Ludwig-Maximilians Universität München

&

Université Paris 1 Panthéon-Sorbonne

(Cotutelle de thèse / Binationale Promotion)

Inference and the structure of concepts

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philosophie der

Ludwig-Maximilians-Universität München

vorgelegt von

Matías Osta Vélez

aus

Montevideo, Uruguay

2020

Referent: Prof. Dr. Stephan Hartmann

Koreferent: Prof. Dr. Max Kistler

Tag der mündlichen Prüfung: 11.12.2020

L'Inférence et la structure des concepts

Thèse pour l'obtention du grade de Docteur en Philosophie de

l'Université Paris 1 Panthéon-Sorbonne

et

l'Université Ludwig-Maximilians de Munich

Présentée et soutenue publiquement par

Matías Osta Vélez

Institut d'Histoire et de Philosophie des Sciences et des Techniques

Munich Center for Mathematical Philosophy

Sous la supervision de:

Max Kistler	*IHPST, Université Panthéon-Sorbonne*
Stephan Hartmann	*MCMP, Ludwig-Maximilians Universität*

Composition du jury:

Igor Douven (rapporteur)	*Sciences Normes Décisions, CNRS, UMR 8011*
Paul Egré (rapporteur)	*Institut Jean Nicod, CNRS, UMR 8129*
Peter Gärdenfors	*Department of Philosophy, Lund University*
Philipp Koralus	*St Catherine's College, Oxford University*

Abstract

Inference and the structure of concepts

This thesis studies the role of conceptual content in inference and reasoning. The first two chapters offer a theoretical and historical overview of the relation between *inference* and *meaning* in philosophy and psychology. In particular, a critical analysis of the *formality thesis*, i.e., the idea that rational inference is a rule-based and topic-neutral mechanism, is advanced. The origins of this idea in logic and its influence in philosophy and cognitive psychology are discussed. Chapter 3 consists of an analysis of the relationship between inference and representation. It is argued that inference has to be studied from a pluralistic perspective due to its dependence on different formats of representing information. The following four chapters apply conceptual spaces, a formal theory of concepts within cognitive semantics, to three concept-based inference-types. First, an explication of Sellars' notion of material inference is advanced. Later, the model is extended to account for nonmonotonic inference by studying the role expectations in reasoning. Finally, a conceptual space-model of category-based induction is presented. This model predicts most of the empirical properties of this psychological phenomenon and subsumes some of the previous theories in psychology. It is stated that the explanatory fruitfulness of this new approach is evidence for the failure of the formality thesis and calls for a unified model of rational inference that puts semantics at center stage. The last chapter of the thesis discusses how inference and concepts interact in scientific reasoning, which makes constant use of hybrid symbolic structures for representing conceptual information. Stephen Toulmin's notions of *method of representation* and *inferential technique* are developed and applied in a case study about the emergence of the notion of instantaneous speed during the passage from geometrical physics to analytical mechanics. It is claimed that this analysis provides support to the pluralistic perspective for theorizing about reasoning.

Résumé

Cette thèse porte sur le rôle du contenu conceptuel dans l'inférence et le raisonnement. Les chapitres 1 et 2 offrent un aperçu théorique et historique de la relation entre "inférence" et "signification" en philosophie et en psychologie cognitive. En particulier, une analyse critique de la "thèse formaliste", i.e., l'idée selon laquelle l'inférence rationnelle est un mécanisme neutre par rapport au sujet (topic-neutral) et qu'il prend appui sur des règles syntaxiques. Les origines de cette idée dans la logique ainsi que son influence dans la philosophie et la psychologie cognitive sont discutées. Le Chapitre 3 porte sur la relation entre l'inférence et la représentation. Il est avancé que l'inférence doit être étudiée depuis une perspective pluraliste en raison de sa dépendance à l'égard de différents formats de représentation des informations qui caractérisent la cognition humaine. Les quatre chapitres suivants sont ceux de la mise en œuvre des espaces conceptuels, une théorie formelle des concepts au sein de la sémantique cognitive, à trois types d'inférence basés sur des concepts. Tout d'abord, une explication formelle de la notion d'inférence matérielle chez Wilfrid Sellars est avancée. Ensuite, le modèle est étendu pour saisir l'inférence non monotone en étudiant le rôle des "attentes" (expectations) dans le raisonnement. Enfin, un nouveau modèle mathématique d'induction avec des concepts (*category-based induction*) est présenté. Ce modèle prédit la plupart des propriétés empiriques de ce phénomène psychologique et fait quelques prédictions nouvelles. Il est indiqué que la fécondité explicative de cette approche novatrice montre l'échec de la thèse formaliste et appelle le développement d'un modèle unifié d'inférence rationnelle centré sur la sémantique. Le dernier chapitre de la thèse porte sur la manière dont l'inférence et les concepts interagissent dans le raisonnement scientifique, qui fait constamment appel à des structures symboliques hybrides pour représenter les informations conceptuelles. Les notions de “methods of

representation" et "inferential techniques," de Stephen Toulmin, sont développées et appliquées dans une étude de cas sur l'émergence de la notion de vitesse instantanée lors du passage de la physique géométrique à la mécanique analytique. On prétend que cette analyse soutient la perspective pluraliste pour théoriser sur le raisonnement.

