of participants

decide to rank the statement (Teller&Feminist) as more probable than the statement (Teller) or (Feminist):

"Linda is a bank teller", (Teller)

"Linda is active in the feminist movement", (Feminist)

"Linda is a bank teller and is active in the feminist movement". (Teller&Feminist)

This ranking conflicts with the laws of probability theory, because the probability of two events (Teller&Feminist) is less than or at most equal to the probability of one of the events (e.g. (Teller)).

6.2.2 Classical Resolution Strategies of Irrationality

Many strategies have been proposed to address the mentioned challenges, ranging from the use of non-classical logics to the modeling of participants' behavior in the Wason selection task (cf. Stenning and van Lambalgen [2008]), to considerations involving reasoning in semantic models instead of (syntactic) deductions (cf. Johnson-Laird [1988]) in the case of the Wason selection task. With respect to the Linda problem, it has been argued that pure probability theory is not appropriate for addressing the problem properly. Busemeyer et al., for example, give a recent explanation for types of probability judgement errors, such as the conjunction fallacy encountered in the Linda problem,[1] but a foundation of the analysis of this problem in coherence theories would be necessary (cf. Pfeifer [2008]).

Another resolution strategy applicable to both puzzles is to question whether tasks were appropriately phrased in the respective experiments. In the Wason selection task the "if-then" rule presented in natural language is usually not equivalent to its interpretation in classical logic, and in the Linda puzzle the term "probable" can be interpreted differently by the participants (cf. Gigerenzer [2005]).

In any case, and although there are many proposals to address the challenges, there is no generally accepted rationality concept available yet. Moreover, specific frameworks

[1] Busemeyer et al.'s work is based on a model of "quantum cognition" (cf. Busemeyer and Bruza [2012]; Rijsbergen [2004]). Although this thesis is not concerned with approaching such type of quantum-based models at any level, it is worth mentioning my opinion that these models seem very promising in modeling human-like cognition, because they have a great potential to provide formal, yet relatively complicated, explanations of many cognitive phenomena. Some are already related to this thesis' accounts on aspects of general intelligence and cognition, such as interpreting novel concept combinations presented in Chapter 7 (cf. Aerts and Gabora [2005a,b], for instance).

can address specific challenges, but do not generalize to the breadth of the mentioned problems.

For a generally intelligent cognitive system, a question that can be raised is: *which principles of rationality can be transferred to (and modeled in) AGI systems, in order to achieve intelligence on a human-like scale?* The discussions in this chapter argue for models that link rationality to the ability of humans to establish analogical relations, and to the ability to adapt to the environment by making good use of previously obtained experiences.

6.3 Non-Standard, CogSci-Based Approaches

The two examples discussed above definitely show that humans have sometimes problems to apply rules of classical logic correctly (at least in rather abstract and artificial situations), and to reason according to the Kolmogorov axioms of probability theory. Nonetheless, the most that can be concluded from the experiments is that human agents are neither classical deduction machines nor probability estimators, but rather perform their indisputable reasoning capabilities by other means, necessarily linked to their cognitive capacities.

6.3.1 Resolving the Selection Task by Cognitive Mechanisms

As mentioned above, participants perform better in the Wason selection task, if content change makes the task easier to access for participants. (Here, "better performance" of participants is interpreted in the sense of "more according to the laws of classical logic".) The conjecture is that the performance of participants seems to have a lot to do with their ability to establish appropriate analogies between already-experienced situations and newly encountered ones. Participants are assumed to perform badly in the classical version of the Wason selection task, because they perhaps fail to establish a correct analogy with their experiences.[1] Therefore, participants fall back to other (less reliable) strategies to solve the problem. In a content-change version of the task the situation is different, because participants can do what they would do in an everyday analogous situation.

Recall that evaluating an analogy as being good depends on whether it involves (i) mappings of relations (not of only mere attributes), and (ii) mappings of coherent sys-

[1] Unless their reasoning indeed follows classical logic, so they can mentally represent the whole situation in terms of a logical implication, then apply modus tollens.

tems of relations (not of only individual relations; cf. section 2.1.3). The content-change version of the Wason problem provides the participants with means by which they can establish a good analogy. Whereas, in the original version of the task, participants are not able to establish as good an analogy between their experience (as a source domain) and the current situation (as a target domain). In addition, and as indicated by Gentner and Forbus, for instance, relations (e.g. causal relations) often serve as higher order relations in analogical processing (cf. Gentner and Forbus [2011]). This seems to allow the reasoners (i.e. the participants) to project parts of their already-known relations (of the source situation) to complete missing parts of corresponding, newly constructed relations (within the target situation):

> "When the antecedents of a causal relation are matched, the consequent is projected to hold in the new (target) situation (prediction); and if instead the consequents are matched, the antecedents are projected to hold in the new situation (explanation or abduction)" (cf. [Gentner and Forbus, 2011, pp. 266]).

In short, cognitive scientific studies further support the claim that the success or failure of managing the task is crucially dependent on the possibility to establish a meaningful analogy.

Another related resolution is to study the *mode of the inference* that should underly a normative theory of rationality. As one of the case studies presented later in this chapter suggests (cf. section 6.4), one way is to follow Wang's accounts on having intelligent reasoning based on insufficient knowledge and resources (cf. Wang [2011]). Wang's form of reasoning utilizes a logical formalism that is claimed in Wang [2004] to avoid some commonly encountered problems when explaining or reproducing cognition using predicate logics.[1] Based on Wang's ideas (and using his terms[2]), when a system has sufficient knowledge and resources (with respect to the problems to be solved), an axiomatic logic (such as classical logic) can be used, which treats the available knowledge as axioms, and derives theorems from them to solve a given problem. When the system has insufficient knowledge, however, it has no absolute truth to be used as axioms, so it has to follow some "non-axiomatic logic" (cf. Wang [2006, 2013]), whose premises and conclusions are all revisable by new "evidence" (cf. Wang [2009]). In Wason's task, the

[1] Wang also gives an implemented reasoning system, called "Non-Axiomatic Reasoning System (NARS)", that can be considered as a cognitive approach towards reasoning (cf. Wang [2004, 2013] and section 6.4.1 below).

[2] An article on which this chapter is based was co-authored with Wang (cf. Abdel-Fattah et al. [2012a]). All descriptions of Wang's ideas and the NARS system (in this chapter) are very minor modifications of his own text contributions.

expected results are the ones assuming an axiomatic system, while the actual results may be consistent with a non-axiomatic one. Therefore, the "mistake" here seems to be mainly the misunderstanding between the cognitive scientists who run the tests and the human participants who take the tests. In such an artificially structured experiment, it is valid for the scientists to assume sufficient knowledge and resources, therefore to expect the application of an axiomatic type of inference mechanism. The mistake, however, is the failure to see the result as possibly coming from another type of inference. On the side of participants, since non-axiomatic reasoning is used more often in everyday life, most of them fail to understand the experiment setting as a testing of their capacity of using an axiomatic inference mechanism. This explains why many participants admit their mistake afterwards, and do better in the content-change task (as soon as they realized that the expected way of reasoning is not their default one, they have less problem to adapt to follow it).

6.3.2 Resolving the Linda Problem by Cognitive Mechanisms

In the case of the Linda problem, a natural explanation of participants' behavior is that there is a "lower degree of coherence" of Linda's profile plus the statement (Teller) in comparison to the degree of coherence of Linda's profile plus the statement (Teller&-Feminist):

"Linda is a bank teller", (Teller)

"Linda is a bank teller and is active in the feminist movement". (Teller&Feminist)

In the conjunctive statement, (Teller&Feminist), at least one conjunct of the statement fits quite well to Linda's profile.

Coherence is a complicated concept (cf. Thagard [2002]) that may need to be discussed in more detail (as does its connection to notions like the idea of representativeness proposed as an explanation for the Linda problem by Tversky and Kahneman themselves). Here, however, it can only be mentioned that coherence is essential for the successful establishment of an analogical relation, as well as for guiding adaptation of obtained knowledge and experiences. In order to make sense out of the task, participants tend to rate statements with a higher probability where facts are arranged in a theory with a higher degree of coherence. Also, this can be thought of as a form of coherently adapting beliefs, which also depends heavily on participants' experiences rather than on their knowledge of Kolmogorov axioms of probability theory.

6.4 Modeling Rationality: Case Studies

Formal and computational models in cognitive science can be roughly divided into two major types:

Descriptive: A descriptive model explains how a system actually works, and its establishment is based on empirical data. A descriptive model's quality is evaluated according to its behavior's *similarity* to that of humans.

Normative: A normative model, on the other hand, specifies how a system should work, and its establishment is based on certain general principles or postulates. Such a normative model's quality is evaluated according to its behavior's *coherence* with these basic assumptions.

Though the two types of models are closely related, they are still built and evaluated differently (cf. Wang [2011]).

When building a model of rationality, a central issue is the selection of the assumptions on which the model is based, since all conclusions about the model are derived from, and justified against, these assumptions. There is a need for focusing on the conceptual analysis of such assumptions, as well as their implications in the model of intelligence, as a form of rationality.

In the following, two examples for cognitively inspired systems are given —namely NARS and HDTP. Both systems stand in a certain tradition to classical cognitive architectures like the well-known models ACT-R and SOAR (cf. Anderson and Lebiere [1998]; Laird et al. [1987] and section 1.2.3), because they attempt to model cognition in breadth and not relative to highly specialized abilities. Nevertheless, and because NARS and HDTP stand in a tradition of modeling the competence aspect of general intelligence, they attempt to integrate a handful of different human-inspired reasoning abilities, and try to integrate these abilities in uniform models. Both systems also differ significantly from the mentioned classical cognitive architectures. In the following, a discussion is given on how these systems can account for "irrational" behaviors in tasks, such as the "Selection Task" and the "Linda Problem". The basically needed details about NARS are briefly introduced first (whereas details about HDTP are already presented in Chapter 3).

6.4.1 NARS: GI with Relative Rationality

The Non-Axiomatic Reasoning System, NARS, is an AGI system designed under the assumption that the system usually has insufficient knowledge and resources with respect

to the problems to be solved, and must adapt to its environment. Therefore, the system realizes a "relative rationality". That is, the solutions are the best the system can get under the current knowledge/resource restriction (cf. Wang [2011]). The system is comprehensively described in Wang [2006, 2013][1], but only the treatments of the "Selection Task" and "Conjunction Fallacy" in NARS are briefly explained here.

NARS assumes insufficient knowledge and resources (cf. Wang [2011]). Beliefs in the NARS model are, therefore, not representing "absolute truths" but rather summarizing the system's experience. In particular, the truth-value of a statement measures its *evidential support*. This evidence can be either *positive* or *negative*, depending on whether the evidence agrees with the statement. In concrete terms, consider the statement (Wason-Cards):

> "If on one side of the card there is a D,
>
> then on the other there is the number 3".
>
> (Wason-Cards)

For this statement, one would have the following types of evidence:

1. The D on the card always provides evidence, which is positive if the other side is 3, and negative otherwise.
2. The 3 on the card may provide positive evidence if the other side is D.
3. The 7 on the card may provide negative evidence if the other side is D.
4. The K card provides no evidence.

To determine the truth-value of the (Wason-Cards) statement, all cards except K, should be checked. But due to insufficient resources, the system may fail to recognize all evidence. In this case, D is the easiest, while 7 the hardest. This result is consistent with the common responses of human beings.

In NARS, the meaning of a concept, such as "Linda" or "feminist bank-teller", is determined by the available information about it, in terms of how it relates to other concepts, as far as the system knows.[2] For a given concept, such information may be either *extensional* (indicating its instances or *special cases*) or *intensional* (indicating its properties or *general cases*). To decide the extent to which a concept, "Linda", is

[1] Also, cf. http://www.cis.temple.edu/~pwang/papers.html for more publications.

[2] For a detailed discussion on the categorization model in NARS, see Wang and Hofstadter [2006]. Particularly, a connection is made to NARS' categorization model in section 7.2, where more aspectual considerations are discussed of how "conceptual entities" and "concepts" are proposed to be interrelated.

a special case of another one, "bank-teller" or "feminist bank-teller", the system will consider all available evidence. In this example, the most accessible evidence about all three concepts are *intensional* (i.e. about their properties), so the system reaches its conclusion by checking if Linda has the properties usually associated with "bank-teller" and "feminist bank-teller", respectively. Since, according to the given information, Linda has more common properties with "feminist bank-teller" than with "bank-teller", Linda's "degree of membership" is higher to the former than to the latter. This is judged as a "fallacy" when probability theory is applied *extensionally* to this situation, so only the *base rates* matters, while the properties do not.

In summary, as soon as a normative model of rationality as a general intelligence aspect makes more realistic assumptions, many "heuristics", "biases", and even "fallacies" can follow from them. In the above examples, there are strong reasons (cf. Wang [2009]) for assuming that the truth-value of a statement should depend on both positive and negative evidence (rather than negative only), and the meaning of a concept should depend on both extensional and intensional relations (rather than extensional only). These examples are believed to mainly show the limitations of traditional models (classical logic and probability theory, for instance), rather than human errors. The practice of NARS or similar systems should show us that it is possible for a new normative model to explain and reproduce similar results in a unified way (as further elaborated on in the rest of this chapter).

6.4.2 HDTP: GI-Based Rationality Through Analogy

As the second case study, a sketch is given of how HDTP can be used to implement some crucial parts of the suggested cognitively based theory of rationality. Heuristic-Driven Theory Projection, HDTP (cf. Chapter 3), provides a framework for computing analogical relations between two domains that are axiomatized in many-sorted first-order logic (cf. Schwering et al. [2009a]). It also provides an explicit generalization of these domains as a by-product of establishing an analogy. Such a generalization can be a base for concept creation by abstraction.

The modeling of the Wason selection task with HDTP is quite simple as long as appropriate background knowledge is available, in case an analogy should be established, or the lack of appropriate background knowledge prevents analogy making, in case no analogy should be established. In other words, the availability of appropriate resources in form of background knowledge is crucial.

If appropriate background knowledge for an analogous case is missing, then there

is no chance to establish an analogical relation or a potential analogy (with low coverage[1] and complex substitutions) is misleading the participant. Hence, participants have to apply other strategies. This is the situation when participants are confronted with the original Wason selection task based on properties of cards. Most participants have problems to establish a meaningful analogy with a well-known domain due to the high degree of abstractness of the task itself.

In the other case, if there is a source theory with sufficient structural commonalities, then the establishment of an analogical relation is straightforward. This happens if the task is changed in the following way. The rule that needs to be checked is now (Wason-Ages):

> "If someone is drinking beer in a bar,
>
> then someone must be older than 21".
>
> (Wason-Ages)

In the experiment, participants can choose between "drinking beer", "drinking coke", "25 years old", and "16 years old" (cf. Cosmides and Tooby [1993]). In the corresponding experiments, participants behave significantly better than in the original selection task.

With analogy making, the improvement of the human subjects in mastering the task can be explained. They can establish an analogy between the sketched set-up of the experiment and a standard situation in daily life, in which they would simply do the necessary actions (based on their background knowledge) to check whether there is someone who is drinking beer in the bar without being older than 21: (i) check people who are drinking beer, and (ii) check what people are drinking who are 16. As both situations are structurally very similar to each other, the generalization is straightforward, substitutions length are small, and coverage[2] is high.

The Linda problem is structurally different in comparison to the Wason selection task. In an analogy making context, an explanation of participants' behavior in terms of coherence maximization is promising. Coherence aspects of input theories are crucial for establishing analogies in several ways. Roughly speaking, the statement (Teller),

> "Linda is a bank teller", (Teller)

[1] Cf. section 3.1.3 for more on the idea of "coverage".

[2] Remember that the higher the coverage the better, because more support for the analogy is provided by the generalization. For more on the idea of "maximizing coverage of the involved domains", refer to the part discussing "ranking heuristics" in section 3.1.3.

has less coherence with Linda's profile than the statement (Teller&Feminist),

> "Linda is a bank teller and is active in the feminist movement".
>
> (Teller&Feminist)

Therefore, it is easier to establish an analogy between Linda as given in Linda's profile and Linda as described in (Teller&Feminist) than in the pure "bank teller" case, statement (Teller). Notice that from an abstract point of view the coherence-based resolution of the task is rather similar with the intensional interpretation of the task in NARS, where "feminist bank teller" has a higher degree of membership with Linda's profile than "bank teller".

6.5 Conclusive Remarks and Related Ideas

There are multiple models of rationality, each with its own assumptions and applicable situations. The traditional models are based on certain idealized assumptions, and thus are limited to the domains where the latter are satisfied. Since human cognition has evolved in (and is usually used in) realistic situations where those idealized assumptions do not hold, those models of rationality express their deviations from actual human behavior (not the other way round). Indeed, the models are what cannot actually be universally applicable, and observed human violations, therefore, should not be deemed "irrational" per se. The seemingly irrational behaviors are there not because the intelligent systems (e.g. humans) are irrational, but because the traditional normative theories do not cover rationality very well.

Instead of normative approaches, the ideas and discussions in this chapter advocate and stress a conceptually different view. In concrete terms, what seems to be precisely needed are new "models of rationality" that should be based on more realistic assumptions, and developed in a more holistic, cognitively inspired "framework". Such models must take humans' multifaceted cognitive capacities, as well as limitations of their cross-domain style of reasoning, into account. The "models of rationality" should be able to provide an adequate and feasible positive account of actual human rationality, without neglecting e.g. "bounded rationality" (cf. Simon [1955]) nor "ecological rationality" (cf. Rieskamp and Reimer [2007]). The models should as well accommodate particularities of human-style reasoning (a.k.a. irrational behaviors) based on the utilization of their cognitive mechanisms. The sought "framework" could form a cornerstone of a closer connection between AGI and cognitive science, embedding important parts of

the AGI program within a cognitive scientific context, whilst making the more general methods and theories of AGI accessible to the cognitive science side. In fact, the overall appeal for a "more cognitive" view on rationality models and systems is infrequent, but not unusual. Amongst others, already Kokinov reaches the conclusion that the concept of rationality as a theory in its own right ought to be replaced by a multilevel theory based on cognitive processes involved in decision-making. On the more technical side, there is a growing body of evidence that analogy engines (like HDTP) and general-purpose reasoning engines (like NARS) can be used for implementing these cognitive mechanisms and, thus, also as foundations of a rationality-guided approach to general intelligence.

This chapter should merely be considered as a point of departure, leaving questions for future research galore. For example with respect to the present proposal concerning HDTP, it seems recommendable to figure out to which extent different types of coherence concepts can be integrated into the framework. In particular, the challenges mentioned above need to be addressed, and a formal treatment of coherence in HDTP needs to be fleshed out (e.g. similar to the presentations given in Joseph [2011]). Furthermore, an implementation of coherence principles for retrieval, mapping, and re-representation purposed in the analogy making process needs to be formulated. Regarding competing theories for rationality, clarifying to what extent cognitive capacities and limitations have already been taken into account (implicitly as well as explicitly) when designing the theories, and to what extent the classical frameworks can be re-instantiated by a cognitively based approach, has to be considered one of the principal questions for future research. Finally, also on a fundamental conceptual level, a broader definition of rational beliefs is still needed, especially when representation and modeling issues are considered about employing beliefs as conceptual entities within concepts (cf. section 1.3 and Chapter 7).

7

Concept-Based Interpretation of Novel Noun Compounds

Cognitive scientists have long been interested in analyzing how people interpret a novel combination of two or more known words, the meanings of which are known, but that of the compound itself may have never been encountered before. In some human languages the combination that involves known words can itself range from the idiomatic or very well-known (e.g. Typewriter, Highway, Railway, Snowflake, etc.) to the unprecedented (e.g. Butterfly Milk, Cactus Finger, Computer Smile, Snake Glass, etc.). (Idiomatic combinations can also be referred to as *lexical compounds*.) Humans can create meaningful interpretations of novel compounds, even if the interpretations distantly differ from that of the two nouns they comprise. For example, people may think of the compound "Butterfly Milk" as reflecting something very difficult or impossible for a person to achieve. Generally intelligent cognitive agents like humans possess this amazing ability to understand such meanings of newly encountered compositions of arbitrary words, or create unprecedented such compositions to reflect composite meanings (e.g. "Brain Wash").