Zusammenfassung

Reasoning (der psychologische Prozess, Schlussfolgerungen aus Daten oder Prämissen zu ziehen) und *Concepts* (im psychologischen Sinne: die Bedeutung von Wörtern, die sich in der menschlichen Denkweise darstellt), sind zwei zentrale Themen der Philosophie und der kognitiven Psychologie. Umso erstaunlicher ist es, dass sie traditionell als unabhängige Forschungsthemen verstanden werden und sich in der Literatur nur selten überschneiden. Das ist besonders verblüffend, wenn wir die weitgehende wissenschaftliche Einigkeit bedenken, die über die zentrale Rolle beider Begriffe für die Erklärung des menschlichen Denkens herrscht. Concepts sind historisch gesehen als ‚Bausteine' des Denkens konzipiert worden. Gleichzeitig bezieht sich Reasoning auf eine bestimmte Art von Gedankenübergängen, die bei rationalen Akteuren das Handeln und die Fixierung von Überzeugungen leiten sollen. Auf den ersten Blick scheinen die Begriffe untrennbar miteinander verbunden zu sein. Es stellt sich also die Frage, warum die Theorien des Reasonings, sowohl in der Psychologie als auch in der Philosophie den Begriff Concepts in ihrer Erklärungsstruktur vermeiden.

Eine mögliche Erklärung dafür ist, dass Studien zu Reasoning von einer logikwissenschaftlichen Sichtweise dominiert werden, die Inferenz als einen rein formal-syntaktischen Prozess —d. h. nicht-semantischen— versteht, der auf einer Reihe bereichsübergreifender und themenneutraler Regeln aufbaut. Aus dieser Sicht werden lexikalische Konzepte als ‚inferentiell inert' angesehen, d. h. sie spielen im Prozess der Inferenz und des Reasoning keine Rolle. Das ist keineswegs zufällig; im Gegenteil, es geht auf ein spezifisches Konstrukt des Begriffs der „logical form" ein, das die Logik seit Aristoteles dominiert (siehe Etchemendy, 1983). Nach dieser Auffassung ist inferentielle Gültigkeit eine Frage der Form und nicht des Inhalts. Deduktive Inferenzen sind gültig aufgrund der Verteilungen der logischen Begriffe, die ihre Struktur charakterisieren,

unabhängig von der Beziehung zwischen den außerlogischen (*extra-logical*) Begriffen (*Concepts*) in der/den Prämisse(n) und der Schlussfolgerung.

Diese Auffassung, ist in Inhelder und Piagets Behauptung zusammengefasst: "[human] reasoning is nothing more than the propositional calculus itself." (1958, S.305). Sie führte dazu, dass kognitive Psychologen und Philosophen das deduktive Reasoning als Paradigma der rationalen Inferenz und den semantischen Inhalt als irrelevant für die Erklärung des Reasoning betrachteten.

Parallel dazu wurde die philosophische Semantik von einer Sichtweise des Meaning (Sinn, Bedeutung) dominiert, die die Trennung zwischen Concepts und Reasoning bestätigte. Zu einem beträchtlichen Teil glaubten die Philosophen erklären zu können, was die lexikalische Bedeutung oder Satzbedeutung ist, ohne den Begriff der Inferenz - oder irgendeinen anderen Begriff, der sich auf einen kognitiven Mechanismus bezieht - einzuführen. Die Semantik wurde dann als etwas gedacht, das ausschließlich von der Beziehung zwischen Sprache und der Welt geleitet ist; und in diesem Sinne wurde sie auf die Begriffe der Referenz und der Wahrheitsbedingungen reduziert.

Ich bin davon überzeugt, dass diese Ideen grundlegend falsch sind und dass es keine Möglichkeit gibt, Reasoning ohne Concepts zu erklären und umgekehrt. Die vorliegende Arbeit ist ein Ansatz, diese Überzeugung zu rechtfertigen.

Ich bin nicht der erste, der davon überzeugt ist. Jonathan Evans, eine zentrale Figur auf dem Gebiet der Psychologie des logischen Denkens, schrieb vor Jahrzehnten, dass Concepts und Inferenz untrennbar miteinander verwoben („inextricably entangled") sind (Evans, 1989, S. 29). Er war davon überzeugt, dass Wissen - d.h. „bodies of concepts" - grundlegend für den Prozess des Reasoning selbst ist; und dass Theorien des Reasoning dies berücksichtigen müssen. Insbesondere behauptete er, dass Reasoning nicht „blind" sein könne, aber dies erfordere ein gewisses Maß an Verständnis des betreffenden Themas. Und da Verständnis den Besitz von Concepts voraussetze, könne es kein Reasoning ohne Concepts geben.

Im Verlauf dieser Arbeit werde ich eine ähnliche Idee auf unterschiedliche

Weise verteidigen. Ich behaupte, dass die Inferenz Eigenschaften von Repräsentationssystemen (representational systems) ausnutzt, die konzeptuelle Informationen kodieren. Es ist zu beachten, dass dies nicht im Widerspruch zu der logistischen Behauptung steht. Logiker glauben, dass die deduktive Inferenz die logische Form ausnutzt, und die logische Form eine (implizite) Eigenschaft der natürlichen Sprache ist —ein Repräsentationssystem. Ich glaube jedoch, dass die logische Inferenz in der alltäglichen Kognition eine eher marginale Rolle spielt; und dass die meisten sprachbasierten Inferenzen auf Eigenschaften der semantischen Repräsentation (semantic representation) aufbauen. Was ist nun ‚semantische Repräsentation'? Einer Tradition aus der kognitiven Semantik folgend, die der wahrheitsbedingten Semantik entgegengesetzt ist, gehe ich davon aus, dass diese Art der Repräsentation jene mentalen Strukturen betrifft, die durch lexikalische Konzepte während der Sprachverarbeitung evoziert werden.