Whether expressed in exact or metaphorical senses, several models were proposed to show how such interpretations could be performed (cf. [Coulson, 2006; Estes, 2003; Keane and Costello, 2001; Wisniewski, 1997; Wisniewski and Gentner, 1991, for instance]). Cognitive psychologists often use the notion of "*conceptual combination*" to refer to the GI aspect that humans have of constructing meaningful novel concepts as combinations of input concepts, based on knowledge of the individual concepts composing such combinations. Linguists usually refer to resulting compounds of more than one word as *compound nominals*, which can act as nouns themselves.[1] For instance,

[1]Compound nominals are nouns in most of the cases, but they need not be. For example, 'get the ball

the compound "High Entrenchment Level Concept" is used in this chapter to refer to a concept that has an entrenchment level with a high value (cf. section 7.2). Both terms (i.e. conceptual combination and compound nominals) are referred to here as *conceptual compounds*, or simply compounds, because words are assumed to be represented as concepts on a language-independent level. The treatment of the process of juxtaposing two nouns is proposed here as a creativity-based production process. A conceptual compound thus denotes a newly established single entity that usually has a different interpretation than that of the (two) composing entities. Two forms of the compositions of arbitrary words have a particular familiarity: (i) adjective-noun compound forms (e.g. "red Nose") and (ii) noun-noun compound forms (e.g. "Book Box"), but this chapter focusses on the latter form. It is even preferred in this chapter to use the more specific notion of *modifier-head, noun-noun compound*, because it reminds the reader of the nature of the two-noun conceptual compound; that is, one noun acts as a modifier to the other head noun.

This chapter expounds a way of realizing "knowledge representation assumptions" in a special purpose concept-based framework, in order to interpret modifier-head compounds. It suggests a method to interpret novel modifier-head conceptual compounds, along with an illustrative example, in a particular type of concept-based models. The methodology to be employed in the interpretation is the blending of concepts. The claim is that a CB-based framework can feasibly be used to interpret novel compounds, where the interpretation of modifier-head compounds is achieved by a language-independent method that suggests elegant interpretations to conceptually blend the corresponding concepts.

7.1 Problem Importance and Challenges

A solution heuristic to interpret modifier-head compounds is proposed in this chapter that, in addition to employing the mechanism of CB, borrows notions from nature-inspired intelligence and belief revision to simulate the development of knowledge acquisition. In this section, a condensed overview is given first to motivate the importance of the problem in hand and to discuss some of its challenges found in the literature.

rolling' can be interpreted as 'Initialization'.

7.1.1 Motivations and Goals

The construction of noun-noun compounds is an indication of human cognition, yet interpreting (or creating) meanings for novel noun-noun constructions is an even more fundamental sign of general intelligence —that is, a GI aspect that is necessary to be captured by AGI models. Cognitive scientists are interested in explaining how the cognitive ability of constructing meanings is performed by human reasoners, while AI researchers aim at developing artificial models or systems that implement the ability. In the context of computational models of cognition, creativity, or general intelligence, a proposal to solve the problem (of interpreting novel noun compounds) eventually helps in affording cognitive agents the ability of creating possible meanings of newly formed combinations of known words.[1] Investigating the way humans interpret previously unseen modifier-head compounds helps in endowing cognitive agents with a simulated mechanism of the interpretation. The specific problem type addressed here is that of interpreting unprecedented modifier-head, noun-noun compounds. That is, previously unseen compounds that comprise exactly two already known nouns, the modifier followed by the head.

Plausibly speaking, a module in a cognitively inspired, computational model may be suggested to interpret previously unseen compounds *à la human*. That is, based on findings in cognitive science that report ways in which humans presumably assign meanings to new compositions of nouns. This necessitates deeper insights into how humans themselves develop interpretations of the compounds, based on their already-existing knowledge of the composing nouns as conceptual knowledge entities (cf. sections 4.2 and 1.3.1). Moreover, the ultimate goal would even be to develop an entire cognitively inspired, concept-based computational model of general intelligence that is based on cross-domain reasoning and accumulation of past experiences. This chapter discusses only concrete aspects of the problem of interpreting novel modifier-head, noun-noun combinations, and suggests only some characteristics needed in a concept-based model that can provide a solution method for the problem. The given discussion and suggestion do not ultimately provide a programming implementation, but rather a conceptual-level proposal of an abstract solution model based on combining knowledge domains in general, concept-based, models of general intelligence and computational creativity. In order to suggest a relationship possibility between the modifier and the head nouns in compounds, the given method utilizes an analogical relation to help generalizing, then blending domain representations of the constituent nouns.

As concepts are essential entities in representing and building the knowledge of

[1] Though the nature of the "interpretation construction" has many applications in various domains (e.g. in natural language processing (NLP) and information retrieval (IR); cf. Gay and Croft [1990]).

cognitive agents, nouns are assumed to be represented as concepts that develop by knowledge acquisition and revision (cf. section 7.2). A plausible artificial mechanization of the interpretation problem is suggested, which works through developing concept representations that give the interpretations of the nouns and their compositions.

7.1.2 An Overview of Problem Challenges

Some proposals related to analyses in the literature are listed next, linking them to previous presentations about concepts and ways of their representation. Beside acknowledging previous work, the analyses help in further clarifying the inherently baffling nature of the problem (and further illustrate that no agreement among researchers on a ubiquitous solution exists). Both the problem's importance and extreme difficulty are implied by the literature dedicated to solving it [Butnariu and Veale, 2008; Gagné, 2002; Gagné and Shoben, 1997; Gay and Croft, 1990; Hampton, 1997; Ryder, 1994, to mention just a few]. Without delving into details of the aforementioned contributions, their influence on inspiring the proposed solution model are highlighted in the following.

- The meaning of a novel compound may not even be simple to interpret by humans because it highly depends on many factors, as extensively discussed in Hampton [1997]; Wisniewski [1997]; Wisniewski and Gentner [1991]. For instance, a novel compound's meaning depends not only on the corresponding meanings of the composing words (which do not always have unique semantic mappings themselves), but also on the particular uses of such meanings, the surrounding context, and an *implicit relationship* between the composing words. The latter is a main challenge, since implicit relationships between a modifier and a head are extremely difficult to abstract. Compare for instance what "Wound" contributes to in a compound like "Hand Wound", to what it contributes to in the compound "Gun Wound" (cf. Coulson [2006]; Hampton [1997]; Levi [1978]). Existing experiences also influence one's comprehension; e.g. a "Decomposing Compound" to a chemist may differ from that to a linguist (cf. Gagné [2002]). A compound does not simply equal the sum of its parts, and its meaning is as sensitive to arbitrary changes as its underlying concepts, which can themselves develop over time by knowledge revision.[1] Specialized contexts and artificial anecdotes, from which a deviated meaning may possibly be inferred, can easily influence the background of a person. A person may say to someone else: "I will bring you whatever you need, even the *'Butterfly*

[1] Concepts in general are relativistic notions, and are sensitive to many sources of change. Think for instance about the relativity of a concept like being big, or about the changes in meaning over time of the concept COMPUTER: clerk, huge machine, PC, laptop, portable or handheld device, and so on.

Milk'", to exaggerate someone's readiness of performing very difficult tasks, for example.

- Having no consensus on what generally relates the head to the modifier, many researchers do have the consensus that comprehension requires the presence of *relational inferences* between the concepts in a conceptual compound. Nine *recoverably deletable* predicates are proposed by Levi, for instance, to characterize the semantic relationships between the composing nouns. Levi suggests that noun compounds "result from a syntactic transformation", in which relative clauses (e.g. "a wound caused by a gun") are converted into compounds (e.g "Gun Wound"). Semantic relationships between the composing nouns can, thus, be characterized as encoding one of these nine predicates: "cause", "have", "make", "use", "be", "in", "for", "from", and "about" (cf. [Coulson, 2006, pp. 126] and Levi [1978]). Furthermore, Gagné and Shoben's *abstract relations theory* indicates a limited number of predicates to relate a modifier with a head (cf. Gagné and Shoben [1997]).

- The *dual process model* (cf. Wisniewski [1997]) claims that attributive and relational combination are two distinct processes resulting from comparison and integration, respectively, whereas other linguistic models raise the possibility that a single-process integration model could account for all concept combinations (cf. Estes [2003]; Gagné [2002]). Other works could also be mentioned, such as the *constraints theory* of Costello and Keane, and the *composite prototype model* of Hampton (cf. Costello and Keane [2000]; Hampton [1997]), but the conclusion remains: the challenge is hard (and there is no consensus). Proposals showing how interpretations (of concept combination) might be performed by humans can be found in [Estes, 2003; Keane and Costello, 2001; Mareschal et al., 2010; Wisniewski and Gentner, 1991, for instance].

- Butnariu and Veale present a concept-centered approach to interpret a modifier-head compound, where the acquisition of *implicit relationships* between the modifier and the head is captured by means of their linguistic relational possibilities (cf. Butnariu and Veale [2008]). Unlike many other approaches, the approach indeed is concept-centered but, unlike this chapter's, it is linguistic-oriented and English-specific, so the approach may be difficult to apply to situations where online concept creation (i.e. on demand) is needed in achieving a GI level.[1] The outlined solution model in this chapter simulates the emergence an interpreta-

[1]However, unlike Butnariu and Veale's, the approach presented in section 7.3 does not use relational possibilities by means of both the modifier and the head.

tion of an unknown composition undergoes by employing multifaceted cognitive mechanisms, and prioritizing past experiences in suggesting the relational inference. It partly follows the claims of Gagné and Shoben; Wisniewski; Wisniewski and Gentner that relational possibilities may only be suggested by the modifier, which is the source concept in our case. In the present contribution, only the modifier plays the big role, and only an analogy-based relation (e.g. "looks-like") is implicitly assumed.

7.2 A Proposed Concept-Based Model

A method that provides a way to overcome several challenges of the problem of interpreting novel modifier-head compounds is presented in Abdel-Fattah [2012]; Abdel-Fattah and Krumnack [2013]. The method suggests, on a conceptual level, a concept-based solution to the problem, based on utilizing the CB framework within a presumptive model of computational creativity and general intelligence.

The reader will hopefully notice that the solution model to be described is cognitively inspired. It complies for example with Chomsky's ideas about the description of language and properties that all natural human languages share, and his proposals for innate predisposition abilities to learn languages (cf. Chomsky [1956]; Cook and Newson [2007]). In addition, the presented ideas agree with Gärdenfors's suggestions about not giving up beliefs that have high *epistemic entrenchment* (cf. Gärdenfors [1988]). That is, the more the beliefs (i.e. the knowledge entities) are used, the higher their epistemic entrenchment, the more difficult it should be for cognitive agents to forget (and vice versa: the less the knowledge entities are used, the less their entrenchment, the easier they can be forgotten). Furthermore, the construction of the suggested model's KB is influenced by Fodor's *language-of-thought* hypothesis (cf. Fodor [1983]), which encourages the assumption that beliefs, frames[1], or concepts are provided in modular groups to serve as input to the blending process. Accordingly, enough established (i.e. highly entrenched) concepts are assumed to be available at the agents' disposal in the model, so that the solution model obeys the case for humans where "a person has a repertoire of available concepts and ways of combining those concepts into higher-order concepts and into propositions" (cf. Hampton [1997]).

The proposed solution method posits, however, concrete (design) assumptions on a special-purpose model wherein the method could function. These posited assumptions

[1] In this chapter, the notion of "frames" is used in Fillmore's sense discussed in section 4.1.1 (also cf. section 7.2.2).

are presented in this section, before a detailed explanation of the method is explained in section 7.3.

7.2.1 Special Assumptions for Knowledge Acquisition in a Concept-Based Model

In modeling aspects of cognition and general intelligence, one may assume that the representation of the entire KB of cognitive agents in a model is built from conceptual entities (or beliefs that can be organized into knowledge concepts; cf. section 1.3). In addition to the related ideas already discussed in sections 1.3 and 4.2, this is also linked to Chalmers et al.'s discussions on aspects of "high-level perception" (cf.Chalmers et al. [1992]). Chalmers et al. consider (mental) representations as "the fruits of perception", with high-level perception enabling the construction of conceptual representations of situations.[1] Chalmers et al. also argue that perceptual processes cannot be separated from other cognitive processes even in principle. Thus, high-level perception is viewed as being "deeply interwoven with other cognitive processes, and that researchers in artificial intelligence must therefore integrate perceptual processing into their modeling of cognition" [Hofstadter and the Fluid Analogies Research Group, 1996, pp. 170]. Moreover, in gestalt psychology, the argument is that "human perception is holistic: instead of collecting every single element of a *spatial object* and afterwards composing all *parts* into one *integrated picture*, people *experience things* as an integral, meaningful whole. The whole contains an *internal structure* described by *relationships* among the *individual elements*" (cf. [Krumnack et al., 2013b, pp. 50; emphasis added]).

Therefore, the special solution model proposed here is assumed to, firstly, enable agents to acquire and store sorts of conceptual knowledge entities as *experiences* that direct the organization process. Some conceptual entities may be innate, some may be perceived facts, or beliefs given as inputs, others may be deduced, and so on. When agents need to make rapid (rather coherent) decisions, some experience-based 'mental shortcuts' enable them to categorize the knowledge they acquired by building schemas (or conceptual spaces; cf. section 4.1). In this way, the organization of the conceptual entities into knowledge concepts comes about. This affects the creation (and the entrenchment level) of another type of (internally organized) knowledge entities that do not result directly, or only, from perception, but rather from the repeated interplay between the already available knowledge and experience. That is, and recalling Gärden-

[1] More explications of this particular use of the notion of "situation" are given in Abdel-Fattah and Schneider [2013]; Schneider et al. [2013].

fors's ideas about beliefs that have high epistemic entrenchment, useful beliefs of all types will keep being reinforced, establishing links and ties to other knowledge where they are of use, whereas knowledge that is not always in use will typically be less remembered. As knowledge undergoes an internal screening, depending on the agents' history of experiences, some of the relations between acquired conceptual entities may fade out over time or partly forgotten. Agents can still form new frames of conceptual entities to compensate knowledge shortage, by means of linking seemingly related or analogical conceptual entities to create new ones.

A way in which agents can combine two existing concepts, for example, in order to create a third one depends on what, and how, entities are organized in each of the former two concepts (this occurs, in particular, when agents construct a meaning of an unknown word combination that consists of two nouns: a modifier- followed by a head-noun). Note that a model in which the proposed solution method can be implemented should be able to interweave the employment of perception and cognitive processes. One possible suggestion to achieve this is by allowing the model to employ ideas, in particular, from swarm and nature-inspired intelligence processes, such as the ant-colony optimization (ACO) techniques (cf. Dorigo and Stützle [2004]), in order to simulate the previously mentioned type of experience intensification and forgetting. In fact, Dorigo and Stützle's ACO approach was originally suggested as an approximation heuristic to solve hard, graph-theoretical optimization problems such as the traveling salesperson and the coloring of graphs (cf. Abdel-Fattah et al. [2005]). Their ideas are based on the way by which a population of ants communicate in a real-life ant-colony, where ants lay "pheromone trails" to mark the paths they take as they move around. Over time, the chemical substance (i.e. the pheromone) is either intensified (in case more ants keep following former ants' same paths) or evaporated. The very same idea can be used to simulate experience intensification (or forgetting) by adding (or decreasing) values that play the role of the pheromone trails.

Agents are also assumed to build and manipulate their KB in this special-purpose model using a KR framework that allows categorizing conceptual entities in (organized) knowledge domains. It should thus be possible for these entities and their organizing relationships to be expressed in a formal language (e.g. first-order logic in the case of the HDTP framework). On the one hand, the KR language should therefore allow more frequently used knowledge parts to be reinforced, establishing a kind of *intensified experience* that links to other knowledge parts where they are of use. On the other hand, knowledge that is not in use for a while will typically become less remembered.

Note: The study of a model of this kind is important in itself from both a theoretical and a practical points of view, and its applications are abound. The underlying assumptions discussed above intend to characterize abstract properties of such a model. This clearly raises at least as many challenging and interesting questions as the number of the aspects that can be considered in the study. For example, the formal descriptions call several ideas from AI and cognitive science, such as knowledge representation, concept learning and formation, and belief change (cf. for instance, Agre [1997]; diSessa [1988]; Gärdenfors [1988]; Zull [2002] and section 1.3). Moreover, there is no general consensus among cognitive psychologists and philosophers as to what concepts are, how they develop, or how they are represented (for an overview, cf. Lamberts and Shanks [1997]; Mareschal et al. [2010]; Murphy [2004]; Wrobel [1994] and section 4.2). In addition to its inherent difficulty, the latter issue of representing concepts is even connected with the expressiveness of the selected formal language. Furthermore, it is worth alerting the reader now that the discussions given below (until the end of this section; §7.2) explain the overall ideas of the proposed solution model in a general way, which employs graph-based illustrations to represent conceptual structures (cf. Sowa [1984, 2011]). This helps to understand the primary ideas about experience intensification and forgetting in an obvious way, and show the feasibility of possibly implementing the proposed solution model in different frameworks. However, in section 7.3, the discussions are centered around HDTP, which uses first-order logic. The chapter does not spell out details regarding how both ways of representation (i.e. graph-based and logic-based) are related, or in which sense they can be equivalent, because this is unnecessary for the current discussion purposes (but more about this can be found in [Chein and Mugnier, 2010, in particular §12.1.4; pp. 344–346]). Limitations of various sorts prevent a satisfying investigation of the model in this chapter, but the needed principles for the current problem focus are addressed as necessary.

7.2.2 Principles and Notations for a Concept-Based Model

The rest of the chapter uses the following principles and notations, which are based on characterizations and assumptions of the posited concept-based model:

1. The KB will be denoted by $\mathbb{K}_B$. It provides all conceptual entities (and their interrelations), and stores experiences as frames, formed from interrelated beliefs. Individual beliefs that belong to $\mathbb{K}_B$ are the main conceptual entities. They can be represented by propositions using the formalism of an underlying KR framework (e.g. a belief b can represent a predicate like $Shape(\texttt{MadameWhiteSnake}, curved)$, or

any axiom). Beliefs may be found to be interrelated, based on inputs or inferences. In such cases, the interrelations should reflect an update to the KB.

2. An acquired "*experience*" is a frame[1], which connects interrelated beliefs. It can be written in a form that combines beliefs $\bigcup_i b_i \in \mathbb{K}_B$, for some index i. But a frame must also (implicitly) determine a network of the related conceptual entities that belong to it (i.e. those beliefs b_i as well as their interrelationship links). A frame can thus be given in a form similar to that of a graph —that is, a set of nodes and another of connecting edges. For example, the ordered-pair $F = \langle \bigcup_i b_i, R_F \rangle$ can indicate that the beliefs $b_i \in \mathbb{K}_B$ are interrelated through relations listed in a set of (unordered) pairs $R_F = \{\ldots, (b_i, b_j), \ldots\}$.

3. Based on the discussions given in sections 1.3.1 and 4.2, a "*concept representation*" can be thought of as a collection of conceptual entities that form two mutually disjoint sets: a set of frames and a set of individual, isolated beliefs. (A visual representation of an arbitrary concept that has two frames and two individual beliefs is shown in Figure 7.1.)

4. A "*concept name*" identifies a concept's representation by assigning a unique string. Denote the set of concept names by $\mathbb{K}_C$ and consider it the model's *lexicon* that contains the names that shorthand the concept representations. For a concept with the name $c \in \mathbb{K}_C$, its representation (which consists of a set of frames and isolated beliefs) will be simply denoted by F^c. This should be thought of as an organized collection of conceptual entities of $\mathbb{K}_B$, but F^c can also be explicitly identified by listing the categorized frames and beliefs composing such a representation as a graph disjoint union (cf. Harary [1994]). For example, the notation $F^c = \left\langle \bigcup_{i_1=1}^{c_b} b_{i_1}, \bigcup_{i_2=1}^{c_f} F_{i_2} \right\rangle$ can be used to indicate that $c_b \geq 1$ individual beliefs and $c_f \geq 1$ entire frames are used in the representation of c, for some indices i_1 and i_2.

The sets $\mathbb{K}_B$ and $\mathbb{K}_C$ have different essences: the former is the set of all low-level conceptual entities, organized in terms of knowledge experiences (frames of interrelated beliefs) and expressed (somehow) in the KR's formal language, while the latter is a set of strings formed using an arbitrary alphabet of symbols. Still, however, for each (concept name) $c \in \mathbb{K}_C$ there is a (concept representation) F^c. Figure 7.1 shows a visual illustration of an arbitrary concept representation with 2 frames and 2 isolated beliefs.

[1]Recall form section 4.1.1 that Fillmore's linguistically motivated account of frames —as collections of categories, whose structure is rooted in motivating context experiences (that play an important role in building word meanings)— parallels Minsky's proposal of frames as data structures —that represent commonly encountered, stereotyped situations (cf. Coulson [2006]; Fillmore [1982]; Minsky [1974]).