Ein zentrales Ziel dieser Arbeit ist es, diese letztere Idee im Detail zu entwickeln. Ich verwende Peter Gärdenfors' Theorie der Conceptual Spaces (Begriffsräume) (2000; 2014), um verschiedene Formen semantisch basierter Inferenzen so zu erklären, dass sie der oben gegebenen Definition entsprechen. Ein Fazit dieser Analyse wird sein, dass Wortklassen aufgrund ihrer Abhängigkeit von verschiedenen Conceptual Spaces bei der Repräsentation während der semantischen Verarbeitung mit spezifischen Inferenzmustern assoziiert werden. Die Auffassung von Inferenz, die ich hier vertrete, ist pluralistisch. Die Idee dahinter ist einfach: Die menschliche Kognition verwendet viele verschiedene Arten von Repräsentationssystemen. Die natürliche Sprache ist wohl die wichtigste. Dennoch verwenden wir auch mentale Bilder und eine Fülle von äußeren Darstellungen wie Diagramme, mathematische Formeln und komplexe wissenschaftliche Modelle zur Darstellung von Phänomenen. Diese Systeme schreiben auch konzeptuelles Wissen fest, und wir verwenden sie beim Schlussfolgern durch Inferenzmechanismen, die für sie typische Eigenschaften ausnutzen. Um diese Behauptung zu untermauern, ist das letzte Kapitel dieser Arbeit dem Studium des modellbasierten logischen Denkens in der Wissenschaft gewidmet; insbesondere der Beziehung zwischen wissenschaftlichen Konzepten, Modellen und Reasoning.

Bevor die Struktur dieser Arbeit erläutert wird, sind zwei Dinge wichtig. Erstens, die Arbeit war ursprünglich als eine Sammlung von Artikeln gedacht, die sich um das zu Beginn dieser Einführung beschriebene Problem drehen. Die letzten Kapitel (Kapitel 4-8) basieren auf diesen Artikeln. Die vorangehenden Kapitel (Kapitel 1-3) zielen darauf ab, den gemeinsamen theoretischen Rahmen aus historischer und philosophischer Sicht zu setzen. Nichtsdestotrotz enthalten diese Kapitel einiges an Kritik an den diskutierten Ansichten sowie einige Ideen, wie man die Beziehung zwischen Inferenz und Repräsentation auflösen kann. Zweitens, bauen die Kapitel 6 und 7 direkt auf zwei Kollaborationen mit Peter Gärdenfors auf. Kapitel 7 basiert auf dem Artikel „Category-based induction in conceptual spaces", der in dem Journal for Mathematical Psychology (Osta-Vélez & Gärdenfors, 2020a) veröffentlicht wurde; während Kapitel 6 auf dem Artikel „Nonmonotonic reasoning, expectation orderings, and conceptual spaces“ (Osta-Vélez & Gärdenfors, n.d.).

Die vorliegende Arbeit ist wie folgt strukturiert. In Kapitel 1 wird die These der Formalität (formality thesis), eine entscheidende Idee für viele Theorien des Reasoning, erörtert, wobei behauptet wird, dass Inferenz ein formaler Prozess ist, der auf einer satzähnlichen Gedankensprache beruht. Hier soll gezeigt werden, wie diese Idee in der aus der klassischen Logik übernommenen begrifflichen Unterscheidung zwischen Form und Inhalt verwurzelt ist. Danach wird der Einfluss dieser Unterscheidung auf die Grundlagen der Kognitionswissenschaft diskutiert, wobei ein besonderer Schwerpunkt auf Jerry Fodors einflussreiche Version der rechnergestützten Theory of Mind gelegt wird. Der letzte Teil ist der Rolle der formality thesis in der kognitiven Psychologie gewidmet. Drei klassische Theorien werden kritisch diskutiert: Piagets entwicklungsorientierte Betrachtung von Reasoning, die Mental Logic Theorie und die Mental Model Theorie.

Kapitel 2 ist den Theories of Meaning gewidmet. Es beginnt mit einer historischen Analyse der Beziehung zwischen Meaning und Inferenz in der allgemeinen Sprachphilosophie. Ziel ist es, den philosophischen Rahmen zu erörtern, der eine ‚Trennung‘ zwischen diesen beiden Begriffen förderte. Danach werden

einige alternative Theorien analysiert, insbesondere die begriffliche Rollensemantik (Conceptual Role Semantics) und der Inferentialismus. Ich komme zu dem Schluss, dass diese Theorien nicht gut geeignet sind, um die Beziehung zwischen Meaning und Inferenz zu erklären. Schließlich wird die Kognitive Semantik eingeführt und als der geeignete Rahmen für die Durchführung der oben genannten Aufgabe dargestellt.

Kapitel 3 analysiert die Beziehung zwischen Repräsentation und Inferenz. Es wird behauptet, dass produktive Inferenzmechanismen voraussetzen, dass konzeptuelle Information vorher geordnet wird; und dass die rechnerischer Effizienz dieser Mechanismen vom Format der Informationsdarstellung abhängt. Ein wichtiger Teil des Kapitels ist eine Kritik an dem, was ich als Repräsentationskonservativismus bezeichne, d.h. die Idee, dass alle Überlegungen auf persönlicher Ebene in einem sprachähnlichen Repräsentationssystem wie Fodors Language of Thought stattfinden. Anschließend wird der inferentielle Pluralismus verteidigt, und es werden einige Beispiele für verschiedene Inferenz-Typen diskutiert.

Während die vorhergehenden Kapitel überwiegend kritisch und theoretisch sind, ist der restliche Teil der Arbeit eher konstruktiv. Es soll gezeigt werden, wie bestimmte Formen von Inferenzen, die aus formalistischer Perspektive schwer zu erklären sind, leicht erläutert werden können, wenn man die Abhängigkeitsbeziehung zwischen Konzepten und Inferenz annimmt. Kapitel 4 führt in die Theorie der Conceptual Spaces ein, das wichtigste formale Werkzeug, das in den folgenden Kapiteln verwendet wird. Kapitel 5 führt eine Erläuterung von Wilfrid Sellars' Begriff der Materiellen Inferenz unter Verwendung von Conceptual Spaces ein. Es wird behauptet, dass, obwohl Sellars' und Brandoms Inferentialismus einen wesentlichen Gebrauch von der Idee der Materiellen Inferenz macht, keiner dieser Autoren eine Erklärung der kognitiven Ursprünge und der dahinter stehenden Mechanismen anbietet. Ich veranschauliche die Analyse Conceptual Spaces, indem ich die inferentiellen Möglichkeiten verschiedener Wortklassen untersuche, insbesondere Substantive, Adjektive, räumliche Präpositionen und einige relationale Konzepte.

Kapitel 6 befasst sich mit nicht-monotischem Denken. Die Hauptthese ist,

dass Schlussfolgerungen unter Unsicherheit auf der Struktur des Hintergrundwissens aufbauen. Das Kapitel baut auf Ideen von Gärdenfors und Makinson (1992, 1994) über die Rolle von Erwartungen beim Reasoning auf. In groben Zügen zeigten sie, dass sich die nicht-monotone Logik auf die klassische Logik sowie eine Ordnung von Sätzen, welche Erwartungen repräsentieren, reduzieren lässt. Es wird argumentiert, dass ihr Rahmenkonzept durch eine auf Conceptual Spaces basierende Analyse der Erwartungen erheblich bereichert werden kann. Insbesondere können die in Conceptual Spaces eingebauten Distanzfunktionen dazu verwendet werden, psychologisch realistische Erwartungsordnungen zu erzeugen, die dazu beitragen, die Dynamik des nicht-monotonen Denkens zu erklären. Gleichzeitig löst diese Analyse einige alte erkenntnistheoretische Probleme der Standardlogik.

In Kapitel 7 wird ein neues mathematisches Modell für kategoriebasierte Induktion vorgeschlagen, eine Art semantisch-basierte Inferenz, die hauptsächlich in der Psychologie untersucht wird. Dieser Inferenzmechanismus nutzt die Kenntnis konzeptueller Beziehungen, um abzuschätzen, wie wahrscheinlich es ist, dass eine Eigenschaft von einer Kategorie auf eine andere projiziert wird. Zum Beispiel wird die Schlussfolgerung ‚Hunde haben Sesambeine; daher haben Wölfe Sesambeine' als stärker empfunden als ‚Hunde haben Sesambeine; daher haben Wale Sesambeine'. Weil Wölfe Hunden ähnlicher sind als Wale. Es wird gezeigt, dass ein Conceptual-Spaces-Modell dieser Art von Induktion die meisten ihrer empirischen Eigenschaften vorhersagen kann und einige neue Vorhersagen machen kann. Am Ende des Kapitels werden die Beziehungen zu anderen Modellen und einige methodologische Ideen diskutiert.

Schließlich lässt Kapitel 8 sprachbasierte Inferenzen beiseite und konzentriert sich auf die Beziehung zwischen Reasoning und wissenschaftlichen Modellen. Aufbauend auf Stephen Toulmins Begriffen der Method of Representation und Inferential Technique (Toulmin, 1972a) analysiere ich die Rolle hybrider symbolischer Strukturen, wie beispielsweise Modelle bei der Zusammensetzung bestimmter Formen des Reasoning zu Phänomenen in der Wissenschaft. Die Analyse wird durch eine Fallstudie über die Entwicklung des Begriffs der augenblicklichen Geschwindigkeit von der geometrischen zur analytischen Mechanik

unterstützt. Die hier verteidigten Ideen werden als Beweis für inferentiellen Pluralismus und für die enge Verbindung zwischen Darstellungsformen (forms of representation) und Formen des Reasoning herangezogen.

Declaration of Authorship

I, Matías Osta Vélez, confirm that:

- This work was done wholly or mainly while in candidature for a joint degree at Université Paris I Panthéon-Sorbonne and Ludwig-Maximilians Universität München.
- Where I have consulted the published work of others, this is always clearly attributed.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have stated it clearly.