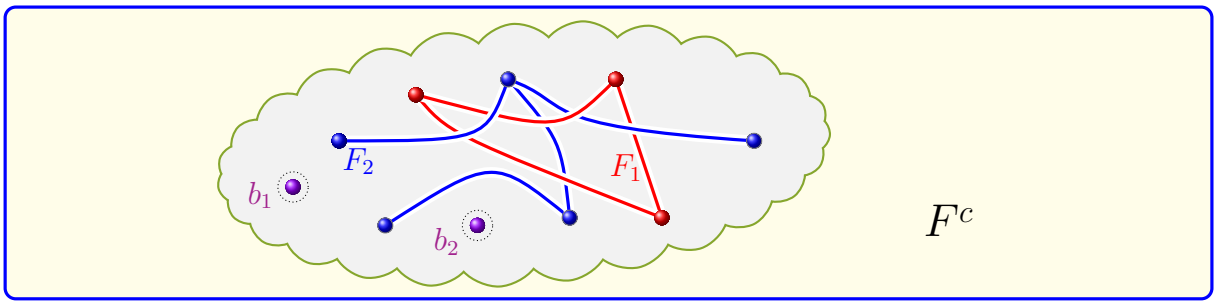


Figure 7.1: A visual illustration of a concept representation F^c (where c is an arbitrary concept). The illustration shows that the representation consists of two frames (F_1 and F_2) and two isolated beliefs (b_1 and b_2).

It is worth pointing out that the modeling of some cognitive processes based on more specific ways in which conceptual knowledge entities can be organized into concepts do already exist. Without giving further low-level representation notations specific to these models, one example to mention is Wang's NARS system (cf. Wang [2006, 2011, 2013] and section 6.4.1), which presumes a model of belief categorization that is close to the described one, though it uses a different representation formalism. Details of how categorization is achieved in the NARS system are thoroughly explained in Wang [2007]; Wang and Hofstadter [2006]. Also, conceptual structures and conceptual graphs would of course be another closely related direction of modeling examples that apply ideas akin to those presented above (cf. Chein and Mugnier [2010]; Sowa [1984, 2011]; Sowa and Majumdar [2003]).

7.2.3 Development of Conceptual Knowledge Entities

Entrenchment Values and Levels: The functioning of the suggested solution model is based on allowing agents not only to store past experiences as organized conceptual entities but to rank them as well (e.g. by assigning numeric values). A function is used here to serve as mnemonics of belief occurrences and rank beliefs according to importance and frequency. Namely, $e_V : \mathbb{K}_B \cup \mathbb{K}_C \rightarrow [0,1]$. Based on this function, *entrenchment values*, denoted by $e_V(b)$, can be assigned to beliefs $b \in \mathbb{K}_B$, depending on any number of factors[1], such as how recently, and how many times, the beliefs have been retrieved by an agent from $\mathbb{K}_B$ (e.g. when beliefs are retrieved in a concept formation process). In addition, the entrenchment values assigned to beliefs $b \in \mathbb{K}_B$ contribute, in turn, to assigning *entrenchment levels*, $e_V(c)$, to concepts $c \in \mathbb{K}_C$ if b occurs in F^c. In other words, entrenchment values of individual conceptual entities that underlie

[1]For example, Wang's system, NARS, makes use of two function values to reflect a *frequency* and a *confidence* for each knowledge entity (cf. Wang [2013]).

representations of concepts contribute to the entrenchment levels of these concepts.[1] This quantitatively reflects what knowledge is already known about a concept, and how 'much' has it been dealt with (i.e. not only *"how 'many' times"* has it been dealt with, but also *"at what frequency"* has it recently been experienced). In a sense, this can be used to parallel the amount of knowledge that is already known about a noun (in case K_C is the English alphabet, for instance).

The (overloaded) function $e_V : \mathbb{K}_B \cup \mathbb{K}_C \rightarrow [0, 1]$ is used for indicating both the entrenchment value of $b \in \mathbb{K}_B$, and the entrenchment level of $c \in \mathbb{K}_C$. By appropriately defining and manipulating e_V, a simulation of knowledge development based on experience can be achieved by mimicking, in particular, the ant-colony optimization approach mentioned previously (cf. Dorigo and Stützle [2004]). That is, it is possible to simulate an effect on updating e_V values by methods that parallel how ACO-based methods update "pheromone trails". Values of e_V can be intensified or weakened while building and organizing the knowledge frames of the concepts (and their blends). The function e_V can thus be used anyway to constantly update the entrenchment values of conceptual entities and, consequently, the entrenchment levels of concepts that involve these entities in their representations. Entrenchment values of more frequently used conceptual entities should be increased, indicating their importance in representing the frames (or concepts) they compose. Values decrease over time to simulate, in some sense, forgetting or a lowering in the importance of the conceptual entities' interrelationships with the frames (or concepts) they belong to.

A concept $c \in \mathbb{K}_C$ will be called a "*high entrenchment level concept*" (abbreviated HELCO) if $e_V(c) \geq \eta$, otherwise a "*low entrenchment level concept*" (abbreviated LEVCO); where $0 < \eta < 1$ is a threshold value.[2] Some conceptual entities (or whole frames or concepts) can be identified as 'innate' (i.e. built-in conceptual entities), with entrenchment values (or levels) equal (or very close to) one. Others obtain by concept formation, as is the case in concept blending, with entrenchment values (or levels) being initially less than η.

Description: In the presumed solution model, the interpretation of a novel compound by means of already-known nouns transfers to the process of forming new LEVCOs $c \in \mathbb{K}_C$ (and their corresponding representations F^c) by the conceptual blending of already-existing HELCOs. This means that when HELCOs combine, LEVCOs result with entrenchment levels that depend on those of the composing HELCOs. For a recently

[1] More precisely, entrenchment values contribute to the entrenchment levels of concept names (since the function, $e_V(c)$ is defined here as an entrenchment level for a concept $c \in \mathbb{K}_C$).

[2] Clearly, a decision will have to be taken about η by the model engineer.

blended LEVCO, say $B \in \mathbb{K}_C$, its entrenchment level $e_V(B)$ would be a function in $e_V(S)$ and $e_V(T)$ of the composing HELCOs S and T, where $S, T \in \mathbb{K}_C$.[1] Of course, many parameters need to be set for such an abstract model. For example, one may need to specify factors on which a calculation of e_V depends, and select an appropriate value for η (as well as a specification of this 'appropriateness'). More importantly would be the specification of how an entrenchment level for a recently blended LEVCO can be found, based on those of its composing HELCOs.

7.3 A Framework for Modeling Interpretations

This chapter is confined to handling the interpretation of a specific set of noun-noun composites, namely the modifier-head compounds. A combination of two already-known nouns, represented as concepts S and T, will be written in the form B = "ST": (i) S represents the modifier noun, (ii) the second concept, T, represents the head noun, and (iii) B refers to a resulting blend of the two input noun concepts. As pointed out by Wisniewski and Gentner, the modifier is used to adapt the meaning of the head, which interprets the combination "S T" as a function application $S(T)$, because S acts, in a sense, as an operator on T that, more or less, changes T's meaning (cf. Wisniewski and Gentner [1991]). The method presented here does not use function application, rather utilizes HDTP as a basis framework for CB, whereby the relational possibilities can be suggested only by the modifier (as suggested in Gagné and Shoben [1997]). An advantage is made of analogical transfer in CB, so that a new enriched domain is created while keeping the original target domain unchanged. This is clarified in the following, where a way is posited, by which HDTP helps in creating blends that represent novel combinations.[2]

7.3.1 From HDTP to CB

According to standard theory, a word is understood by the company of words it keeps (cf. Firth [1957]) or, according to the HDTP's jargon, by the *background knowledge* an

[1] In fact, $e_V(B)$ should be a function in $e_V(S)$, $e_V(T)$, and the entrenchment values of the conceptual entities defining (i) the representations of S and T, and (i) the representation of the generalization that results (cf. discussions at the end of next section; § 7.3.1).

[2] As briefly mentioned in the notes at the end of section 7.2.1, a tension should be avoided by the reader in the transition from the previous section to this one. Unlike in section 7.2, the representations in this section are not based on general or graph-based conceptual structures but rather centered around HDTP (which is logic-based). In addition to visual illustrations that help the purpose of explanation, concrete representations of concepts are therefore given in first-order logic.

agent possesses about the words as well as about the context in which they appear. For a combination B="S T", once S and T are appropriately axiomatized (i.e. represented) in sorted, first-order logic, they are provided to HDTP as source and target concepts, respectively. Accordingly, an axiomatization of (the operator) S is used as the source domain for HDTP, and an axiomatization of (the head) T as the target. In this way, HDTP provides a step towards the blending of the two given nouns (as concepts) and the usage of resulting candidate blends to interpret the combination.

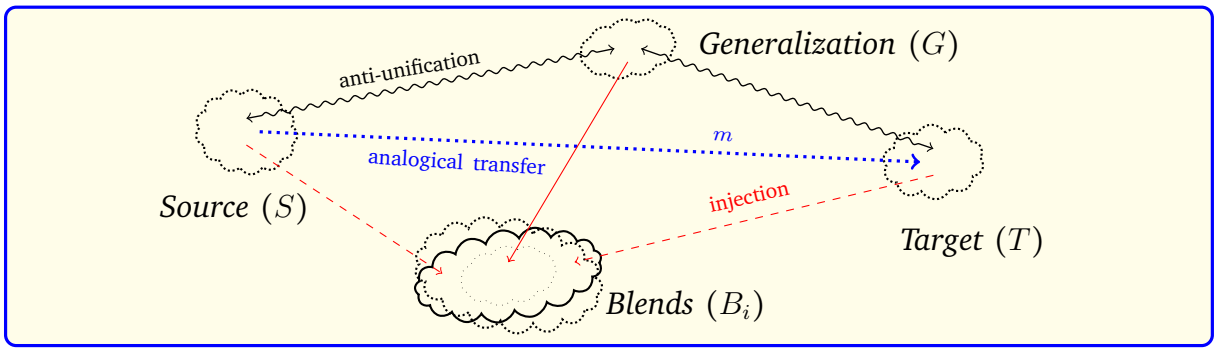


Figure 7.2: HDTP's overall approach to creating analogies and CB. S and T are source and target inputs, m represents the analogical transfer relation from S to T, and G is the generalization computed by anti-unifying S and T. The dashed arrows $S \rightarrow B_i$ and $T \rightarrow B_i$ describe the (partial) injections of facts and rules from the source and target into candidate blend spaces B_i.

The blending starts after providing S and T to HDTP, where an analogy is established and an explicit generalization, G, is computed (cf. Figure 7.2), which can be a base for concept creation by abstraction. When HDTP is applied to the domain inputs, blend candidates result that give possible interpretations of the compound (referred to by B_i in Figure 7.2). The transfer of knowledge, during analogical reasoning, is allowed in only one direction (and not the other) to pave the way for the "composition" and "emergence" steps of CB to come into play (cf. section 4.3.1). According to the outline in section 3.1.3, what happens is that HDTP proceeds during its two-phase analogy-making as follows:

1. in the *mapping phase*, S and T are compared to find structural commonalities[1], and to create a generalized description that subsumes the matching parts of both domains, and

2. after unmatched knowledge in the source domain is mapped to the target domain (during the *transfer phase*), blend hypotheses B_i, where $i \geq 1$, can be established as

[1] As can be viewed now, structural commonalities correspond to the 'identification' between SPACE_1 and SPACE_2 shown in Figure 4.3, page 86.

Source Axiomatization S ="SNAKE"	
$\forall x\ \exists w$ *Width*(x, w)	$(1a)$
$\forall x\ \exists l$ *Length*(x, l)	$(1b)$
$\forall x$ *Typical*$_S(x) \rightarrow$ *Shape*$(x, curved) \wedge$ *Skin*$(x, scaled)$	$(1c)$
$\forall x \exists l\ \exists w$ *Length*$(x, l) \wedge$ *Width*$(x, w) \rightarrow l > w$	$(1d)$
Target Axiomatization T ="GLASS"	
$\forall x\ \exists w$ *Width*(x, w)	$(2a)$
$\forall x\ \exists h$ *Height*(x, h)	$(2b)$
$\forall x$ *Typical*$_T(x) \rightarrow$ *Transparent*$(x) \wedge$ *Fragile*(x)	$(2e)$
Blend B ="SNAKE GLASS"	
$\forall x\ \exists w$ *Width*(x, w)	$(3a)$
$\forall x\ \exists l$ *Length*(x, l)	$(3b)$
$\forall x$ *Typical*$_S(x) \rightarrow$ *Shape*$(x, curved) \wedge$ *Skin*$(x, scaled)$	$(3c)$
$\forall x \exists l\ \exists w$ *Length*$(x, l) \wedge$ *Width*$(x, w) \rightarrow l > w$	$(3d)$
$\forall x$ *Typical*$_T(x) \rightarrow$ *Transparent*$(x) \wedge$ *Fragile*(x)	$(3e)$

Table 7.1: Parts of suggested noun axiomatizations and their combination.

interpretation suggestions of the composition "$S\ T$". Blends are injected with facts and rules from the inputs, based on the abstraction obtained by a generalization.

Additional types of implicit relationships between the modifier and the head may later be suggested and established during the transfer phase.

An Example: As a specific instance, consider the "SNAKE GLASS" compound. According to Wisniewski and Gentner, a group of human participants described the compound as a "tall, very thin drinking glass" (cf. Wisniewski and Gentner [1991]). The example given below illustrates a possible blend of (partial formalizations of) the domains representing the source and target nouns SNAKE and GLASS, respectively (cf. Table 7.1).

Irrespective of whether or not other constituents are included in the formalization, a representation of the concept SNAKE should normally emphasize the existence of some *salient* SNAKE characteristics. A suggested formalization is given in Table 7.1, in which the common-sense emphasis is on a SNAKE having a length that is much bigger than its width, a curved body shape, and a skin that is covered in scales. Also, the characteristics that a typical GLASS exemplar must have, among other salient characteristics, are its transparency and fragility. A GLASS object has dimensions determining its width and height. The blended domain, SNAKE GLASS, is an expansion of GLASS, the target, in which notions of 'shape' and 'skin' frames are added, taken from SNAKE (i.e. transferred). In principle, the blended domain theory can be thought of as coming from enriching the

first-order theory by which the target is represented with new notions (or frames) taken from the source, and then importing the axioms of the source into it (cf. Figure 7.3).

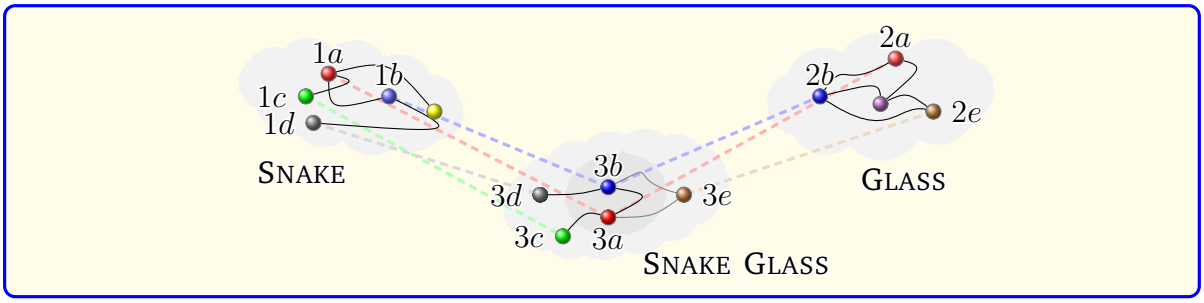


Figure 7.3: The form of the noun-noun blend, 'SNAKE GLASS', that results from the transfer phase of the blending between 'SNAKE' and 'GLASS' (cf. Table 7.1).

A blend of the two concepts that represent SNAKE and GLASS would import salient properties of SNAKE (i.e. highly entrenched conceptual entities or frames) that do not 'conflict' with GLASS's. In particular, based on Table 7.1, a blend candidate should indicate a relation between the *dimensions* of the SNAKE GLASS. From Table 7.1 and Figure 7.3, one can see that $(1a)$ and $(1b)$ are identified with $(2a)$ and $(2b)$, respectively, using HDTP, which also enables the inference (by injecting $(1d)$ into $(3d)$) that one dimension of SNAKE GLASS is much larger than the other. Furthermore, SNAKE GLASS would have conceptual entities indicating a curved shape (by injecting $(1c)$ into $(3c)$), and other non-conflicting constituents of SNAKE (in addition to injected, non-conflicting GLASS constituents, such as $(3e)$).

7.3.2 From CB to Interpretations

It is worth emphasizing that the suggested framework does not aim to function in the sense that two given nouns will only (or always) produce a unique result to interpret the compound under consideration. Cognitive science experiments show that humans too do not always agree on one meaning of the same given noun-noun combination, neither do they exactly follow one particular model each time they encounter a similar combination (cf. Mareschal et al. [2010]; Wisniewski [1997]; Wisniewski and Gentner [1991]).

The proposed framework rather enumerates alternative blends, ranked by the complexity of the underlying mappings. This is a desirable property in my view, because it (i) allows various possible interpretations instead of only one, and (ii) leaves space for experiences to play a role in deciding whether or not a specific blend is favored over

another.[1] People also interpret novel combinations by drawing on past experience with similar combinations (cf. Gagné [2002]). Moreover, combinations exist that are unlikely to be encountered in life, such as BOOK TIGER (cf. Wisniewski and Gentner [1991]), or have a meaning that distantly deviates from what the forming concepts refer to (e.g. a PIT BULL is a breed of dogs, and can be considered as a lexical compound).

What Values Can Affect Interpretations? In the example given above, every resulting SNAKE GLASS blend is intended to be interpreted (or represented) by an enumerated potential LEVCO, $B_i \in \mathbb{K}_C$, with $0 < e_V(B_i) < \eta$ and $i \geq 1$. Broadly speaking, and for any hypothesized blend interpretation B_i that can be constructed from a generalization G of modifier and head nouns, represented by S and T, respectively; the calculation of each $e_V(B_i)$ depends on the calculation of many other values. (Note that, in the suggested model, the represented characteristics of a given concept become more salient when their corresponding conceptual entities get reinforced most of the (more recent) times the concept is retrieved.) The following are particular values that need to be considered among those on which $e_V(B_i)$ depends:

1. *Entrenchment values of the conceptual entities that contribute to the representation of each input concept*: In fact, the values $e_V(b)$ for each conceptual entity $b \in F^S \cup F^T$ tell us how much each conceptual entity from the inputs should contribute to the overall entrenchment level of a resulting generalization from HDTP.[2]

2. *Entrenchment values of the conceptual entities that contribute to the representation of the constructing generalization*: Since resulting blends are to be constructed based on generalizations, entrenchment values of entities representing a generalization will affect the construction (and the entrenchment level of) blends that are constructed based on this generalization.

3. *Salient characteristics imported from the inputs*: The more salient a characteristic is (in an input concept), the more likely it should affect a blend.[3] Salient characteristics may not only have very high entrenchment values, but also appear within a frame, in which most of its composing conceptual entities do also have high

[1] A concept would be less favored for example if entrenchment values of most of its conceptual entities are low, or if it contains many sparse, individual conceptual entities that do not form a frame.

[2] Note that unions of concept representations (or even of frames, such as in $\bigcup_{i=1}^{cf} F_i$) may be interpreted as a graph disjoint union (cf. Harary [1994]).

[3] Being "infectious" and "self-replicating" are more salient characteristics of VIRUS than, for example, being an organic structure or a form of life (scientific opinions are different, in any case). Thus, in a blend such as COMPUTER VIRUS (or even BOAT VIRUS), one would highly likely think more about the former two characteristics of VIRUS than about the latter.

entrenchment values. This would typically result in most of the frame appearing in the generalization (i.e. the whole frame having salient characteristics is itself highly entrenched).

4. *Conflicting entities imported from the inputs*: One should in general avoid injecting an entity from one of the inputs that could conflict with another entity imported from the other input. When an entity is selected as a candidate for injection into a potential blend, this means it might have not already been part of generalizations offered by HDTP. Therefore, the relative importance of the entity should be assessed with respect to completing the frame within which it appears.[1]

How precisely $e_V(c)$ values of LEVCOs $c \in \mathbb{K}_C$ can be computed, or how several implicit relationships can be retrieved during the transfer phase in the analogy-making process, seem to be crucial questions left for a later, deeper analyses and formalizations of the proposed solution framework.