Signed:

_____

Date: 12.10.2020

_____

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisors, Max Kistler and Stephan Hartmann. Max has been helping me immensely since the early days of my master's studies in Paris. I have benefited greatly from his lectures, advice, and the many discussions we had during the last years. Stephan has been a source of constant support and guidance over the past three years. His insightful comments and suggestions have significantly improved the quality of my work. Besides my supervisors, I'm deeply indebted to Peter Gärdenfors. I have been reading Peter's work since my undergraduate years. I started collaborating with him after visiting Lund University in April 2019; since then, I have learned tremendously from his vast knowledge, creativity, and generosity. Two chapters of this thesis result from this collaboration. Yet, Peter's ideas permeate this dissertation through and through.

I am deeply grateful to Maite Rodríguez Apólito for her support and kindness over many years. Without her, this thesis would not have been possible. I am also indebted to Denis Merklen, for his friendship, advice, and guidance. Denis and his family have been a constant source of affection since my very first day in Paris.

Furthermore, I would like to thank my dear friends and colleagues at the University of the Republic, in Uruguay, for their sustained support during my studies. A special thank to Alejandro Chmiel, Guillermo Nigro, Washington Morales, Horacio Lena, Paulo Cardozo, Lucía Lewowicz, José Seoane, María Laura Martínez, and Juan Queijo.

During my studies in Paris, I have benefited from the stimulating environment of the IHPST and the DEC. I am especially grateful to my colleagues and friends Marina Imocrante, Marie Michon, Fernando Valenzuela, Karina Mendez, Marco Casali, Sophia Rousseau-Mermans, Caroline Angleraux, Armano Lavalle, and Gianni Gestaldi. I am also thankful to many other members of the IHPST who have helped me during that time. Among these are Pierre Wagner, Marco Panza, David Waszek, Victor Lefèvre, Alireza Bani Sadr, Cecilia Bognon, Henri Salha, Gaelle Pontarotti, Daniel Kostic, and Méven Cadet.

I spent half of my PhD studies at the Munich Center for Mathematical Philosophy, and I am very grateful to all its members for providing such a friendly and stimulating work environment. I particularly thank Elio La Rosa, Sena Bozdag, Marco Forgioni, Gianluca Grilletti, Borut Trpin, Mel Ludwig, Naftali Weinberger, Alessandra Marra, Timo Freiesleben, Killian McGrath, and Maria Csauscher. My special thanks go to Michal Hladky, Matteo de Benedetto, and Lorenzo Rossi, for countless hours of exciting and amusing conversations.

I would also like to express my gratitude to the members of the examination committee, Igor Douven, Paul Egré, Philipp Koralus, and once again, to Peter Gärdenfors, as well as to the many other researchers who had read or discuss with me the ideas developed in this dissertation. Among these are Hugo Mercier, James Hampton, Mauricio Suárez, and Salvador Mascarenhas. I am especially grateful to Lara Kirfel for her encouragement and efforts to curb my ignorance about experimental psychology.

I especially want to thank Lena Peichl for her kindness, help, constant support, and patience during this intense past year in Germany.

Last but certainly not least, I want to thank my family for their unconditional support and affection. To my mother, my father, Lody, Pablo, and Hebe. This work is dedicated to them.

During my PhD studies, I was generously supported by the National Agency for Research and Innovation (ANII) and Campus France, by the Deutscher Akademischer Austauschdienst (DAAD), and by the Faculty of Humanities of the University of the Republic.

Co-authored and single-authored publications

Chapters 6 and Chapter 7 are joint work with Peter Gärdenfors. The content of Chapter 6 comes from the paper "Nonmonotonic reasoning, expectation orderings, and conceptual spaces," (Osta-Vélez & Gärdenfors, n.d., submitted). Chapter 7 is based on the article "Category-based induction in conceptual spaces," published in *Journal of Mathematical Psychology*, vol. 96, 102357, (Osta-Vélez & Gärdenfors, 2020a). I take full responsibility for the way that material is presented in this dissertation.

Chapter 8 is based on my article "Methods of representation as inferential devices," published in *Journal for General Philosophy of Science*, vol. 50, pp. 231–245, (Osta-Vélez, 2019). All other parts of this thesis were written solely by myself.

Contents

List of Figures

List of Abbreviations

CRS	Conceptual Role Semantics
CS	Conceptual Spaces
CTM	Computational Theory of Mind
IRS	Inferential Role Semantics
LOT	Language of Thought
MBRA	Model-Based Reasoning Approach
MLT	Mental Logic Theory
MMT	Mental Model Theory
SCM	Similarity Coverage Model
TC	Typicality Criterion

List of Symbols

$\mathscr{C}(M)$	Conceptual space of concept M.
$C(X_1 \cup \cdots \cup X_n)$	Convex Hull of categories $X_1 \cdots X_n$.
$d(x, y)$	Distance between x and y.
$Exp(M)$	Set of expected properties of an object falling under concept M.
$ExpS(Y \rightarrow Y)_Z$	Expectations that property S is projected from X to Y, with $X \subseteq Z$ and $Y \subseteq Z$.
p^M	Prototype of concept M.
$sim(x, y)$	Similarity among concepts x and y.
$V(M)$	Volume of concept M.

Introduction

Reasoning and concepts are two central issues in philosophy and cognitive psychology. Surprisingly enough, they have traditionally been understood as independent research topics, and they rarely intersect in the literature. That is particularly baffling if we consider the widespread agreement on both notions' essential role in explaining human thinking. *Concepts* have been historically conceived as the "building blocks" of thoughts. At the same time, *reasoning* refers to a particular kind of thought-transitions which is supposed to guide action and belief fixation in rational agents. At first glance, the notions seem to be intrinsically connected. The question is, then, why do theories of reasoning, both in psychology and in philosophy, obviate the notion of concept in their explanatory structure?