7.4 Conclusive Remarks and Related Ideas

Finding a meaning of a novel combination is a difficult (creative) task, yet providing a computational account that simulates the task in human cognition is an even more difficult one. Humans employ cross-domain cognitive mechanisms, especially analogy-making and concept blending, in developing their understanding of newly introduced conceptions that are basically combinations of already-known ones. Inspired by this claim, the chapter shows how it could be possible to propose a computational model of (creativity and) general intelligence that employs both mechanisms to contribute to solving the presented problem.

Basic challenges of the modifier-head interpretation problem (cf. section 7.1.2) incite not only a proposal to approach its solution using a non-classical method, but also encouraged preliminary characterizations of a conceptual model, suggested to implement the proposed method. A concept-based, logic-based, language-independent, cognitively inspired approach is presented, which has the potential to contribute to tackling the problem. The feasibility of constructing a blend in the described manner exemplifies the suggestion of how this form of noun-noun combinations could be approached in models similar to the one given in this chapter. Nevertheless, the chapter proposes

[1] One way is to compare the percentage of the entity's enclosing frame that is already in a generalization to the rest of the frame not included in this generalization.

ideas to confront some challenges but raises others, which need further elaboration and formalization.

The presented approach amounts to developing conceptual interpretations for creating new concepts. This is directed to emulate (creative) thinking via utilizing cognitive capabilities in making analogies and blends of concepts. The intuition behind the emulation goes as follows. While the mental process of interpretation (of a new modifier-head compound) is evolving, a meaning is 'invented' online using a cross-domain reasoning process. In this process, a virtual copy of the head noun is first thought of or imagined. Then, the copy is tweaked in a sense affected by the modifier noun's essential traits or salient characteristics.[1] In such a process, it makes sense to further assume that the newly created meaning can be a combination of the *salient characteristics* of the two words appearing in the compound, depending on how much in common the two words have, and on one's background knowledge. In the suggested model's terms, one can say that the frames defining the newly created concept result from blending the salient frames defining the composing concepts. The resulting features depend on (i) the organized entities representing the modifier and head concepts, (ii) previous encounters of the meanings (that is, enforcing previous experiences), as well as (iii) how a head noun may "look like" when it is attributed to the modifier noun. For instance, how may a BOX look like when it is attributed to a BOOK in the compound BOOK BOX, and how may a GLASS look like when it is attributed to a SNAKE in the compound SNAKE GLASS, etc. Note that the "saliency" of a concept's feature or trait results from enforcing and re-enforcing repeated experiences that are related to the concept's defining frames, and this is one reason why notions about the ant colony optimizations's "pheromone trails" are recalled here.

I believe this agrees with people's continuous re-conceptualization of their understandings of conceptions as they are encountered over time (and in different contexts). I do not claim, however, that this is precisely how concepts are processed mentally, neither do I claim that this way always gives best meaningful outputs. There are some inspiring and motivational reasons, however, why a combination is proposed to be possibly modeled as explained. For example, the principles given in Costello and Keane [2000]; Keane and Costello [2001], the developmental psychology literature in [Lamberts and Shanks, 1997; Mareschal et al., 2010, for instance], the discussions in Gust et al. [2011], the studies and experimental results in Wisniewski [1997]; Wisniewski and Gentner [1991], the discussions of concepts in Chapters 1.2.2 and 4, and the

[1] When one encounters BOAT VIRUS, for instance, one would normally start to think about VIRUS and tweak it by BOAT, but not about BOAT first.

other ideas (of Chomsky; Fodor; Hampton) mentioned within this chapter's discussions (cf. section 7.2.3). All this provides a support from research in cognitive science why novel combinations are proposed to evolve the given way. Moreover, this way allows a form of blending that respects the dual process of comparison and integration, on which famous models are already based (cf. Estes [2003]; Gagné [2002]; Keane and Costello [2001]; Levi [1978]), yet relational possibilities still can only be suggested by the modifier, which is the source concept in the current case (but it is not unusual; cf. Gagné and Shoben [1997]; Wisniewski [1997]; Wisniewski and Gentner [1991]).

From a modeling point of view, the way analogy-making is made use of in identifying common parts of the source and target concepts of a modifier-head compound, in generalizing them, and creating blends, serves maintaining relational and attributive combinations at the same time. However, the implicit relational possibility that analogy provides us with between the head and the modifier still does not account for many of the different cases that can be encountered (because a combination B="S T" is interpreted as "T that looks-like S" or "T that is in-the-form-of S"), though it seems promising and could be improved by using the relationships between the underlying frames of the given concepts.

Of course, neither HDTP nor the CB framework alone intend to solve the challenges altogether, but the method presented here is considered a first starting step towards the interpretation of noun-noun compounds using a new concept-based perspective. It presumably overcomes some representation challenges that are usually faced in designing cognitively inspired models of (computational creativity and) general intelligence; the model is promising and can be used in other applications, I presume. The encoding of rated experiences and the use of levels of entrenchment for concepts in comparable concept-based models can help in achieving solutions to other challenges, such as when concepts get changed or externally affected by newly observed facts.

8

An Implementation-Oriented Explication of Analyzing Counterfactual Conditionals

8.1 '*Being Smart*': Essences and Mechanisms

There is a growing need to identify various benchmark aspects of what makes humans more generally intelligent than other cognitive beings. Particular aspects that characterize human-level cognition (e.g. creativity and rationality) are identified in former chapters (cf. Chapters 5 and 6). This chapter identifies an additional benchmark aspect: how humans analyze counterfactual conditionals? The chapter emphasizes that the problem of analyzing counterfactual conditionals is crucial to be identified as a benchmark aspect in artificial systems that aim at modeling general intelligence.

Proposing methods (or entire systems) that can abstractly or computationally model identified aspects is of no less importance, since a better understanding of how a specific aspect operates may better be realized when the methods efficiently describe how an aspect works. Aiming at an implementation-oriented explication, the chapter thus investigates the roles of cognitive mechanisms responsible for reasonable analyses of counterfactuals. It proposes how to computationally contribute to solving the problem by AGI systems, and point out some challenges that artificial systems may encounter in computationally solving this problem. As in former chapters, the given arguments in the current chapter show that the operational utilization of analogical mapping and conceptual blending is helpful in overcoming these challenges. In fact, this utilization leads to reasonable analyses of counterfactual conditionals in artificial cognitive systems.

This chapter seeks to achieve three connected goals concerning the problem of ana-

lyzing counterfactual conditionals:

1. First, it attracts the reader's attention to humans' cognitive competency of analyzing the reasonability of counterfactual conditionals. This sheds light on the importance of a problem that has been maltreated in artificial systems that aim at modeling human-comparable intelligence, despite its wide importance and despite its long history across several fields (cf. [Byrne, 2005; Fauconnier and Turner, 1998; Lewis, 2001; Turner and Fauconnier, 1998, for example]).

2. Secondly, it discusses cognitive phenomena that could be responsible for this particular competency in humans. Here, it is argued that the ability to analyze this kind of conditionals is one of the essential aspects of intelligence, which needs to be better treated and better understood when building artificial cognitive systems.

3. Finally, the chapter shows that the analyzability has the potential to be represented and computed by integrating the functionalities of analogy-making and conceptual blending; two of the fundamental, multifaceted cognitive mechanisms that have proved to play important roles in endowing cognitive systems with essences of cognition (cf. [Fauconnier and Turner, 2002; Hofstadter, 2001; Martínez et al., 2011, for example] and Chapters 5 and 6).

The rest of the chapter is structured as follows. The problem of counterfactuals is introduced in section 8.1.1. An elaboration on how a cognitive system might approach the problem is conceptually discussed from a high-level perspective in section 8.2. In section 8.3, a proposal on how to formally achieve this is presented. A detailed worked-out example is given in section 8.4, before section 8.5 concludes the chapter with final remarks.

8.1.1 Counterfactual Conditionals (CFC)

A *counterfactual conditional* (from here on CFC), is a conditional sentence in the subjunctive mood: an assumption-conclusion conditional that designates what would be (or could have been) the case if its hypothetical antecedent were true. CFCs are also known as subjunctive conditionals or remote conditionals. They are contrasted with both

1. "material conditionals": in which the antecedent and the consequent may have no relation in common, yet the conditional itself can be true (because its truth value depends only on those of the antecedent and the consequent); and

2. "indicative conditionals": which can be thought of as operations given by statements of the form "If *antecedent*, (then) *consequent*".

Although indicative conditionals, too, may be sometimes seen as contrary-to-fact statements, a straightforward comprehension difference between indicative and subjunctive conditionals is classically shown by the well-known Oswald/Kennedy pair of examples given by the sentences 8.1 and 8.2 (cf. Adams [1970]):

If Oswald did not kill Kennedy, someone else did. (8.1)

If Oswald had not killed Kennedy, someone else would have. (8.2)

Sentence 8.1 shows an indicative conditional, while sentence 8.2 is a subjunctive version. The majority of people would accept the former as reasonable yet reject the latter (cf. Adams [1970, 1975]; Pearl [2011]). Another difference is given in Santamaría et al. [2005], where participants read presupposed facts very rapidly, indicating that a "priming effect" occurs when human participants read CFCs and not when they read indicative conditionals.

Table 8.1 gives a general form and some examples of other (sentences that paraphrase) counterfactual conditionals. A major part of the CFCs can be given in the general form of sentence 8.4, but other sentences may also be paraphrased to agree with this form. For example, sentence 8.5 can be written as:

If Nashwa had not cooked the dinner,
then Ahmed would have cooked the dinner. (8.3)

The general form of sentence 8.4 has two parts: an *antecedent* (i.e. the assumption) and a *consequent* (i.e. the conclusion), which are both hypothetical statements. According to standard semantics, both parts could be 'false' (at least the assumption is a known falsehood). The concern, thus, is not with binary truth values of CFCs, like the case for material implications, but rather with analyzing and verifying CFCs and their conditions for being considered meaningful or reasonable.

In addition to the importance of their computational evaluation per se, CFCs situate themselves within entertaining scopes of end-to-end artificial systems. Counterfactual reasoning is involved, and plays an important role (one way or another), in problems and puzzles of domains as diverse as learning, theory-of-mind, moral judgement, or decision-making under risk and uncertainty. In the field of theory-of-mind, for example, the children in the famous muddy-children problem (cf. Shoham and Leyton-Brown

If (it were the case that) *antecedent*, then (it would be the case that) *consequent*.	(8.4)
Ahmed would have cooked the dinner if Nashwa had not done so.	(8.5)
If Mubark had not stepped down after the revolution in 2011, Egypt would have suffered from a military coup in 2011.	(8.6)
If Mursi had stepped down in 2012, Egypt would not have suffered from the military coup in 2013.	(8.7)
In France, Watergate would not have harmed Nixon.	(8.8)
If Julius Caesar was in command during the Korean war, then he would have used the atomic bomb.	(8.9)
If Julius Caesar was in command during the Korean war, then he would have used the catapult.	(8.10)

Table 8.1: A list of sentences that represent or paraphrase counterfactual conditionals.

[2009]) take actions because they fail to verify "what-if" situations that are contrary to their (common) knowledge. The general case of the problem is described in Table 8.2, where n honest, logical-reasoner children commonly know that both (i) $1 \leq k \leq n$ of them is muddy, and (ii) the question: "*do you know whether you are muddy?*" have already been asked publicly for k rounds. A child would know she is muddy not only by the common knowledge but also by first thinking to herself: "if I were not muddy, I would have known by the $(k-1)^{st}$ repetition of the question (the common knowledge) that the rest already know they are muddy". Then, and as the child fails to verify this CFC, the child concludes she must be muddy.

A group of n children played in the mud. Their father notices that k of them have mud on their foreheads and says: "*at least one of you has mud on his forehead*". The children can all see each other's foreheads, but not their own. All of the children are intelligent, logical reasoners, honest, and answer simultaneously. The father keeps repeating the question: "*do any of you know that you have mud on your forehead?*" without receiving any responses for exactly $k-1$ rounds. Immediately after the k^{th} repetition of the question, all the children with muddy foreheads raise their hands simultaneously, indicating that they now know they are the k muddy children.

Table 8.2: A description of the "Muddy Children" situation (adopted from [Shoham and Leyton-Brown, 2009, pp. 393–394]).

8.1.2 Analyzing CFCs by Humans and in Artificial Systems

A CFC is considered to be *verifiable* if its contrary-to-fact conclusion consistently follows from its contrary-to-fact assumption by a reasonable judgement. The analysis of a CFC is the reasoning process that leads to the judgment, which is assumed to hold in a (third) contrary-to-fact world that, in turn, depends on the reasoner's background and reasoning strategies. The verification of a CFC is a *judgement of reasonability* that involves the subjective importation of knowledge-based facts (cf. [Lee and Barnden, 2001, p. 8]) and is weaker than logical validation. Yet this judgement can always be disputed (cf. Goodman [1947]; Quine [1960]), using CFCs like sentence 8.9 and sentence 8.10, for instance (cf. section 8.4).

The reasonable analysis of CFCs is seen as a fundamental cognitive competency that may be used to designate, evaluate, and compare superior cognitive systems. It obviously requires a cognitive system to proficiently *create contrary-to-fact conceptions*, in order to reasonably analyze a given CFC. Humans, the ultimate exemplar of cognitive beings, are without any doubts the unique species that can perform such a reasonable analysis. They can do this because, in particular, they utilize logical reasoning, create alternatives to reality, communicate with language, hold rational beliefs, show rational behavior, as well as employ several cognitive capacities (cf. Abdel-Fattah et al. [2012a,b]). It is dazzling how humans smoothly analyze a given CFC and may convincingly estimate a rough truth degree, and even argue about it. In general terms, this can be achieved in humans by the imagination of a whole set of alternative conceptualizations that differ in certain aspects from their real world counterparts, but in which the CFC's antecedent holds. The reasoning process is then carried out in creatively imagined worlds, yielding coherent results (cf. Byrne [2005]). It is proposed that this process can be achieved in artificial systems when the system is endowed with (computationally plausible versions of) such abilities. In the following, a short literature overview identifies the most important ones of these abilities.

8.1.3 A Crisp View of Specific Treatments

The representation and verification of CFCs have always delivered debates within many disciplines, like philosophy, psychology, computer science, and linguistics. Important contributions in the literature are mentioned to back up the ideas in the later discussion.

Philosophical treatments: Beside Goodman's discussion of CFCs (cf. Goodman [1947]), another classical line of work by Lewis and Stalnaker uses possible world semantics of modal logic to model CFCs based on a similarity relation between possible

worlds. According to Lewis's account (cf. Lewis [2001]), the truth type of a CFC in the form of sentence 8.4 can be either vacuously true, non-vacuously true, or false. This depends on the existence of a closely similar possible world to the real world, in which the antecedent and the consequent are true. The account is unclear as to what 'similarity' (or 'closeness') mean, and it did not use values to represent truth degrees.

Psychological treatments: Many cognitive scientists would agree that reasoning, in general, requires the creative production of mentally constructed conceptual entities (cf. Gentner and Stevens [1983]; Johnson-Laird [1983]; Johnson-Laird and Byrne [1991]; Mareschal et al. [2010]). The creation and verification of CFCs, in particular, as alternatives to reality are widely explored in the pioneering work of Byrne (cf. Byrne [2005]), where many experiments about reasoning and imagination are carried out. In Byrne's context, human imagination is seen as rational thinking, where people rely on background knowledge when they try to think logically. Accordingly, "a key principle is that people think about some ideas by keeping in mind two possibilities" (cf. Santamaría et al. [2005]). This means that *two mentally constructed* domains are needed in assessing the truth of a given CFC (which are treated in this thesis as conceptual spaces, and referred to as source and target domains).

Linguistic treatments: Classical approaches view language as consisting of statements that can be reasoned about in terms of their truth functions. But some linguists also deal with meaning construction in natural language by means of mentally constructed spaces and their blending (cf. Coulson [2006]; Fauconnier [1994]). Of a particular interest to this chapter is Lee and Barnden's analysis of CFCs in cognitive linguistics (cf. Lee and Barnden [2001]), which based on the mapping between different reasoning spaces and the drawing of analogies between these spaces. This analysis is also implemented in an AI reasoning system and applied to the verification of certain CFCs (cf. Lee and Barnden [2001]), which further indicates that a form of analyzing CFCs can already be computed by artificial systems.

Algorithmic treatments: Recently, an algorithmic approach towards CFCs was presented by Pearl (cf. Pearl [2011]). Complete procedures for discerning whether a given counterfactual is 'testable' and, if so, expressing its probability in terms of experimental data are given in Shpitser and Pearl [2007]. Pearl's basic thesis of treating counterfactuals states that their generation and evaluation is done by

means of "symbolic operations on a model" that represents the beliefs an agent has about the "functional relationships in the world" (cf. Pearl [2011]). In this way, Pearl views the procedure as a concrete implementation of Ramsey's idea (cf. Ramsey [1929]), in which a conditional is accepted if its consequent is true after its *antecedent is (hypothetically) added* to the background knowledge, making whatever *minimal adjustments* that are required to *maintain consistency*.

8.2 A Tale of Two Multifaceted Mechanisms

The modeling of counterfactual reasoning is not only highly disputed, but can also be considered to be AI complete: while seemingly easy for humans, the treatment of CFCs poses a hard problem for artificial systems. However, the utilization of computationally plausible cognitive mechanisms in the analysis of CFCs appears to be achievable in artificial systems. The analysis of CFCs is a clear competency of humans (cf. section 8.1.2), which obviously requires a high level of artificial intelligence if cognitive agents were to acquire this competency (or approximate it) in any cognitive system that models it. Thus, and particularly when it comes to developing computational cognitive systems that can analyze the reasonability of CFCs, this competency is considered as a complex-structured mechanism. The verification of CFCs is proposed to be achieved by means of reducing this complex mechanism to simpler, rather essential, cognitively motivated, and computationally plausible mechanisms (such as analogy-making and conceptual blending).

By abstracting the major ideas of the various treatments given in section 8.1.3, one can discover that 'similarity' between 'domain worlds' (or creatively imagined 'conceptions') plays a shared role in all the treatments. One can also note that an artificial modeling system may need to, at least, develop processes that (i) consider 'conceptual domains' as inputs (cf. section 1.3), (ii) compare the 'similarity' between these domains (cf. Chapter 2), and (iii) judge the reasonability of a given CFC by deciding whether or not a 'blend' of these concepts 'remain consistent' after 'adding the antecedent' of a given CFC to the background knowledge.

Analogy Making: The Role of *"The Core of Cognition"*: Analogies are an important aspect of reasoning and "a core of cognition" (cf. Hofstadter [2001]), so they can be used to explain various types of behavior and decisions (cf. Abdel-Fattah et al. [2012a]; Kokinov [2003] and Chapter 6). Analogy is important for concept learning and can also be seen as a framework for creativity (cf. Abdel-Fattah et al. [2012b]; Hofstadter

and the Fluid Analogies Research Group [1996]; Holyoak and Thagard [1996] and Chapter 5). The ability to see two dissimilar domains as similar, based on their common relational structure, is fundamental and ubiquitous for human cognition (cf. Gentner et al. [2001]). Former chapters show that an analogy engine can be useful in modeling several aspects of cognition. This chapter continues the same trend and shows that it also helps in analyzing CFCs in computational cognitive systems. Like in former chapters, HDTP is used as an example of an analogy-making system for computing analogical relations between two domains (cf. Chapter 3).

Creation by Integration: The Role of *"The Way We Think"*: Conceptual integration (conceptual blending, or CB) is claimed by Fauconnier and Turner to underly "the way we think" and explain "the nature and origin of cognitively modern human beings" (cf. Fauconnier and Turner [2002]). Chapter 4 thoroughly presents CB as a multifaceted mechanism that facilitates the creation of new concepts by a constrained integration of available knowledge.

The Combined Role: Analyzing CFCs by Employing Analogies and CB: Based on section 8.1.3[1], the treatments along the various directions appear to utilize humans' cognitive abilities of:

1. conceptualizing hypothetical domains (as alternatives to reality) that contain the necessary background knowledge,

2. intelligently drawing analogies between parts of the domains (and associating some of their constituting elements with each other), and

3. constructing a variety of possible consistent conceptualizations, in which the given CFC can be verified.

Therefore, the ideas of CB may be used, side by side with analogy-making, to analyze the reasonability of CFCs by blending two input mental spaces and constructing a space, in which the analysis of CFCs can take place, referred to as *"counterfactual blend spaces"* (cf. section 8.3.2). In section 8.3, the main idea is explained. The basic argument is that the combination of (i) a powerful analogy engine and (ii) the ideas of CB, potentially endows cognitive systems with the ability to reasonably analyze (some) CFCs in an intuitive way. From an implementation-oriented perspective, this implies that artificial

[1] See also Byrne [2005], where many cognition experiments are given that further supports the proposed view.

models can analyze the reasonability of CFCs as long as computational versions of the aforementioned cognitive mechanisms, in particular, can be utilized.