One possible explanation for this is that reasoning studies have been dominated by a *logicist* approach that understands *inference* as a purely formal-syntactic process —i.e., non-semantic— which builds on some set of domain-general and topic-neutral rules. From that point of view, lexical concepts are considered "*inferentially inert*," i.e., not playing any role in the process of inference. That is by no means casual; on the contrary, it responds to a specific construal of the notion of "logical form" that has dominated logic since Aristotle (see Etchemendy, 1983). In this view, inferential validity is a matter of *form*, and not of *content*. Deductive inferences are valid in virtue of the distributions of logical terms characterizing their structure, regardless of the relation between the extra-logical terms (concepts) in the premises(s) and the conclusion.

This view, summarized in Inhelder and Piaget's claim that "[human] reasoning is nothing more than the propositional calculus itself." (1958, p. 305), led cognitive psychologists and philosophers to consider deductive reasoning as the paradigm of rational inference, and to see semantic content as irrelevant for the explanation of reasoning.

In parallel to this, philosophical semantics was dominated by a view of *meaning* that confirmed the disconnection between concepts and reasoning. To a significant extent, philosophers thought they could explain what lexical or sentential meaning was, without introducing the notion of inference —or any other notion referring to a cognitive mechanism. Semantics was then thought as something exclusively concerned by the relation between language and the world; and, in this sense, reduced to the notions of *reference* and *truth-conditions*.

I am persuaded that these ideas are fundamentally wrong and that there is no way to explain reasoning without concepts and vice versa. This whole thesis is an attempt to justify this conviction.

I am not the first to believe so. Jonathan Evans, a central figure in the field of the psychology of reasoning, wrote decades ago that concepts and inference are "*inextricably entangled*" (Evans, 1989, p. 29). He was convinced that *knowledge* —i.e., "bodies of concepts"— is constitutive of the very process of reasoning; and that psychological theories must account for this. In particular, he claimed that reasoning could not be "blind," but that requires some degree of understanding of the topic in question; and since understanding requires concept possession, there is no reasoning without concepts.

In a similar vein, Hugo Mercier and Dan Sperber have recently developed a comprehensive theory of reasoning that seeks to explain its social, individual, and evolutionary dimensions (Mercier & Sperber, 2017). A fundamental idea behind it is that inferential mechanisms are meant to exploit empirical regularities in the environment that are codified in representational systems of different sorts. Their theory is essentially anti-formalist, and understands inference as *inextricably entangled* with representation. However, it does not explain how this entanglement would work, nor does it explains how conceptual information is structured within representation systems. In fact, —and once again— their approach overlooks the notion of *concept* (see Osta-Vélez, 2019).

Throughout this thesis, I will defend —in different ways— ideas which are similar to those mentioned above. My focus will be on the point Mercier's and Sperber's theory left unexplained, i.e., the issue of how inferential mechanisms are entangled to representational structures. In particular, I claim that inference

exploits different properties of the representational systems used by everyday cognition for encoding conceptual information.

Notice that this is not in contradiction to the logicist claims. Logicists believe that deductive inference exploits logical form, and logical form is an (implicit) property of natural language, i.e., of a representational system. However, I believe that logical inference plays a rather marginal role in everyday cognition and that most language-based inferences build on properties of semantic representation. Now, what is "semantic representation"? Following a tradition from cognitive semantics opposed to truth-conditional approaches, I assume that this kind of representation concerns the mental structures evoked by lexical concepts during language processing.

A central aim of this thesis is to develop this latter idea in some detail. I use Peter Gärdenfors' theory of conceptual spaces (2000; 2014) to explicate different forms of semantic-based inferences in such a way that fits the definition given above.

Conceptual Spaces is a research program in cognitive science and knowledge representation claiming that conceptual content is organized in different topological and geometrical structures at a sub-symbolic level of representation of information. It provides many formal and theoretical tools for explaining how concepts are used in cognitive processes like categorization, induction, concept formation, or language learning.

In this dissertation, I intend to show how this approach can be successfully used for explaining the role of concepts in reasoning. Regarding semantic-based inference, an upshot of the analysis here developed will be that word classes have specific inferential patterns associated with them due to their reliance on different conceptual spaces while being represented during semantic processing. Furthermore, it will be shown how inductive inference and nonmonotonic reasoning rely on very specific properties of conceptual structures like similarity and typicality.

The view of inference defended here is *pluralistic*. The idea behind it is straightforward: human cognition uses many different kinds of representational systems. Natural language is, arguably, the most important one. Still, we also

use mental images and a plethora of external symbolic structures like diagrams, mathematical formulas, and scientific models to represent phenomena. These systems also codify conceptual knowledge, and we use them in reasoning through inferential mechanisms that exploit properties that are typical of them. To offer some support for this claim, the last chapter of this work is devoted to studying model-based reasoning in science; and, in particular, the relationship between scientific concepts, models, and reasoning.

Before going on to explain the structure of this work, two things are important to note. First, the content of several chapters has already been published in the form of articles. In particular, chapters six and seven build on two collaborations with Peter Gärdenfors (Osta-Vélez & Gärdenfors, n.d., 2020a); while chapter eight is based on (Osta-Vélez, 2019). Second, the thesis unfolds in two stages. The first three chapters offer a critical analysis of the historical and philosophical framework motivating this work; and they are supposed be the "glue" connecting the rest of the content. On the other hand, the last five chapters are rather constructive and propose different ways in which the issues that this dissertation attends can be approached.