8.3 Towards a Treatment Formalization: Constructing Counterfactual Blends

A Summarized Description of the Treatment: This section explains that (at least some) CFCs can be analyzed by constructing appropriate blend spaces, using analogy between input domains that correspond to the antecedent and the consequent of a given CFC (the form of sentence 8.4). The given procedure is based on a structural mapping of two input domains that correspond to the antecedent and the consequent of a given CFC. The structural mapping gives rise, in turn, to several *blend candidates*, which import major elements (i.e. knowledge or conceptual entities; cf. section 1.3) from one or the other input domain. The importation may render some blend candidates (logically) inconsistent, which reflects a non-reasonability of the given CFC. But those blend candidates that satisfy specific criteria (beside being consistent) will reflect a given CFC's reasonability. In this way, a heuristics is formulated to choose the most plausible candidates, guided by the (logical) structure of the given CFC based on some fixed principles (cf. section 8.3.2).

In the given treatment[1], the analysis of a given CFC (in the general form of sentence 8.4) requires the creation of two mental domains for each of the involved parts (i.e. the antecedent and the consequent). In order to find similarities and suggest common background between the two parts, analogical mapping is used to compare the structural aspects in both domains. Associations between the two mentally constructed domains can thus be found. Finally, a logically consistent combination of the two domains can be suggested, as a newly created blend of them, in which the reasoning process can hold. This cross-domain reasoning process will take place in a blend space that forms the setting to verify the CFC. Constraints could be imposed to give preference to one blend over another. Additionally, each conceptualization may be given a rank reflecting its relative plausibility.

To put these (and section 8.2's) ideas into a formal framework, the process will be split into two steps:

[1]The approach may seem to have common characteristics with Lee and Barnden's or Fauconnier's, because all of them are more or less inspired by analogy and blending. However, the treatment in this chapter adopts a more general blending procedure and use a different method and heuristics to suggest the construction of blends.

1. the generalization of the given domains of a CFC (via analogical mapping), and
2. the construction of a counterfactual space (via conceptual blending).

Each one is explained in more details in the following.

8.3.1 Generalization and Structural Mapping

The mapping is based on a representational structure used to describe the two domains. In a computational system these descriptions may be given in a formal language, like first-order logic. The strategy applied here is based on the HDTP framework (cf. Schwering et al. [2009a] and section 3.1), but this chapter uses a schematic form of natural language for the given examples, in order to improve readability.

The basic idea is to detect *structural commonalities* in both domain descriptions by a generalization process. Then, based on this generalization, objects from both domains that have corresponding major roles can be identified. As an example consider the following parts of the real and hypothetical worlds according to sentence 8.3:

Nashwa cooked the dinner	(REL)
Ahmed cooked the dinner	(HYP)
X cooked the dinner	(GEN)

Sentences such as (REL) and (HYP) can be generalized by keeping their common structure and replacing differing elements by variables in a generalization (GEN). This generalization gives rise to associations, in particular:

$$X : \textit{Nashwa} \triangleq \textit{Ahmed}.$$

In Figure 8.1, common parts of representations of (REL) and (HYP) (particularly, *Nashwa* and *Ahmed*) are assumed to be identified by an analogy between their enclosing, structured, input domains (REAL and HYPO, respectively), which represent the antecedent and the consequent, respectively, of this example's CFC (cf. sentences 8.5 and 8.3).

It is clear that the richer the conceptualizations of the domains, the more correspondences may arise. However, an essential point in constructing the generalization is the principle of "coherence", which states that if a term occurs in multiple statements of a domain description, it should always be mapped to the same corresponding term of the other domain (i.e. consistent reusability of mapped terms). Such a reusable mapping of terms is a good indicator for structural correspondence.

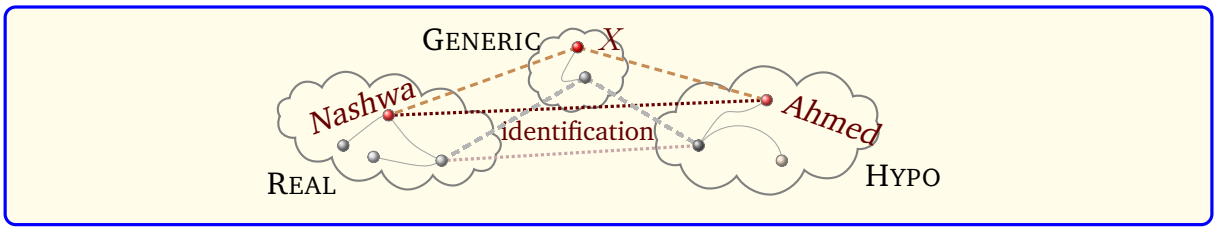


Figure 8.1: In the GENERIC space, the element/term X generalizes/anti-unifies two elements/terms that play similar roles in their corresponding domains. This illustration is based on sentence 8.3, where SPACE_1 and SPACE_2 of Figure 4.3 (cf. page 86) are replaced by REAL and HYPO, respectively (also cf. section 8.3.1).

8.3.2 Reasonability Principles for Counterfactual Blend Construction

The established mapping is used as a basis for constructing *counterfactual blend* candidates. (A "counterfactual blend" will henceforth be denoted CFB.) Statements from both input domains can be imported, and the mapping is applied for merging them. But one should note that the objects, which are covered by the mapping, must play the same role in both input domains. Therefore, their simultaneous existence in a CFB is considered incompatible, although normal CB explicitly allows simultaneous occurrence of corresponding entities from both domains in the blend space (cf. Coulson [2006]; Fauconnier and Turner [2002] and the explanations given in section 4.3.1). Thus, for each such object, the proposed treatment must have a way to reasonably choose one of the alternatives in a systematic way. The following "*reasonability principles*" are proposed to guide the construction of CFBs:

(P1) "*Counterfactuality*": A CFB candidate should satisfy the antecedent of the given CFC.

(P2) "*Choice*": For every matching pair, one alternative is allowed to be imported into a CFB candidate.

(P3) "*Consistency*": A CFB candidate should sustain (logical) consistency.

(P4) "*Maximality*": A CFB candidate should contain as many imported instances of the original axioms as possible.

As it rules out many meaningless and unneeded possibilities from the beginning, (P1), the principle of counterfactuality, will be the starting point of departure to achieve a reasonable CFB. It forces the antecedent of the CFC to hold in a CFB candidate and

thereby provides the first reasonability criterion for selecting alternatives from the mapping pairs. In a next step, an initial description of a CFB candidate can be enriched by importing additional statements from any of the two input domains, keeping all the principles satisfied. During importation, all terms covered by the mapping have to be replaced coherently by the chosen alternative.[1] If no alternative for a term has been chosen yet, a choice has to be made and marked for *all subsequent occurrences* of that term. In general, the process should try to maximize the number of imported statements to allow for inferences of concern. One however has to assure that the constructed CFB stays (logically) consistent.

These reasonability principles do not always lead to a unique CFB by allowing for multiple variants. This should not be considered as a downside, but rather an essentially desirable feature in implementations of artificial systems. Indeed, this feature may not be easily achieved in classical AI or cognitive systems without emphasizing the role that CB plays. Thanks to the ideas of CB, this feature allows for alternative verifications of a given CFC (cf. section 8.4), where the existence of multiple (reasonable) CFB spaces simulates the indecisiveness of humans in judging a given CFC. Remember that the judgement of a given CFC may always be disputed (cf. Goodman [1947]; Quine [1960], sentence 8.9 and sentence 8.10), which means that a modeling system may need to allow the possibility of having several (reasonable) CFB candidates for arguing about the same given CFC in several ways. A more concrete explanation is given at the end of section 8.4. A simple example is first presented in the following to demonstrate the discussed ideas, leaving the thorough discussion to the more detailed, worked out example in section 8.4.

A Simple Example

The following is a simplified explanation of how a CFC can be formalized. The CFC used in the explanation is a metaphor discussed in Turner and Fauconnier [2003]:

If Clinton were the Titanic, the iceberg would sink. (CLT)

[1] This discussion implies that (P1), in particular, has a remarkable effect, not only on ruling out the importation of many meaningless and unneeded possibilities (e.g. inconsistent statements), but also on keeping a CFB candidate reasonable by enforcing modifications on some of the statements that may be imported: in this way, an inconsistent statement may have the potential to be modified to another version that can be imported. The imported version of the modified statement reasonably cohere with all the imported statements in the CFB candidate in hand (also cf. footnote 1 on page 167).

The metaphor introduces two input domains. The first domain is that of political affairs in Washington, which contains among other things knowledge entities given in sentences (CL1) and (CL2), whereas the second domain comprises the events around the Titanic in some facts, including (TI1) and (TI2) in particular:

Clinton hits *the scandal*, (CL1)

Clinton does not sink, (CL2)

The Titanic hits *the iceberg*, and (TI1)

The Titanic sinks. (TI2)

From this data alone, a generalization can be constructed, consisting of generalized facts that can be instantiated in both input domains. In the current case, the generalization would contain only one proposition:

X hits Y. (Gen1)

The generalization in (Gen1) gives rise to an analogical mapping:

$$X : \textit{Clinton} \triangleq \textit{The Titanic} \qquad Y : \textit{the scandal} \triangleq \textit{the iceberg}.$$

Based on this analogy, a CFB can now be constructed using the four principles stated earlier. According to the principle of counterfactuality (P1), the antecedent of the current CFC (i.e. statement (CLT)) has to be introduced into the blend:

Clinton hits *the iceberg*. (B1)

This instantiation allows already to choose alternatives according to the principle of choice (P2):

$$X \mapsto \textit{Clinton} \qquad Y \mapsto \textit{The Titanic}.$$

The principle of maximality (P4) invites the importation of additional facts from the input domains into the CFB (substituting terms as necessary) such as:

Clinton does not sink, or (B2)

Clinton sinks. (B3)

Here (B2) is imported from (CL2), and (B3) is imported from (TI2), by applying the substitution. However, the resulting blend candidate can be rendered inconsistent in case it contains both (contradictory) facts (B2) and (B3). The principle of consistency (P3) forces the removal of (at least) one of the conflicting facts, (B2) or (B3), from the blend candidate. For the intended interpretation, one would remove (B3). Assuming suitable background knowledge, like "if A hits B, then A or B sinks", one can logically derive the the conclusion of the original CFC in statement (CLT). Note that if (B2) were selected to be removed from the (inconsistent) blend candidate instead of (B3), keeping (B3) itself in the constructed blend candidate, one could still derive another conclusion of the original CFC in statement (CLT). For example, it would then be the case that "Clinton hits the iceberg but sinks", which could be argued as a less-acceptable conclusion.

Digression: Some Remarks

Based on Abdel-Fattah et al. [2013a,b], the approach is presented here in a very general way that avoids discussing issues of deeper (mostly philosophical) nature. For instance, no explicit constraints are mentioned here on what counts as an admissible set of inputs. Absence of such constraints can allow the anchoring of input domains to be in metaphorical, impossible, or phantasy worlds, such as "Star Wars". Still, in such cases it would also be unclear whether to consider the conditionals as *counterfactuals* (with "factuality" being described by "impossible" worlds) or *counterpossibles* (with impossible antecedents). The latter notion of "counterpossibles" is defined by Lewis to refer to conditionals with impossible antecedents, and are always vacuously true regardless of the consequent. Some approaches would, in addition, argue in favor of a representation language for CFCs that is different from the one the HDTP framework considers. One natural proposal would then be to express axioms in the domain theory as weighted constraints capable of being broken at a cost, such as the case in Bello [2012].

This type of questions is not particular to the work presented here, and is commonly encountered, yet known in the history of CFCs to be rigorous and painstaking. In the current chapter, thus, no distinction is made between kinds of CFCs nor between possible categorizations of concepts representing their antecedents and consequents. To recall, the presentation only focusses on considering the three facets of the problem stated in section 8.1, namely: (i) emphasizing that the problem is crucial for consideration by systems aiming to imitate aspects of humans GI, (ii) investigating the cognitive mechanisms responsible for reasonable analyses of counterfactuals, and (iii) proposing

how to computationally contribute to solving the problem by utilizing these mechanisms in such systems.

8.4 An HDTP-Based Explication

A worked out example is given in this section to provide a more detailed explanation of the procedure and constraints described briefly in the previous section. The example also provides different lines of argumentation for verifying one given CFC.

The Caesar-Korean CFC Example: Recall the following conditional from section 8.1.1, which is already listed in Table 8.1 and based on [Quine, 1960, p. 222] and Goodman [1947]:

> If Julius Caesar was in command during the Korean war,
> then he would have used the atomic bomb. (8.9)

This conditional is to be interpreted in a "hypothetical world", as it combines elements (*Caesar* and the *Korean War*) that do not belong together in the real world. Such hypothetical world can be constructed by blending two domains, the *Gallic Wars* (RE) (for *Roman Empire*), on the one hand and the *Korean War* (KW), on the other. To formalize the example, the background knowledge in the two domains that are believed to be relevant to this discussion are stated first (N.B. temporal and tense aspects are disregarded in the given statements and representations). For the (RE) domain, this background knowledge can include the axioms:

> *Caesar* is in command of the *Roman* army in the *Gallic Wars*, (RE1)
> The *catapult* is the *most devastating weapon*, and (RE2)
> *Caesar* uses the *most devastating weapon*. (RE3)

On the other hand, the (KW) domain can include the axioms:

> *McArthur* is in command of the *American* army in the *Korean War*, (KW1)
> The *atomic bomb* is the *most devastating weapon*, and (KW2)
> *McArthur* does not use the *atomic bomb*. (KW3)

Based on these axiomatizations, and according to the ideas discussed in section 8.3, a generalization can be computed. The statements that will enter the generalization by applying HDTP are only those, for which instances are present in both domains. According to the given representation, this includes:

X is in command of the Y army in Z, and (G1)

W is the *most devastating weapon*. (G2)

From the generalization, a mapping of corresponding terms in both domains can be derived:

X : *Ceasar* $\triangleq$ *McArthur* Y : *Roman* $\triangleq$ *American*

Z : *Gallic Wars* $\triangleq$ *Korean War* W : *catapult* $\triangleq$ *atomic bomb*.

Recall from section 5.4, in particular Figure 5.3 (cf. page 108), the way in which the blending process is proposed to function based on the HDTP framework. Accordingly, CFB candidates can now be constructed by merging the (RE) and (KW) domains, identifying axioms and entities matched by the generalization, and keeping the four reasonability principles for CFB satisfied (cf. section 8.3.2). For the current example, this implies that the antecedent of the CFC under consideration (sentence 8.9) must be satisfied in each reasonable CFB candidate (according to (P1), the principle of counterfactuality). It also implies that the CFB is enriched by "injecting" non-contradicting knowledge entities from the (RE) and (KW) domains. The generalized part that will appear in any CFB must therefore replace each occurrence of X by *Caesar*, each occurrence of Y by *American*, and each occurrence of Z by *Korean War*. So, one may start by insisting that a CFB candidate contains:

Caesar is in command of the *American* army in the *Korean War*, (B1)

then continue enriching the CFB candidate by importing further statements from the input domains, such as:

The *atomic bomb* is the *most devastating weapon*, (B2)

Caesar uses the *most devastating weapon*, (B3)

Caesar does not use the *atomic bomb*, (B4)

and so on. However, the unconstrained enrichment of a CFB by the mere injection of

knowledge entities from both domains may render it inconsistent or unreasonable. For example, a CFB that injects all of (B1), (B2), (B3), and (B4) violates the consistency principle (P3) because (B4) contradicts what could be inferred from (B2) and (B3), namely (B5):

Caesar uses the *atomic bomb*. (B5)

In Figure 8.2, depictions are given to illustrate: (i) the Korean War domain (KW), (ii) the Gallic Wars domain (RE), (iii) the given generalization, and (iv) two (reasonable) CFB candidates (their construction is discussed below). For simplicity, Figure 8.2 does not identify the terms but only the statements that are composed of those terms.

The principle of counterfactuality (P1) can single-handedly prevent the importation of many (implausible) sentences that can be injected into a CFB candidate.[1] In the current example, (P1) already enforces the choice for three of the mapped terms:

$$X \mapsto \textit{Caesar}, \qquad Y \mapsto \textit{American}, \qquad Z \mapsto \text{Korean War}.$$

Therefore, one can no longer import (implausible) statements such as:

McArthur does not use the *atomic bomb*, nor (KW3)

McArthur is in command of the *Roman* army in the *Gallic Wars*, (NoWay)

into any CFB candidate (otherwise the candidate is clearly unreasonable). Nevertheless, many CFB candidates can, in principle, still be constructed. A CFB candidate may import as many (plausible) statements as possible from any (or both) of the input domains (but perhaps not, simultaneously, all of them); sustaining its reasonability by making use of the guiding principles. However, the importation of one (plausible) statement or another may be found to cause reasonability problems. That is, a statement can have the potential to be imported into a CFB candidate, but may not be imported into such a candidate because this is practically prevented by the reasonability principles (otherwise the importation will render the CFB candidate unreasonable). For instance, consider (RE2) and (RE3) which infer (by classical deduction in the (RE) domain):

Caesar uses the *catapult*. (RE4)

[1] In fact, the principle of counterfacuality does more than that. One may have noticed that (B2) is an imported version of (KW2), and (B3) is an imported version of (RE3), whereas (B4) is not a directly imported version of (RE4), nor of any other statement. (B4) is a rather restrictedly imported version of (KW3), in which the term *Caesar* replaces *McArthur* (according to (P1)).

An unmodified version of (RE4) can, in principle, be imported into a CFB candidate (unlike (KW3), which cannot be imported into any CFB candidate unless it is modified by (P1)). But it is possible that one of the reasonability principles disallow the importation of (RE4) into a specific CFB candidate, especially when contradicting statements have already been imported into the same CFB candidate[1].

One could in general get several, reasonable CFB candidates for the same CFC, but some of them may eventually be (logically) equivalent, according to our principles and heuristics. Two (non-equivalent) blend spaces for the CFC in hand are given in Figure 8.2, and described in the next paragraphs.

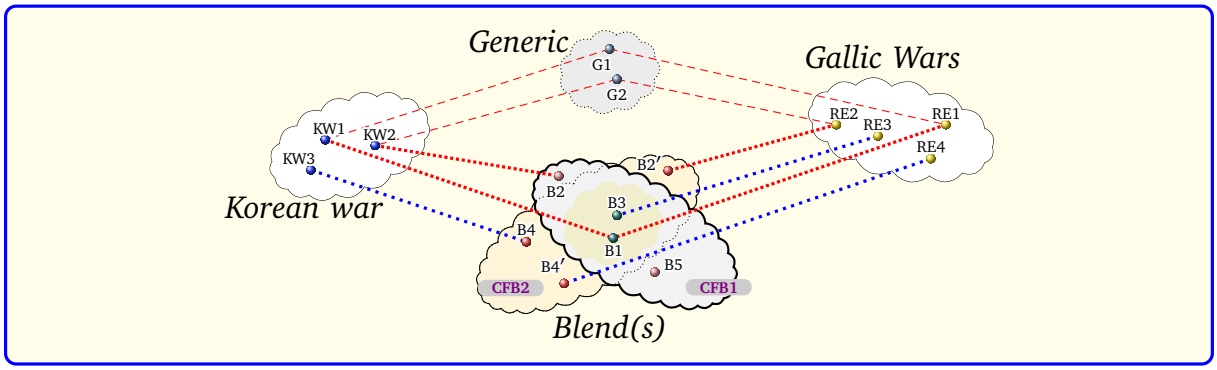


Figure 8.2: An illustration of two possible blend spaces for the CFC of sentence 8.9. For the sake of simplicity, the illustration does not show the mapped terms but rather depicts some of the representing sentences given in section 8.4.

(CFB1): The main representational sentences of this blend candidate include (B1), (B2), (B3), and (B5) (cf. Figure 8.2). This blend reasonably verifies the CFC because it implies that: "*Caesar* is in command of the *American* army in the *Korean War* and uses the *most devastating weapon*, which is the *atomic bomb*". This CFB could be equivalent to another one that only contains (B1), (B2), and (B3) as axioms, since (B5) is (consistently) deducible from (B2) and (B3).

Note that (B1) is supported by (P1) and (P2). (B2) is imported using (P2); similarly (B3). Finally, (B5) is a direct inference of (B2) and (B3). Note that (P3) prohibits the injection of (B4) into (CFB1): (B4) is an instantiation of (KW3) in which *Caesar* instantiates X, but (B4) has a potential clash with (B5).

[1]The candidate (CFB1) described below, for instance, prevents (RE4) to be directly imported as it is, because the directly imported versions of (RE2) and (RE3), namely (B2) and (B3), respectively, infer the statement (B5) that contradicts the imported version of (RE4) (namely, (B4′)). Whilst, (RE4) can be directly imported, as it is, into another candidate, (CFB2), which does not include "both" imported versions of (RE2) and (RE3).

(CFB2): This is an alternative blend space, which reflects the possibility that *Caesar* would use the *catapult* and not the *atomic bomb*. Its axioms include (B1), (B3), (B4), and the following sentence:

The *catapult* is the *most devastating weapon*, (B2′)

which is a directly imported version of (RE2). Also, in (CFB2):

Caesar uses the *catapult*, (B4′)

results either as an inference from (B2′) and (B3), or as a directly imported version of (RE4) (which, itself, can already be inferred from (RE2) and (RE3) in the (RE) domain). In this (CFB2) blend (cf. Figure 8.2):

1. *Caesar* is in command of the *American* army according to (B1),
2. the *catapult* is considered the *most devastating weapon* according to (B2′),
3. *Caesar* does not use the *atomic bomb* according to (B4), but rather
4. *Caesar* uses the *catapult* according to (B4′).

In addition, according to the proposed maximality principle[1], (CFB2) is more 'maximal' than (CFB1). According to the illustrations in Figure 8.2, (B4′) does not belong to (CFB1), which means that *Caesar* cannot use the *catapult* as an alternative in (CFB1).

8.5 Conclusive Remarks and Related Ideas

The problem of analyzing CFCs has a long history in many disciplines, yet very few computational solution frameworks exist (especially as part of an AGI system). This chapter emphasizes the importance and argues for the feasibility of considering cognitive mechanisms in attacking this challenging problem. It focusses on the two cross-domain, multifaceted, cognitive mechanisms, with which the other chapters of the thesis are also concerned with and proposes a computational strategy to contribute to solving the problem by cognitive systems. As a proof of concept, and to give a concrete explication example of applying the main ideas of the suggested strategy, the presentation was based on the HDTP framework, using a schematic form of natural language to improve readability. But a desirable characteristic of the suggested strategy is its generality, which allows for applying the presented ideas in other frameworks as well.

[1] As well as according to the currently given representations, of course.

The chapter expounds the opinion that the general problem of analyzing CFCs deserves to be a benchmark problem for comparing modeling approaches of aspects of general intelligence and cognition, by considering the evaluation of how they propose to analyze different types of the CFCs. Not only are certain CFCs clearly quite harder to imagine or to reason about than others, but the relationships of the entities appearing in these CFCs (among themselves and between their counterparts in the factual world) can be hard to handle as well. Nevertheless, a distinction in treating different CFCs needs to be reflected by such modeling approaches in verifying such CFCs. This would require considering, among many, many other profound issues, a precise schematization of CFCs (cf. Lewis [2001]), combined with a cognitive architecture capable of handling natural language processing and dynamic outlooks on semantics (cf. Veltman [2005]), in addition to a precise formalization of several characterizing facets of the conceptual integration paradigm itself (cf. Fauconnier and Turner [2002] and section 4.4).

In future work, the focus should be on answering some of these and other related questions. For example, in the process of analyzing a CFC, the aspects in which the real and the hypothetical worlds differ may not be very obvious to identify. Even in his possible-world semantics treatment of CFCs, Lewis did not give a precise definition of what a "miracle" is (cf. Lewis [2001]). In any case, the setting of an adequate alternative CFB space calls for the construction of a (temporary) knowledge domain that may contain counterfactual knowledge entities. A construction–analysis process, like the outlined one, could be what one might expect from an artificial cognitive system. Also, the presentations in this chapter tried to restrict the form of the CFC to that of sentence 8.4, though it is still important to identify characteristics of the CFCs, to which the proposed approach can (or cannot) always be applied. No doubt that this is a completely non-trivial issue, in particular because a unified representational scheme may also be required. Moreover, actual computational models still need to be deeper investigated in order to get more practical insights into implementing the presented ideas.

9

Conclusions

A summary of the essential issues discussed in the preceding chapters, and some final elaborations and comments, are given in the following to conclude the text in hand.

Utilization Pervasiveness of Multifaceted, Cross-Domain Mechanisms: The text has focused primarily on analogical reasoning and concept blending as multifaceted cognitive mechanisms, providing an overview on how they can be utilized for modeling an abundance of general intelligence aspects. This was aiming to show that the utilization can model solutions to baffling cognition problems in a wide range of domains, including reasoning, concept representation and concept manipulation. Both mechanisms (i.e. analogy and blending) can be seen from an abstract perspective as types of cross-domain reasoning, where the former discussions help in appreciating why establishing cross-domain connections between seemingly unconnected domains is thought to be a highly important part of modeling aspects of human-level intelligence. The discussions also emphasize the importance of assuming why knowledge ought to be organized in some form of domains, and that an underlying KB has to provide conceptual entities in groups that serve as input to cross-domain processes. The HDTP framework supplied a robust and flexible setting for achieving such emphasis.

The concrete set of problems and their suggested solution ideas are expected to encourage the utilization of organized fusion of conceptual entities, abstraction of concepts, and analogy-based reasoning, in building computational models of cognitive systems that are able to demonstrate several GI aspects. Employing cognitively motivated techniques in overcoming challenges related to solving such cognition and general intelligence problems is to a large extent still in its very early infancy —no doubt, this affected the level of presentation throughout the thesis. But the presented application directions show a notable pervasiveness. In order to continue the overall argument,

spread over the preceding chapters, this conclusion also argues why accounting for an integration of these pervasively utilized capacities (when designing a general intelligent system) might be a good and quite rewarding idea. The following overview of different domains and scenarios puts into further perspective the significant difference to state-of-the-art artificial, cognitive systems that the pervasiveness of multifaceted cognitive mechanisms, predominantly analogy and concept blending, are expected to realize.

Offering Novel Problem-Solving Routes: It is undeniable that AI has made significant progress in the field of problem solving over the last decades (cf. section 1.2). In most cases, however, the power of the current "approaches" is limited. There is, for example, the ubiquitous danger of facing the often feared "combinatorial explosion" and the abyss of underspecification of problems. Applying ideas via utilizing cognitive mechanisms might offer an alternative approach to a solution for certain problems, avoiding thus some classical pitfalls that most AI researchers used to struggle with:[1]

1. Conceptual blending, or CB (cf. Chapter 4), offers elegant solution ways to look at puzzles like Smullyan's "*Rate-Time Puzzle*" (cf. [Smullyan, 1978, pp. 12]) and Koestler's "*Riddle of the Buddhist Monk*" (cf. Fauconnier and Turner [1998, 2002]). Particularly, note in solving the "Riddle of the Buddhist Monk" (cf. section 4.3.1) that, instead of treating two days as separate time spans, they are blended into one, resulting in a scenario in which there seem to be two monks, moving towards each other, and meeting exactly in the place the riddle asks for an answer to. Following this approach, the CB technique does not only give an answer in terms of existence of the place, but also features a "constructive way" to provide implicit information on its location.

2. A particularly great benefit is expected from integrating and developing cognitive mechanisms into general intelligent systems in the branch of (artificial) productive creativity (cf. Chapter 5 and the model of concept construction, proposed in Abdel-Fattah and Schneider [2013]; Schneider et al. [2013]). Blending, in particular, has already been identified as a key element in the concept generation stage in creative design processes (cf. Nagai [2009]). It is undeniable that a form of blending is pervasive in most occurrences of human (productive) creativity over all time (cf. Martínez et al. [2012]): This ranged from old mythology (e.g. "Pegasus" and "Centaurs") to old and modern storytelling (e.g. , *"Alice's Adventures*

[1] Needless to say, some ubiquitous challenges would remain, such as the time complexity needed for obtaining satisfying solutions.

In Wonderland"; cf. Carroll [1869], and "The Hunger Games"; cf. Collins [2010]) to modern product design, lifestyle , and even gaming devices (e.g. the famous company's name came to be "LEGO" from the Danish phrase "leg godt", which means "play well"; tablet computers can also be seen as blends of journals and computers).

3. The mechanisms under concern can be of a novel use in the area of rationality and rational behavior (cf. Abdel-Fattah et al. [2012a]; Besold et al. [2011] and Chapter 6). For many years, researchers from different areas (e.g. game theory, psychology, logic, and probability theory) have tried to find feasible theories and models for human reasoning and human-style rationality. The results of these efforts, though being highly elaborate and theoretically well-founded, mostly fall short when actually predicting human behavior. Analogical reasoning and CB can offer a way out (cf. Chapter 6 for a more detailed treatment). Thus, the consideration of the mechanisms gives an intuitive and comparatively simple explanation to some classical challenges to existing theories of rationality, making it also an interesting candidate for inclusion in future general intelligent systems (which undoubtedly will have to deal with issues related to rationality and human behavior).

4. Chapter 7 discusses that CB in particular can provide additional functionality in language understanding and production within (concept-based) AI systems. The capabilities of humans that are exhibited in making sense of neologisms and previously unknown word combinations are impressive. In many cases, actual natural language interface systems are stretched to their limits to capture interpretations of noun compounds, although the individual meanings of the combined terms might already be known to the system (in particular if a concept-based representation approach is used as KR; cf. sections 1.3, 4.2, and 7.2). Again, CB here offers a guideline for combining concepts into blends, thus making accessible also the resulting blended concept, as e.g. in Goguen's by now established, classical example concerning the combined words BOATHOUSE and HOUSEBOAT (cf. section 4.3.2). From a natural language interface point of view, integrating the faculties of concept representation and blending into a general intelligent system is expected to be advantageous in both aspects of language production and language understanding: Whilst the output of the system might seem more human-like due to the occasional use of combined words based on blends, also the language input understanding part might profit from these capabilities, making use of blend-

ing techniques when trying to find the meaning of unknown, possibly combined words. (For some further considerations concerning CB and noun–noun combinations cf. Abdel-Fattah [2012]; Abdel-Fattah and Krumnack [2013]; Martínez et al. [2011] and Chapter 7). Humans cannot only induce appropriate meanings of unknown words, both within and without a given context, but can also produce words that reflect particular meanings or descriptively summarize specific situations (e.g. giving the name "Facebook" to the online social networking service, yet a "face book" is, physically, a directory that helps people get to know each other).

5. As a further application, consider the analysis of counterfactual conditionals (cf. Chapter 8), where humans can flexibly restrict the judgement of whether or not a given counterfactual conditional is verifiable, although the conditional might indicate many, or possibly infinite, possibilities (cf. Abdel-Fattah et al. [2013a,b]). A cross-domain reasoning ability possibly allows for unfamiliar results by blending two (or more) familiar concepts, opening thus the door to non-deductively achieve thinking outside the box through cross-domain reasoning. Counterfactual reasoning is yet another example that emphasizes the importance of utilizing mechanisms other than deductive reasoning in (creative) concept construction. (In fact, deduction may not be preferred in the first place when computationally analyzing counterfactual conditionals.)

Not surprisingly, many ideas, of what currently is regarded standard in narrow-AI (cf. section 1.2.2), have already been laid out in Newell and Simon's article (cf. Newell and Simon [1976]). They very early anticipated improvements that seem inevitable from today's perspective, and aimed at implementing a system that has "the same scope of intelligence as we see in human action" [Newell and Simon, 1976, pp. 116]. On the one hand, works such as Newell and Simon's inspired the scientific society to begin studying human-like intelligence, but, on the other hand, it was "search" that has mainly been proposed as a cornerstone of computing since the beginning of AI (cf. Newell and Simon [1976]). Although ideas for "intelligence without much search" have been proposed in Newell and Simon [1976], which later flowed into the cognitive architecture SOAR (cf. Laird [2008] and section 1.2.3), much less interest has been dedicated since then to computationally plausible cognitive mechanisms, which may rather utilize cross-domain reasoning (cf. Abdel-Fattah and Schneider [2013]). Still, even recent projects that focus on simulating human intelligence are clearly lacking generality. Despite their undeniable, impressive success, IBM's chess-playing computer DeepBlue (cf. Hsu [2002b]) and question-answering system Watson (cf. Ferrucci et al. [2010]), for example, lack any aspect of creativity with regards to the tasks they can solve. Achieving

aspects of general intelligence calls for a plethora of complex capacities that are closely connected with the utilization of various multifaceted cognitive mechanisms. Creative concept construction, as an example of a fundamental GI aspect, may definitely need searching in performing specific sub-processes, such as finding analogies to newly available concepts before modifying (or adapting) them. But, anyway, it certainly differs from mere search in several sides.

An Incomplete List of Promising Mechanisms: In the context of cross-domain reasoning, several cognitive mechanisms are summarized in Table 9.1. Blending, analogy-based, as well as additional mechanisms (like re-representation, frequency effects, and abduction; cf. Martínez et al. [2012]) already cover many interesting aspects of higher cognitive abilities (cf. Table 9.1). There is however no claim that the list is in any sense complete. This list further supports the claim in this thesis that it is necessary to endow cognitive systems with the mentioned mechanisms, in order to cover varieties of higher-order cognitive abilities. Additional several low-level abilities are, of course, still needed for a cognitive system to model general intelligence in a holistic and complete sense. The presented approaches do not cover aspects like reactive behaviors, sensorimotor loops, or certain attention mechanisms. These aspects are challenges, addressed by other mechanisms that are neither directly related to analogy making nor concept blending. The integration of higher and lower cognitive mechanisms could be achieved by a hybrid architecture, presently a rather standard approach for integrating heterogeneous methodologies.[1]

The Final Crowning and the Prospective Research: Investigating, testing, and implementing cognitive mechanisms need to attract more attention, in particular when one is interested in modeling as perplexing aspects of GI as those discussed in this thesis. We may agree that good science always requires good observations, but their cross-linking can be far more advantageous. Furthermore, I expect that AGI should basically work on delivering unconventional ways to achieve a level of human-comparable intelligence, which should not be bounded by normal limits that (biologically) constrain human intelligence. Taking all previous evidence together, multifaceted, cross-domain mechanisms seem to form a cornerstone in modeling architectures for general intelligent systems.

[1] An argument is also given in [Hofstadter and the Fluid Analogies Research Group, 1996, Chapter 4] on the inability of models to lead to satisfactory understanding of the human mind if they separate conceptual processes from perceptual processes. Recall also Chalmers et al.'s argument, mentioned briefly in section 7.2.1, that perceptual processes cannot be separated from other cognitive processes even in principle.

Cross-Domain Mechanism:	**Sample GI Aspect:**
Analogical Mapping	Understanding new domains; creation and comprehension of metaphorical expressions and analogies in general
Analogical Transfer	Problem solving; introduction of new concepts into a domain; creative invention of new concepts
Generalization	Learning of abstract concepts; compression of knowledge
Specialization	Applying abstract knowledge in concrete situations; problem solving by realizing a general strategy
Re-representation	Adaptation of the input domains in order to compute appropriate analogies and blend spaces
Blending	Creation of new concepts and theories; problem solving; human style rationality; understanding of compound nouns; analysis of CFCs
Abduction	Finding explanations of certain facts

Table 9.1: Many desirable functions of intelligent systems can be explained by cross-domain reasoning mechanisms. The left column lists mechanisms, most of them are introduced in this thesis, and associates them with GI aspects that can be based on them. (Adapted from [Martínez et al., 2012, pp. 237].)

Therefore, the thesis stresses that next generation computational systems (that aim at achieving human-comparable intelligence) need to focus more on the utilization of as many cognitive mechanisms as possible. A wide range of theoretical, practical, formal, representational, and computational characterizations of such systems is still ahead, though.

Of course, there are still related issues that need to be more investigated or build upon the work presented in this thesis. The following adds to the conclusive remarks and perspective ideas spread over the preceding chapters:

1. It is crucial to focus more on aspects concerning knowledge representation of conceptual entities and their categorization into interrelated concepts. I assume that a computational manifestation of any aspect of human-comparable, intelligent thinking cannot be achieved without such focus. The focus should, moreover, facilitate the computational implementation of AGI systems that endow its under-

lying agents with a variety of GI aspects. KR is already a big research direction, but what AGI systems may suffer from is the lack of integrated, generalized KR frameworks and methodologies dedicated to overcoming the type of challenges encountered when utilizing higher-level cognitive mechanisms (e.g. in dealing with concept blending and representing nouns as concepts).

2. It should indisputably be clear how much weak CB is, with respect to its characterization and formalization aspects. As already highlighted in section 4.4, this weakness has at least two negative consequences: less appreciation of CB by the scientific community, and blocking further advancements. Therefore, future work necessarily has to provide more comprehensive characterization and formalization aspects of CB.

3. Integrating the previous two points would also be of a substantially great benefit. In particular, the ideas presented (on a conceptual level) in Chapters 7 and 8 still lack a glimpse of realization by recording and analyzing actual computations in cognitive models. This realization does seem feasible, but it necessitates a careful interaction between the issue of concept representation for cognition and formalization of CB aspects.

References

Abdel-Fattah, A. M. H. (2012). On the feasibility of concept blending in interpreting novel noun compounds. In *Proceedings of the Workshop "Computational Creativity, Concept Invention, and General Intelligence"*, volume 01-2012 of *Publications of the Institute of Cognitive Science (PICS)*, pages 27–32, Osnabrück, Germany. Institute of Cognitive Science. 29, 91, 136, 174

Abdel-Fattah, A. M. H., Besold, T. R., Gust, H., Krumnack, U., Schmidt, M., Kühnberger, K.-U., and Wang, P. (2012a). Rationality-Guided AGI as Cognitive Systems. In *Proc. of the 34th annual meeting of the Cognitive Science Society*, pages 1242–1247. 29, 109, 123, 155, 157, 173

Abdel-Fattah, A. M. H., Besold, T. R., and Kühnberger, K.-U. (2012b). Creativity, Cognitive Mechanisms, and Logic. In Bach et al. [2012], pages 1–10. 28, 91, 155, 157

Abdel-Fattah, A. M. H., El-Zahar, M. H., Ghaleb, F. M., and Khamis, S. M. (2005). Graph algorithms and their applications. Masters dissertation, Faculty of Science, Ain Shams University, Cairo. 138

Abdel-Fattah, A. M. H. and Krumnack, U. (2013). Creating Analogy-Based Interpretations of Blended Noun Concepts. In *AAAI Spring Symposium: Creativity and (Early) Cognitive Development*, volume SS-13-02 of *AAAI Technical Report*. AAAI. 29, 91, 136, 174

Abdel-Fattah, A. M. H., Krumnack, U., and Kühnberger, K.-U. (2013a). The Importance of Two Cognitive Mechanisms in Analyzing Counterfactuals: An Implementation-Oriented Explication. In *Advances in Cognitive Systems, Baltimore, Maryland, USA, December 12-14, 2013*. 29, 91, 164, 174

Abdel-Fattah, A. M. H., Krumnack, U., and Kühnberger, K.-U. (2013b). Utilizing Cognitive Mechanisms in the Analysis of Counterfactual Conditionals by AGI Systems. In Kühnberger et al. [2013], pages 1–10. 29, 91, 164, 174

Abdel-Fattah, A. M. H. and Schneider, S. (2013). Back-and-forth inception: Towards a cognitively-inspired model of concept construction via conceptual intellection. In *Proceedings of the Workshop "Computational Creativity, Concept Invention, and General Intelligence", C3GI at IJCAI 2013*. 19, 29, 35, 47, 75, 79, 104, 137, 172, 174

Adams, E. W. (1970). Subjunctive and Indicative Conditionals. *Foundations of Language*, 6(1):89–94. 153

Adams, E. W. (1975). *The Logic of Conditionals*. D. Reidel Publishing Co., Dordrecht. 153

Aerts, D. and Gabora, L. (2005a). A theory of concepts and their combinations i: The structure of the sets of contexts and properties. *Kybernetes*, 34:167–191. 121

Aerts, D. and Gabora, L. (2005b). A theory of concepts and their combinations ii: A hilbert space representation. *Kybernetes*, 34:192–221. 121

Agre, P. (1997). *Computation and Human Experience*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press. 11, 12, 139

Aho, A. and Ullman, J. (1995). *Foundations of Computer Science: C Edition*. Principles of Computer Science Series. W. H. Freeman. 10

Alexander, J. (2011). Blending in Mathematics. *Semiotica*, 2011(187):1–48. 91, 92, 109

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111:1036–1060. 19

Anderson, J. R. and Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum. 125

Argamon, S., Burns, K., and Dubnov, S. (2010). *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*. Springer. 89, 90, 188