The thesis is structured in the following way. Chapter 1 discusses the *formality thesis*, a crucial idea for many theories of reasoning, claiming that inference is a formal process carried out on a sentence-like language of thought. Here, the aim is to show how this idea is rooted in the conceptual distinction between *form*, and *content* inherited from classical logic. After this, the influence of this distinction in the foundations of cognitive science is discussed, with a particular focus on Jerry Fodor's influential version of the computational theory of mind. The last part is devoted to the role of the formality thesis in cognitive psychology. Three classical theories are critically discussed: Piaget's developmental view of reasoning; Mental Logic theory, and Mental Model theory.

Chapter 2 is devoted to theories of meaning. It begins with a historical analysis of the relation between *meaning* and *inference* in mainstream philosophy of language. The aim is to discuss the philosophical framework which promoted a "divorce" between these two notions. Afterward, some alternative theories

are analyzed, in particular conceptual role semantics and inferentialism. I conclude that these theories are not well equipped for explaining the relationship between meaning and inference. At the end, cognitive semantics is introduced and defended as the appropriate framework for doing the aforementioned job.

Chapter 3 is an analysis of the relation between representation and inference. It is claimed that "productive" inferential mechanisms require that conceptual information is previously organized; and that the computational efficiency of these mechanisms depends on the format of representation of information. An important part of the chapter is a criticism of what I call *representational conservativism*, i.e., the idea that all personal-level reasoning takes place in a language-like representational system like Fodor's *language of thought*. Finally, inferential pluralism is defended, and some examples of different inference-types are discussed.

While the previous chapters are mostly critical and theoretical, the remainder of the thesis is rather constructive. It aims at showing how certain forms of inferences —that are difficult to explain from a formalist perspective— are easily explicated if we assume the dependency relation between concepts, representation, and inference. Chapter 4 introduces the theory of conceptual spaces, the main formal tool used in the next chapters. Chapter 5 advances an explication of Wilfrid Sellars' notion of *material inference* using the aforementioned theory. It is claimed that while Sellars' and Brandom's views makes essential use of the idea of material inference, none of these authors offers an explanation of the cognitive origins and mechanisms behind it. I exemplify the conceptual space-analysis by studying the inferential affordances of various word classes; notably, nouns, spatial prepositions, and some relational concepts.

Chapter 6 deals with nonmonotic reasoning. The main claim is that inference under uncertainty builds on the structure of background knowledge. The chapter builds on ideas from Gärdenfors and Makinson (1992; 1994) regarding the role of expectations in reasoning. Roughly, they showed that nonmonotonic logic can be reduced to classical logic plus an ordering of propositions representing expectations. It is argued that their framework can be significantly enriched by a conceptual spaces-based analysis of expectations. In particular,

the built-in distance functions in conceptual spaces can be used to generate psychologically realistic expectation orderings that help to account for the dynamics of nonmonotonic reasoning. At the same time, this analysis solves some old epistemological issues of default logic.

Chapter 7 propose a new mathematical model for category-based induction, a kind of semantic-based inference mainly studied in psychology. This inferential mechanism uses knowledge of conceptual relations to estimate how likely it is for a property to be projected from one category to another. For instance, the inference "Dogs have sesamoid bones; thus wolves have sesamoid bones" is perceived as stronger than "Dogs have sesamoid bones; thus, whales have sesamoid bones" because wolves are more similar to dogs than whales. It will be shown that a conceptual spaces-model of this kind of induction can predict most of its empirical properties and make some new predictions. At the end of the chapter, the relations with other models and some methodological ideas will be discussed.

Chapter 8 leaves aside language-based inferences and focuses on the relationship between reasoning and scientific models. Building on Stephen Toulmin's notions of *method of representation* and *inferential technique* (Toulmin, 1972a), I analyze the role of hybrid symbolic structures in the constitution of particular forms of reasoning about phenomena in science. The analysis is supported by a case study about the development of the notion of instantaneous speed from geometrical to analytical mechanics. The ideas defended here are taken as evidence for inferential pluralism and for the intimate connection between forms of representation and forms of reasoning.

Finally, Chapter 9 summarizes the main ideas defended in the dissertation and points out various research lines that open up from them.

Chapter 1

Formality: reasoning without meaning

Summary

In this chapter, I discuss the *formality thesis*, a crucial idea for many theories of reasoning claiming that inference is a formal process carried out on a sentence-like language of thought. I aim to show how this idea is rooted in the conceptual distinction between form and content inherited from classical logic. And to discuss its interpretation in the foundations of cognitive science and the psychology of reasoning. I will analyze the limitations of this thesis and its contribution to the influential idea that reasoning can be studied from a syntactic perspective, without considering any semantic notion.

1.1 Logic, cognition, and the formality thesis

Logic has been historically conceived as central to reasoning. It has played an essential part in the development of cognitive science in general, and theories of reasoning in particular (Harman, 1984; Henle, 1962). It has been used as a competence model for deductive reasoning (Overton, 1990); as a normative framework for evaluating our performance in reasoning tasks (Osherson, 1975b; Stenning & van Lambalgen, 2011); or as a methodological tool for modeling the formal structure of high-level cognitive operations (Piaget, 1957). In general, the various forms in which logic has influenced the study of reasoning over the

years share an underlying assumption: logical properties are formal-syntactic properties and, if human inference is logical, then it must be formal in some similar way.