Argand, J.-R. (1813). Philosophie mathématique. essay sur une manière de représenter les quantités imaginaires, dans les constructions géométriques. *Annales de Mathématiques pures et appliquées*, 4:133–146. 103

Bach, J., Goertzel, B., and Iklé, M., editors (2012). *Artificial General Intelligence - 5th International Conference, AGI 2012, Oxford, UK, December 8-11, 2012. Proceedings*, volume 7716 of *Lecture Notes in Computer Science*. Springer. 17, 179

Barnden, J. A., Glasbey, S., Lee, M. G., and Wallington, A. M. (2002). Reasoning in Metaphor Understanding: The ATT-Meta Approach and System. In *COLING*. 88

Bartha, P. (2013). Analogy and analogical reasoning. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Stanford University, fall 2013 edition. 31, 36

Baum, E., Hutter, M., and Kitzelmann, E., editors (2010). *Artificial General Intelligence*. Atlantis Press, Lugano, Switzerland. 17, 116, 117

Baum, W. M. (1994). *Understanding Behaviorism: Science, Behavior, and Culture*. Behavior analysis and society series. HarperCollins College Publishers. 5, 10

Becker, J. D. (1969). he Modeling of Simple Analogic and Inductive Processes in a Semantic Memory System. In *Proceedings of the 1st IJCAI*, pages 655–668, Washington, DC. 48, 50

Bello, P. (2012). Pretense and cognitive architecture. *Advances in Cognitive Systems*, 2:43–58. 164

Bermúdez, J. L. (2010). *Cognitive Science: An Introduction to the Science of the Mind*. Cambridge University Press. 6, 7, 9

Besold, T. R., Gust, H., Krumnack, U., Abdel-Fattah, A. M. H., Schmidt, M., and Kühnberger, K.-U. (2011). An argument for an analogical perspective on rationality & decision-making. In van Eijck, J. and Verbrugge, R., editors, *Proceedings of the Workshop on Reasoning About Other Minds: Logical and Cognitive Perspectives (RAOM-2011), Groningen, The Netherlands*, volume 751 of *CEUR Workshop Proceedings*, pages 20–31. CEUR-WS.org. 173

Blomberg, O. (2011). Conceptions of Cognition for Cognitive Engineering. *International Journal of Aviation Psychology*, 21(1):85–104. 10, 27

Boden, M. A. (1984). What is computational psychology? *Proceedings of the Aristotelian Society*, 58:17–35. 17

Boden, M. A. (1996). *Artificial Intelligence*. Handbook Of Perception And Cognition. Elsevier Science. 104

Boden, M. A. (2003). *The Creative Mind: Myths and Mechanisms*. Taylor & Francis. 109, 112

Boden, M. A. (2006a). *Mind As Machine: A History of Cognitive Science*. Number Bd. 1 in Mind as Machine: A History of Cognitive Science. Oxford University Press. 5, 7

Boden, M. A. (2006b). *Mind As Machine: A History of Cognitive Science*. Number Bd. 2 in Mind as Machine: A History of Cognitive Science. Oxford University Press. 5, 7

Bringsjord, S. and Licato, J. (2012). Psychometric Artificial General Intelligence: The Piaget-MacGuyver Room. In Wang and Goertzel [2012], chapter 3, pages 25–48. 15

Burki, L. and Cavalluci, D. (2011). Measuring the results of creative acts in r & d: Literature review and perspectives. In D. Cavalluci, R. de Guio, G. C., editor, *Building Innovation Pipelines through Computer-Aided Innovation, CAI 2011*, pages 163–177. Heidelberg: Springer. 102

Busemeyer, J. R. and Bruza, P. D. (2012). *Quantum Models of Cognition and Decision*. Cambridge University Press. 4, 121

Busemeyer, J. R., Franco, R., Pothos, E. M., and Trueblood, J. S. (2011). A Quantum Theoretical Explanation for Probability Judgment Errors. *Psychological Review*, 118(2):193–218. 121

Butnariu, C. and Veale, T. (2008). A concept-centered approach to noun-compound interpretation. In Scott, D. and Uszkoreit, H., editors, *COLING*, pages 81–88. 134, 135

Byrne, R. M. (2005). *The Rational Imagination: How People Create Alternatives To Reality*. Bradford Books. MIT Press. 152, 155, 156, 158

Campbell, N. R. (2013). *Physics: The Elements*. Cambridge University Press. 36

Carroll, L. (1869). *Alice's Adventures in Wonderland*. Mac Millan. 173

Chalmers, D. J., French, R. M., and Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, 4(3):185–211. Also appears as Chapter 4 in [Hofstadter and the Fluid Analogies Research Group, 1996, pp. 169–193]. 44, 45, 47, 137, 175

Chein, M. and Mugnier, M.-L. (2010). *Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs*. Advanced Information and Knowledge Processing. Springer. 139, 141

Cheng, P. W. and Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17:391–416. 38

Chomsky, N. (1956). Three Models for the Description of Language. *IRE Transactions on Information Theory*, 2:113–124. A version is available online at: http://www.chomsky.info/articles/195609--.pdf (last visited 31^{st} January 2014). 136, 150

Chomsky, N. (1957). *Syntactic structure*. The Hague/Paris: Mouton. 103

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19. 26

Clement, C. A. and Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15(1):89–132. 37, 38

Clement, J. J. (2008). *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Stimulation*. Springer London, Limited. 35, 36, 37, 39, 40, 41, 43, 62, 79, 80, 109

Collins, S. (2010). *The Hunger Games*. Scholastic. 173

Colton, S. (2011). The painting fool in new dimensions. In Show and Tell, editors, *Proceedings of the 2nd International Conference on Computational Creativity*. 112

Colton, S., Charnley, J., and Pease, A. (2011). Computational Creativity Theory: The FACE and IDEA Descriptive Models. In *In Second International Conference on Computational Creativity (ICCC2011)*. 104, 111, 112

Cook, V. J. and Newson, M. (2007). *Chomsky's Universal Grammar: An Introduction*. Wiley, 3rd edition. 136

Cosmides, L. and Tooby, J. (1993). Cognitive adaptations for social exchange. In J. H. Barkow and L. Cosmides and J. Tooby, editor, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pages 163–228. Oxford. 120, 128

Costello, F. J. and Keane, M. T. (2000). Efficient creativity: constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349. 135, 149

Coulson, S. (2006). *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge University Press. 76, 81, 85, 91, 131, 134, 135, 140, 156, 161

Davis, R., Shrobe, H., and Szolowits, P. (1993). What is a knowledge representation? *AI Magazine*, 14(1):17–33. 20, 22, 23

de Groot, A. D. (2008a). The Main Features of the Theory of Selz. In *Thought and Choice in Chess*, chapter 2B, pages 52–76. Amsterdam University Press, Amsterdam. 101

de Groot, A. D. (2008b). *Thought and Choice in Chess*. Amsterdam Academic Archive Series. Amsterdam University Press. 101

Diaconescu, R. (2008). *Institution-independent Model Theory*. Studies in Universal Logic. Birkhäuser, Basel. 88, 89

diSessa, A. A. (1988). Knowledge in pieces. In Forman, G. E. and Pufall, P. B., editors, *Constructivism in the computer age*, Communication Textbook, chapter 4, pages 49–70. Lawrence Erlbaum Associates, Mahwah, NJ. 20, 36, 139

Dorigo, M. and Stützle, T. (2004). *Ant Colony Optimization*. Bradford Books. MIT Press. 138, 142

Einstein, A. and Infeld, L. (1966). *The evolution of physics: from early concepts to relativity and quanta*. Touchstone books. Simon and Schuster. 36

Engesser, K., Gabbay, D. M., and Lehmann, D. (2007). *A New Approach to Quantum Logic*. Studies in Logic. College Publications. 4

Estes, Z. (2003). A tale of two similarities: comparison and integration in conceptual combination. *Cognitive Science*, 27:911–921. 131, 135, 150

Evans, J. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, 128:978–996. 118

Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. In *Proceedings of the April 21-23, 1964, spring joint computer conference*, AFIPS '64 (Spring), pages 327–338, New York, NY, USA. ACM. 32, 48, 49, 50, 52

Evans, T. G. (1968). A program for the solution of a class of geometric-analogy intelligence-test questions. In Minsky [1968], pages 271–353. 48, 52

Falkenhainer, B. C., Forbus, K. D., and Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41:1–63. 37, 39, 40, 50, 51, 52, 62, 65

Fauconnier, G. (1994). *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press. 75, 76, 89, 156

Fauconnier, G. (1997). *Mappings in Thought and Language*. Cambridge University Press. 75, 159

Fauconnier, G. and Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, 22(2):133–187. 74, 75, 88, 152, 172

Fauconnier, G. and Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, New York. 22, 74, 75, 76, 77, 80, 85, 86, 87, 88, 89, 92, 93, 94, 95, 96, 109, 152, 158, 161, 170, 172

Fauconnier, G. and Turner, M. (2008). Rethinking Metaphor. In Gibbs, R., editor, *Cambridge Handbook of Metaphor and Thought*, pages 53–66. Cambridge University Press, New York. 90

Feigenbaum, E., Feldman, J., and Armer, P. (1995). *Computers and Thought*. AAAI Press Series. AAAI Press. 13

Fellbaum, C. D. (2013). Obituary george a. miller. *Computational Linguistics*, 39(1):1–3. 11

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79. 15, 101, 115, 174

Fillmore, C. J. (1982). Frame semantics. In The Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, Seoul, Korea. 76, 136, 140

Finke, R. A., Ward, T. B., and Smith, S. M. (1992). *Creative Cognition: Theory, Research and Applications*. Bradford Books. Bradford Books. 109

Firth, J. R. (1957). *Papers in linguistics 1934–51*. Oxford University Press. 143

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Bradford Books. MIT Press. 136, 150

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford Cognitive Science Series. Clarendon Press. 80

Forbus, K. D. (2001). Exploring Analogy in the Large. In Gentner et al. [2001], pages 20–58. 51, 52

Forbus, K. D., Gentner, D., Everett, J. O., and Wu, M. (1997). Towards a Computational Model of Evaluating and Using Analogical Inferences. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, Cognitive Science Society (US) Conference/Proceedings, pages 229–234. Lawrence Erlbaum Associates. 36, 37, 42, 43

Forbus, K. D., Gentner, D., and Law, K. (1995). MAC/FAC: A Model of Similarity-Based Retrieval. *Cognitive Science*, 19(2):141–205. 39, 51, 52, 62

Frankish, K. and Ramsey, W. (2012). *The Cambridge Handbook of Cognitive Science*. The Cambridge Handbook of Cognitive Science. Cambridge University Press. 5, 7

French, R. M. (2002). The computational modeling of analogy-making. *Trends in Cognitive Sciences*, 6(5):200–205. 37, 39, 42, 45, 49, 50, 51, 52

Gagné, C. L. (2002). Lexical and relational influences on the processing of novel compounds. *Brain and Language*, 81:723 – 735. 134, 135, 147, 150

Gagné, C. L. and Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23(1):71–87. 134, 135, 136, 143, 150

Gärdenfors, P. (1988). *Knowledge in Flux : Modeling the Dynamics of Epistemic States*. MIT Press: Cambridge, Massachusetts. 20, 136, 137, 139

Gardner, H. (1987). *The Mind's New Science: A History of the Cognitive Revolution*. Psychology: History. Basic Books. 5, 6, 7, 9, 10, 11

Gay, L. and Croft, W. (1990). Interpreting nominal compounds for information retrieval. *Information Processing and Management*, 26(1):21–38. 133, 134

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7:155–170. 33, 36, 37, 38, 42, 43, 50, 51, 63

Gentner, D. (1989). The Mechanisms of Analogical Learning. In Vosniadou, S. and Ortony, A., editors, *Similarity and Analogical Reasoning*, chapter 7, pages 199–241. Academic Press. 34, 37, 44

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., and Forbus, K. D. (1997). Analogical Reasoning and Conceptual Change: A Case Study of Johannes Kepler. *The Journal of the Learning Sciences*, 6(1):3–40. 36, 39

Gentner, D. and Forbus, K. D. (1991). MAC/FAC: A Model of Similarity-Based Retrieval. In *Proceedings of the Cognitive Science Society*. 51

Gentner, D. and Forbus, K. D. (2011). Computational models of analogy. *WIREs Cogn Sci*, 2(3):266–276. 31, 38, 40, 42, 43, 44, 45, 54, 123

Gentner, D., Holyoak, K. J., and Kokinov, B. N., editors (2001). *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press. 35, 158, 186, 189, 191

Gentner, D., Rattermann, M. J., and Forbus, K. D. (1993). The Roles of Similarity in Transfer: Separating Retrievability from Inferential Soundness. *Cognitive Psychology*, 25:524–575. 34, 40, 43

Gentner, D. and Stevens, A. L. (1983). *Mental Models*. Cognitive Science. L. Erlbaum Associates. 156

Gick, M. L. and Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15(1):1–38. 43

Gigerenzer, G. (2005). I think, therefore I err. *Social Research*, 72(1):195–218. 121

Gigerenzer, G. (2010). *Rationality for mortals: how people cope with uncertainty*. Evolution and cognition. Oxford University Press. 118, 119

Goertzel, B. and Pennachin, C. (2010). *Artificial General Intelligence*. Cognitive Technologies. Springer. 117

Goertzel, B. and Wang, P. (2007). *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms - Proceedings of the AGI Workshop 2006*. Frontiers in Artificial Interlligence and Applications. IOS Press Inc. 16, 193

Goguen, J. A. (1999). An Introduction to Algebraic Semiotics, with Application to User Interface Design. In *Computation for Metaphors, Analogy, and Agents*, volume 1562 of *Lecture Notes in Computer Science*, pages 242–291. Springer. ix, 77, 88, 89, 90, 91, 96, 173

Goguen, J. A. (2005). What is a concept? In Dau, F., Mugnier, M.-L., and Stumme, G., editors, *ICCS*, volume 3596 of *Lecture Notes in Computer Science*, pages 52–77. Springer. 81, 89

Goguen, J. A. (2006). Mathematical Models of Cognitive Space and Time. In Andler, D., Ogawa, Y., Okada, M., and Watanabe, S., editors, *Reasoning and Cognition: Proc.*

of the Interdisciplinary Conference on Reasoning and Cognition, pages 125–128. Keio University Press. ix, 75, 77, 78, 79, 89, 90, 91, 92, 109

Goguen, J. A. and Harrell, D. F. (2004). Style as a Choice of Blending Principles. In Argamon, S., Dubnov, S., and Jupp, J., editors, *Style and Meaning in Language, Art Music and Design*, pages 49–56. AAAI Press. 89

Goguen, J. A. and Harrell, D. F. (2010). Style: A Computational and Conceptual Blending-Based Approach. In Argamon et al. [2010], pages 291–316. 89, 90

Goguen, J. A. and Malcolm, G. (1996). *Algebraic Semantics of Imperative Programs*. Foundations of Computing. MIT Press. 89

Goodman, N. (1947). The problem of counterfactual conditionals. *The Journal of Philosophy*, 44:113–118. 155, 162, 165

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In Sun, R., editor, *The Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press. 118

Guhe, M., Pease, A., Smaill, A., Martínez, M., Schmidt, M., Gust, H., Kühnberger, K.-U., and Krumnack, U. (2011). A computational account of conceptual blending in basic mathematics. *Cognitive Systems Research*, 12(3–4):249–265. 47, 59, 91, 105, 109, 111

Guhe, M., Pease, A., Smaill, A., Schmidt, M., Gust, H., Kühnberger, K.-U., and Krumnack, U. (2010). Mathematical reasoning with higher-order anti-unification. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1992–1997. 47, 62

Guilford, J. P. (1950). Creativity. *The American Psychologist*, 5(9):444–454. 100, 109

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill series in psychology. McGraw-Hill. 100

Gust, H., Krumnack, U., Martínez, M., Abdel-Fattah, A. M. H., Schmidt, M., and Kühnberger, K.-U. (2011). Rationality and General Intelligence. In Schmidhuber et al. [2011], pages 174–183. 29, 118, 149

Gust, H., Kühnberger, K.-U., and Schmid, U. (2006). Metaphors and Heuristic-Driven Theory Projection (HDTP). *Theoretical Computer Science*, 354:98–117. 35, 56, 60, 68

Gutiérrez, C. R. (2012). *Advances in Knowledge Representation*. InTech. 20

Hall, R. P. (1989). Computational Approaches to Analogical Reasoning: A Comparative Analysis. *AI*, 39:39–120. 42, 45, 48, 50, 52

Hampton, J. A. (1997). Conceptual combination. In Lamberts and Shanks [1997], pages 133–161. 134, 135, 136, 150

Harary, F. (1994). *Graph Theory*. Addison-Wesley series in mathematics. Perseus Books. 140, 147

Helman, D. H. (1988). *Analogical Reasoning: Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*. Synthese Library. Kluwer Academic Publishers Dordrecht. 35, 36

Hersh, R. (2011). From Counting to Quaternions – The Agonies and Ecstasies of the Student Repeat Those of D'Alembert and Hamilton. *Journal of Humanistic Mathematics*, 1(1):65–93. 91

Hesse, M. B. (1966). *Models and analogies in science*. University of Notre Dame Press. 36

Hofstadter, D. R. (1984). The Copycat Project: An Experiment in Non Determinism and Creative Analogies. *Massachusetts Institute of Technology*, 755. Also see Chapter 5 in [Hofstadter and the Fluid Analogies Research Group, 1996, pp. 205–268]. 42, 47, 53, 83, 112

Hofstadter, D. R. (1995). A review of mental leaps: Analogy in creative thought. *AI Magazine*, 16(3). 36

Hofstadter, D. R. (2001). Epilogue: Analogy as the core of cognition. In Gentner et al. [2001], pages 499–538. 36, 152, 157

Hofstadter, D. R. and Sander, E. (2013). *Surfaces and Essences: Analogy As the Fuel and Fire of Thinking*. Basic Books. BasicBooks. 31, 36

Hofstadter, D. R. and the Fluid Analogies Research Group (1996). *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Science-Psychology. Basic Books. 35, 37, 42, 49, 53, 66, 101, 137, 175, 182, 189

Hofstadter, D. R. and the Fluid Analogies Research Group (1996). *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Science-Psychology. Basic Books, New York. 157

Holding, D. H. (1985). *The psychology of chess skill*. L. Erlbaum Assoc Hillsdale, N.J. 100, 101

Holyoak, K. J. and Thagard, P. (1989). Analogical Mapping by Constraint Satisfaction. *Cognitive Science*, 13:295–355. 38, 39, 46

Holyoak, K. J. and Thagard, P. (1996). *Mental Leaps: Analogy in Creative Thought*. A Bradford book. Bradford Books. 35, 36, 39, 158

Hsu, F. (2002a). *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton Univ. 15, 101

Hsu, F. H. (2002b). *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton Univ. 174

Hummel, J. E. and Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3):427–466. 36

Hummel, J. E. and Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110:220–264. 106

Indurkhya, B. (1989). Modes of Analogy. In Jantke, K. P., editor, *Proceedings of the International Workshop on Analogical and Inductive Inference*, volume 397 of *Lecture Notes in Computer Science*, pages 217–230. Springer-Verlag. 32, 33, 34, 35

Indurkhya, B. (1992). *Metaphor and Cognition: An Interactionist Approach*. Mathematics and Its Applications. Springer. 35, 43, 62, 68

Jain, A. K., Mao, J., and Mohiuddin, K. (1996). Artificial Neural Networks: A Tutorial. *IEEE Computer*, 29(3):31–44. 46

James, W. (1950). *The Principles of Psychology*. Number Bd. 1 in Dover Books on Western Philosophy. Dover Publications. First published by Henry Holt & Co. in 1890. 36

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T. F., and Sageman, B. (2013). Finding faults: analogical comparison supports spatial concept learning in geoscience. *Cognitive Processing*, 14(2):175–187. 33, 34

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA. 79, 156

Johnson-Laird, P. N. (1988). *Cognitive science*. Cambridge University Press. 121

Johnson-Laird, P. N. (1995). *Mental Models, Deductive Reasoning, and the Brain*, pages 999–1008. MIT Press, 3 edition. 79

Johnson-Laird, P. N. and Byrne, R. M. (1991). *Deduction*. Erlbaum Hillsdale, NJ. 156

Joseph, S. (2011). *Coherence-Based Computational Agency*. PhD thesis, Universitat Autónoma de Barcelona. 96, 130

Keane, M. T. and Costello, F. J. (2001). Setting limits on analogy: Why conceptual combination is not structural alignment. In Gentner et al. [2001], pages 172–198. 131, 135, 149, 150

Kerber, M. (1989). Some Aspects of Analogy in Mathematical Reasoning. In Jantke, K. P., editor, *AII*, volume 397 of *Lecture Notes in Computer Science*, pages 231–242. Springer. 34

Keyser, S. J., Miller, G. A., and Walker, E. (1978). Report of the State of the Art Committee to the Advisors of the Alfred P. Sloan Foundation. An unpublished report submitted to the Alfred P. Sloan Foundation, New York. 7, 8, 9

Koestler, A. (1964). *The Act of Creation*. Arkana Series. Arkana. 74, 88, 172

Kokinov, B. N. (2003). Analogy in Decision-Making, Social Interaction, and Emergent Rationality. *Behavioral and Brain Sciences*, 26(2):167–168. 130, 157

Kokinov, B. N. and French, R. M. (2003). Computational Models of Analogy-Making. In *Encyclopedia of cognitive science*, volume 1, pages 113–118. Nature Publishing Group. 31, 34, 35, 39, 42, 43, 44, 45, 46, 47, 50, 51

Kokinov, B. N. and Petrov, A. (2001). Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In Gentner et al. [2001]. 43, 47, 106

Krumnack, U., Abdel-Fattah, A. M. H., and Kühnberger, K.-U. (2013a). Formal Magic for Analogies. In Kühnberger, K.-U., König, P., and Walter, S., editors, *Proceedings of the Workshop "Formalizing Mechanisms for Artificial General Intelligence and Cognition (Formal MAGiC)"*, volume 01-2013 of *Publications of the Institute of Cognitive Science (PICS)*, Osnabrück. Institute of Cognitive Science. 62

Krumnack, U., Gust, H., Schwering, A., and Kühnberger, K.-U. (2010). Remarks on the meaning of analogical relations. In Baum, E., Hutter, M., and Kitzelmann, E., editors, *Artificial General Intelligence, 3rd International Conference AGI*. Atlantis Press. 62

Krumnack, U., Schwering, A., Gust, H., and Kühnberger, K.-U. (2007). Restricted higher-order anti-unification for analogy making. In *Twenties Australian Joint Conference on Artificial Intelligence*, pages 273–282. Springer. 55, 56, 57, 58, 62, 107

Krumnack, U., Schwering, A., Kühnberger, K.-U., Gust, H., Abdel-Fattah, A. M. H., Besold, T., Schmidt, M., and Schneider, S. (2013b). Sketch learning by analogy. In Kutz, O., Bhatt, M., Borgo, S., and Santos, P., editors, *SHAPES*, volume 1007 of *CEUR Workshop Proceedings*, pages 49–58. CEUR-WS.org. 137

Kühnberger, K.-U., Rodolph, S., and Wang, P., editors (2013). *Artificial General Intelligence - 6th International Conference, AGI 2013, Beijing, China, 31 Juli – 3 August, 2013. Proceedings*, volume 7999 of *Lecture Notes in Computer Science*. Springer-Verlag Berlin Heidelberg. 179, 197

Laird, J. E. (2008). Extending the Soar Cognitive Architecture. *Frontiers in Artificial Intelligence and Applications*, 171. 19, 174

Laird, J. E. (2012). *The Soar Cognitive Architecture*. MIT Press. 19

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: An Architecture for General Intelligence. *Artificial Intelligence*, pages 1–64. 19, 125

Lakoff, G. and Johnsen, M. (2003). *Metaphors we live by*. The University of Chicago Press, London. 35

Lakoff, G. and Núñez, R. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. Basic Books, New York. 91

Lakoff, G. and Núñez, R. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. Basic Books, New York. 111

Lamberts, K. and Shanks, D., editors (1997). *Knowledge, Concepts, and Categories*. MIT Press. 80, 81, 139, 149, 189

Lee, M. (2010). Truth, metaphor and counterfactual meaning. In Burkhardt, A. and Nerlich, B., editors, *Tropical Truth(s): The Epistemology of Metaphor and other Tropes*, pages 123–136. De Gruyter. 88

Lee, M. G. and Barnden, J. A. (2001). A computational approach to conceptual blending within counterfactuals. Cognitive Science Research Papers CSRP-01-10, School of Computer Science, University of Birmingham. 88, 91, 155, 156, 159

Legg, S. and Hutter, M. (2007). A collection of definitions of intelligence. In Goertzel and Wang [2007], pages 17–24. 17

Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals*. Academic Press, New York. 134, 135, 150

Lewis, D. (2001). *Counterfactuals*. Library of philosophy and logic. Wiley. 152, 155, 156, 164, 170

Magnini, B. and Strapparava, C. (1990). Computational representation of mental spaces: A functional approach. In *ECAI*, pages 419–424. 75

Mareschal, D., Quinn, P. C., and Lea, S. E. G. (2010). *The Making of Human Concepts*. Oxford Series in Developmental Cognitive Neuroscience. Oxford University Press. 80, 81, 135, 139, 146, 149, 156

Martínez, M., Besold, T., Abdel-Fattah, A. M. H., Kühnberger, K.-U., Gust, H., Schmidt, M., and Krumnack, U. (2011). Towards a Domain-Independent Computational Framework for Theory Blending. In *AAAI Fall Symposium: Advances in Cognitive Systems*, volume FS-11-01 of *AAAI Technical Report*, pages 210–217. AAAI. 29, 47, 74, 105, 111, 152, 174

Martínez, M., Besold, T. R., Abdel-Fattah, A. M. H., Gust, H., Schmidt, M., Krumnack, U., and Kühnberger, K.-U. (2012). Theory blending as a framework for creativity in systems for general intelligence. In Wang and Goertzel [2012], chapter 12, pages 219–239. 47, 59, 62, 91, 92, 103, 105, 111, 172, 175, 176

McCarthy, J. (1998). What is artificial intelligence? 12, 13

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. Note that this article reproduces only the August 31, 1955 proposal. The original typescript consisted of 17 pages plus a title page, and is housed in the archives at Dartmouth College and Stanford University. 13, 15

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry Into the History and Prospects of Artificial Intelligence*. Ak Peters Series. A.K. Peters. 11, 13, 100

McCormack, J. and d'Inverno, M. (2012). *Computers and Creativity*. SpringerLink : Bücher. Springer-Verlag GmbH. 100, 104, 109

Medin, D. L. and Schaffer, M. M. (1978). Context Theory of Classification Learning. *Psychological Review*, 85(3):207–238. 82

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63(2):81–97. 11

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, 7(3):141–144. 5, 6, 7, 9, 11

Minsky, M. L. (1968). *Semantic Information Processing*. MIT Press. 49, 184

Minsky, M. L. (1974). A framework for representing knowledge. Technical report, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. 76, 140

Mitchell, M. (1993). *Analogy-Making As Perception: A Computer Model*. Neural Network Modeling and Connectionism. MIT Press, Cambridge, MA, USA. 36, 47, 83

Murphy, G. L. (2004). *The Big Book of Concepts*. Bradford Books. MIT Press. 21, 80, 81, 82, 83, 84, 139

Murphy, G. L. and Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, 92:289–316. 80, 82, 83, 84, 96

Nagai, Y. (2009). Concept blending and dissimilarity: factors for creative concept generation process. *Design Studies*, 30:648–675. 172

Newell, A., Shaw, J. C., and Simon, H. A. (1963). The process of creative thinking. In Gruber, H., Terrell, G., and Wertheimer, M., editors, *Contemporary Approaches to Creative Thinking*, pages 63–119. Atherton, New York. 5, 109, 112

Newell, A. and Simon, H. A. (1956). The Logic Theory Machine –A Complex Information Processing System. *IRE Transactions on Information Theory*, 2(3):61–79. 13, 100

Newell, A. and Simon, H. A. (1963). Gps, a program that simulates human thought. In Feigenbaum, E. and Feldmann, J., editors, *Computers and Thought*, pages 279–293. McGraw-Hill. 5, 49, 100

Newell, A. and Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ. 5

Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, 19(3):113–126. 5, 15, 19, 174

Nilsson, N. J. (1998). *Artificial Intelligence: A New Synthesis*. The Morgan Kaufmann Series in Artificial Intelligence Series. MORGAN KAUFMAN PUBL Incorporated. 13, 19

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press. 118

Pearl, J. (2011). The algorithmization of counterfactuals. *Annals of Mathematics and Artificial Intelligence*, 61(1):29–39. 153, 156, 157

Pease, A. and Colton, S. (2011). Computational Creativity Theory: Inspirations behind the FACE and the IDEA models. In *In Second International Conference on Computational Creativity (ICCC2011)*. 104, 111

Penrose, L. S. and Penrose, R. (1958). Impossible objects: A special type of visual illusion. *British J. of Psychology*, 49:31–33. 34

Pereira, F. C. (2007). *Creativity and Artificial Intelligence: A Conceptual Blending Approach, Applications of Cognitive Linguistics*. Mouton de Gruyter, Amsterdam. 74, 76, 77, 78, 82, 85, 88, 89, 90, 91, 92, 94, 95, 100

Pereira, F. C. and Cardoso, A. (2003). Optimality principles for conceptual blending: A first computational approach. *AISB Journal*, 1. 94

Pfeifer, N. (2008). A Probability Logical Interpretation of Fallacies. In Kreuzbauer, G., Gratzl, N., and Hiebl, E., editors, *Rhetorische Wissenschaft: Rede und Argumentation in Theorie und Praxis*, pages 225–244. LIT-Verlag. 121

Plotkin, G. D. (1970). A Note on Inductive Generalization. *Machine Intelligence*, 5:153–163. 54, 55

Plotkin, G. D. (1971). A Further Note on Inductive Generalization. *Machine Intelligence*, 6:101–124. 54, 55

Pólya, G. (1954). *Induction and analogy in mathematics*, volume 1 of *Mathematics and Plausible Reasoning*. Princeton University Press. 35

Poole, D. L., Mackworth, A. K., and GOEBEL, R. A. (1998). *Computational Intelligence: A Logical Approach*. Oxford University Press on Demand. 13, 19

Quine, W. V. (1960). *Word and Object*. MIT Press. 155, 162, 165

Ramsey, F. P. (1929). General propositions and causality. In Mellor, D. H., editor, *Philosophical Papers*, pages 145–153. Cambridge University Press. 157

Reitman, W. R., Grove, R. B., and Shoup, R. G. (1964). Argus: An Information-Processing Model of Thinking. *Behavioral Science*, 9(3):270–281. 32, 48, 49, 50

Reynolds, J. C. (1969). Transformational Systems and the Algebraic Structure of Atomic Formulas. In Meltzer, B. and Michie, D., editors, *Machine Intelligence*, volume 5, pages 135–153. Edinburgh University Press, Edinburgh, Scotland. 54

Rieskamp, J. and Reimer, T. (2007). Ecological rationality. In Baumeister, R. F. and Vohs, K. D., editors, *Encyclopedia of Social Psychology*, pages 274–276. SAGE Publications, Inc. 129

Rijsbergen, C. J. v. (2004). *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA. 121

Ringle, M. (1979). *Philosophical Perspectives in Artificial Intelligence*. Harvester Studies in Cognitive Science. Humanities Press. 48

Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104:192–223. 81

Rumelhart, D. E. and Norman, D. A. (1981). Analogical processes in learning. In Anderson, J. R., editor, *Cognitive Skills and Their Acquisition*, Carnegie Mellon Symposia on Cognition Series. Lawrence Erlbaum Associates, Hillsdale, NJ. 36

Runco, M. A. and Pritzker, S. R. (1999). *Encyclopedia of Creativity*. Number Volume 1 in Encyclopedia of Creativity. Academic Press. 99, 100, 104, 109

Russell, S. J. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice-Hall Series in Artificial Intelligence. Pearson Education/Prentice-Hall, 3rd edition. 12, 13, 14, 18, 19

Rutherford, E. (1911). The Scattering of α and β Particles by Matter and the structure of the atom. *Philosophical Magazine*, 21:669–688. 63

Ryder, M. E. (1994). *Ordered chaos: the interpretation of English noun-noun compounds*, volume 123 of *Linguistics*. University of California Press. 134

Santamaría, C., Espino, O., and Byrne, R. M. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *Journal of Experimental Psychology*, 31(5):1149–1154. 153, 156

Schank, R. (1975). *Conceptual Information Processing*. Fundamental Studies in Computer Science. North-Holland. 21

Schmidhuber, J., Thórisson, K. R., and Looks, M., editors (2011). *Artificial General Intelligence - 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, volume 6830 of *Lecture Notes in Computer Science*. Springer. 17, 188

Schmidt, M. (2010). *Strategies in Structural Alignment for Heuristic-Driven Theory Projection*. Master's thesis, University of Osnabrück. 53, 59

Schneider, S., Abdel-Fattah, A. M. H., Angerer, B., and Weber, F. (2013). Model Construction in General Intelligence. In Kühnberger et al. [2013], pages 109–118. 79, 137, 172

Schwering, A., Krumnack, U., Kühnberger, K.-U., and Gust, H. (2009a). Syntactic Principles of Heuristic-Driven Theory Projection. *Cognitive Systems Research*, 10(3):251–269. 40, 42, 43, 44, 45, 52, 53, 55, 56, 57, 58, 59, 60, 62, 63, 64, 65, 66, 106, 127, 160

Schwering, A., Kühnberger, K.-U., and Kokinov, B. N. (2009b). Analogies: Integrating Multiple Cognitive Abilities (Guest Editorial). *Special Issue on Analogies - Integrating Cognitive Abilities in Journal of Cognitive Systems Research*, 10(3):175–177. 35, 36

Schwering, A., Kühnberger, K.-U., Krumnack, U., Gust, H., Wandmacher, T., Indurkhya, B., and Ojha, A. (2009c). A computational model for visual metaphors. interpreting creative visual advertisements. In Filipe, J., Fred, A. L. N., and Sharp, B., editors, *ICAART*, pages 339–344. INSTICC Press. 105, 106

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–457. 101

Searle, J. R. (2011). Watson doesn't know it won on 'jeopardy!'. Wall Street Journal (online version: 23-Feb.-2011, last retrieved: 22-Feb.-2013). 101

Seckel, A. (2004). *Masters of Deception: Escher, Dalí & the Artists of Optical Illusion*. Sterling Publishing Company Incorporated. 35

Shoham, Y. and Leyton-Brown, K. (2009). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press. 153, 154

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In Parr, R. and van der Gaag, L. C., editors, *UAI*, pages 352–359. AUAI Press. 156

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118. 129

Skinner, B. F. (1984). The operational analysis of psychological terms. *Behavioral and Brain Sciences*, 7:547–553. 10

Smullyan, R. M. (1978). *What Is the Name of This Book?: The Riddle of Dracula and Other Logical Puzzles*. Prentice-Hall. ix, 73, 77, 78, 86, 87, 172

Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley. 139, 141

Sowa, J. F. (2011). Cognitive architectures for conceptual structures. In Andrews, S., Polovina, S., Hill, R., and Akhgar, B., editors, *ICCS*, volume 6828 of *Lecture Notes in Computer Science*, pages 35–49. Springer. 139, 141

Sowa, J. F. and Majumdar, A. K. (2003). Analogical reasoning. In de Moor, A., Lex, W., and Ganter, B., editors, *Proceedings of the 11th International Conference on Conceptual Structures (ICCS 2003)*, volume 2746 of *Lecture Notes in Computer Science*, pages 16–36. Springer. 35, 41, 141

Stalnaker, R. (1968). A theory of conditionals. In *Studies in Logical Theory*, volume 2, pages 98–112. Oxford: Blackwell. 155

Steels, L. (1993). The Artificial Life Roots of Artificial Intelligence. *Artif. Life*, 1(1–2):75–110. 14

Stenning, K. and van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Bradford Books. MIT Press. 121

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning : the componential analysis of human abilities*. Lawrence Erlbaum Associates. 35

Sternberg, R. J. and Lubart, T. I. (1999). *The Concept of Creativity: Prospects and Paradigms*, chapter 1, pages 3–15. Cambridge University Press, Cambridge. 100, 109

Stillings, N., Chase, C., Feinstein, M., Garfied, J. L., and Rissland, E. (1995). *Cognitive Science: An Introduction*. A Bradford book. MIT Press. 4, 7

Sun, R. (2008). Introduction to computational cognitive modeling. In *The Cambridge Handbook of Computational Psychology*, pages 3–20. Cambridge University Press. 18

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, 10(2):124–140. 18

Sun, R. and Ling, C. X. (1998). Computational cognitive modeling, the source of power, and other related issues. *AI Magazine*, 19(2):113–120. 18

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12:435–467. 96

Thagard, P. (2002). *Coherence in Thought and Action*. Life and Mind: Philosophical Issues in Biology and Psychology. MIT Press. 96, 124

Thagard, P. (2005). *Mind: Introduction to cognitive science*. MIT Press, Cambridge, MA, 2 edition. 5, 6, 7

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–265. 11

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59:433–460. One of the most influential papers in the history of the cognitive sciences: `http://cogsci.umn.edu/millennium/final.html`. 11, 13

Turner, M. and Fauconnier, G. (1998). *Conceptual Integration in Counterfactuals*, pages 285–296. CSLI Publications. Stanford CSLI. 74, 152

Turner, M. and Fauconnier, G. (2003). Metaphor, metonymy, and binding. In Dirven, R. and Pörings, R., editors, *Metonymy and Metaphor in Comparison and Contrast*, pages 469–487. Mouton de Gruyter. 162

Tversky, A. and Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90(4):293–315. 119, 120, 124

Veale, T. and O'Donoghue, D. (2000). Computation and Blending. *Computational Linguistics*, 11(3–4):253–282. Special Issue on Conceptual Blending. 88, 90, 91, 92

Veltman, F. (2005). Making counterfactual assumptions. *Journal of Semantics*, 22:159–180. 170

Wagman, M. (1996). *Human Intellect and Cognitive Science: Toward a General Unified Theory of Intelligence*. Praeger. 22

Wallas, G. (1926). *The art of thought*. C.A. Watts & Co. Ltd, London. 103, 109

Wang, P. (2004). Cognitive logic versus mathematical logic. 123

Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence*. Springer, Dordrecht. 123, 126, 141

Wang, P. (2007). The logic of intelligence. In *Artificial general intelligence*, Cognitive technologies, pages 31–62. Springer-Verlag Berlin Heidelberg. 141

Wang, P. (2008). What do you mean by "ai"? In Wang, P., Goertzel, B., and Franklin, S., editors, *AGI*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pages 362–373. IOS Press. 17

Wang, P. (2009). Formalization of Evidence: A Comparative Study. *Journal of Artificial General Intelligence*, 1:25–53. 123, 127

Wang, P. (2011). The Assumption on Knowledge and Resources in Models of Rationality. *International Journal of Machine Consciousness (IJMC)*, 3:193–218. 118, 123, 125, 126, 141

Wang, P. (2013). *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. World Scientific Publishing Company, Incorporated. 123, 126, 141

Wang, P. and Goertzel, B. (2012). *Theoretical Foundations of Artificial General Intelligence*, volume 4 of *Atlantis Thinking Machines*. Atlantis Press. 17, 29, 182, 193

Wang, P. and Hofstadter, D. R. (2006). A logic of categorization. *Journal of Experimental and Theoretical Artificial Intelligence*, 18(2):193–213. 126, 141

Wason, P. C. and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, 23:63–71. 119, 120

Weisberg, R. (1993). *Creativity: beyond the myth of genius*. Books in psychology. W H Freeman & Company. 109

Wichert, A. (2013). *Principles of Quantum Artificial Intelligence*. World Scientific Publishing Co. 4

Wilson, R. and Keil, F. (2001). *The MIT Encyclopedia of the Cognitive Sciences (MITECS)*. A Bradford book. MIT Press. 5

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review*, 4(2):167–183. 131, 134, 135, 136, 146, 149, 150

Wisniewski, E. J. and Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. In Simpson, G., editor, *Understanding Word and Sentence*. Elsevier Science Publishers B.V. (North-Holland). 131, 134, 135, 136, 143, 145, 146, 147, 149, 150

Wooldridge, M. and Jennings, N. R., editors (1995a). *Intelligent Agents, ECAI-94 Workshop on Agent Theories, Architectures, and Languages, Amsterdam, The Netherlands, August 8-9, 1994, Proceedings*, volume 890 of *Lecture Notes in Computer Science*. Springer. 13

Wooldridge, M. and Jennings, N. R. (1995b). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10(2):115–152. 13, 14

Wrobel, S. (1994). *Concept Formation and Knowledge Revision*. Kluwer. 81, 139

Zull, J. E. (2002). *The Art of Changing the Brain: Enriching Teaching by Exploring the Biology of Learning*. Stylus Pub. 139