That idea is part of a long tradition of conceiving the mind as a machine/computer. It started with Thomas Hobbes, and was further developed by George Boole, Charles Babbage, Alan Turing, Warren McCulloch and Walter Pitts, and Jerry Fodor —among many others— until becoming one of the central paradigms in cognitive science (Boden, 1988; Gigerenzer & Goldstein, 1996). The general thesis behind this influential view is that some set of formal/mathematical operations is what underlies human cognition. If we can model them *via* the right mathematical algorithms, then thinking can be formally explicated, and eventually replicated by some non-biological device.

However, the idea that reasoning can be formal in a logical sense is not exactly equivalent to the idea that cognition can be described by some mathematical formalism. The first thesis applies exclusively to rational (personal-level) inference, and it claims that classical logic can specify the mechanisms behind it. In contrast, the second applies to any cognitive process whatsoever, and it is not committed to one particular mathematical structure as a model.

These two notions of formality can be associated with the distinction made by Dutilh Novaes between *the formal as computable* and *the formal as de-semantification* (2012). The former notion emphasizes those features of formal systems allowing to specify a mechanism *via* some algorithmic procedure; while the latter refers to formal systems as structures devoid of any meaning or any semantic property whatsoever.[1]

Both versions of formality are clearly compatible and often go hand by hand. However, the emphasis on de-semantification is something that concerns especially to logical formality, since classical logic is built upon the assumption that

[1]Carnap was one of the main defenders of the formal as de-semantification. We can find a nice definition of this idea in the following passage:

> A theory [...] is to be called formal when no reference is made in it either to the meaning of the symbols (for example, the words) or to the sense of the expressions (e.g., the sentences), but simply and solely to the kinds and order of the symbols from which the expressions are constructed. (Carnap, 2000, p. 1).

a sharp line between syntax and semantics can be drawn. For this reason, I will focus exclusively on this last idea, trying to explain its role in cognitive science and cognitive psychology *via* the influence of what I call the "formality thesis."

Roughly, the formality thesis claims that personal-level inference is a syntax-driven, content-independent, and rule-based mechanism. Due to that, formalists see no room for semantic notions (like *meaning*) in psychological explanations of reasoning. When properly articulated, the formality thesis requires to make two crucial assumptions about the nature of mental representation: (i) that it has a language-like structure; and (ii) that its syntactic properties are entirely independent of its semantic properties.

Understanding the role of this idea in psychology is crucial for having a better insight on what researchers now call "the old paradigm" in the psychology of reasoning (Elqayam & Over, 2013): a family of theories that used classical logic as a model of deductive competence and/or performance (Chater & Oaksford, 1993; Overton, 1990). In what follows, I will explain the roots of the formality thesis in classical logic and its role in the development of the computational view of the mind. After that, I will discuss its influence on some important theories in the psychology of reasoning.

1.1.1 Logical formality and the hylomorphic tradition

The contemporary assumption that reasoning is formal is an heir of the idea that logic is the theory of correct inference *plus* the definition of deductive validity as a function of logical form. Understanding this requires digging a bit into the history of logic. In particular, into what John MacFarlane called the *hylomorphic* tradition (MacFarlane, 2000). According to MacFarlane, our contemporary view of logic is rooted in a *hylomorphic* conception that assumes the existence of a sharp distinction between *form* and *content* in reasoning and argumentation. In particular, this tradition assumes that those properties of reasoning that are logically interesting are formal properties, i.e., independent of the content or *topics* of arguments, and that are reflected in their grammatical or syntactic structure.

The *hylomorphic* view finds its main source in Aristotle's famous distinction between *form* and *matter* (Conway, 1995; MacFarlane, 2000). Aristotle brought the distinction from the *Physics* to the study of reasoning and argumentation by claiming that these two are also constituted by both *formal* and *material* properties. He then advanced the idea that *validity* was a formal property of arguments, as a result of his studies on everyday argumentation (Aristoteles & Ross, 1965, p. 29). Roughly, he saw structural similarities among intuitively valid arguments with varied contents, and he was able to pinpoint these similarities by using schematic letters as placeholders for lexical concepts (see, Corcoran, 2006, for a detailed explanation). The upshot was a classification of argument schemes representing different *forms* of valid arguments. Aristotle called the systematic study of these schemes *formal logic*, and wrote the *Prior Analytics* as a study of them.

However, Aristotle knew that a theory of reasoning and argumentation needed more than a theory of the formal conditions of deductive consequence. Thus, he proposed a complementary discipline, *material logic*, as the theory which studies those features of inference that rely on (*material*) knowledge. Aristotle devoted to that topic the *Posterior Analytics* (Aristoteles & Ross, 1965; Conway, 1995).

The hylomorphic view was strengthened during the *mathematization* of logic thanks to Boole's and Frege's works (see, Van Heijenoort, 1967), leading to the what Warren Goldfarb has called the *schematic conception of logic* (Goldfarb, 2001). In his words, according to the schematic conception:

> ...the subject matter of logic consists of logical properties of sentences and logical relations among sentences. Sentences have such properties and bear such relations to each other by dint of their having the logical forms they do. Hence, logical properties and relations are defined by way of the logical forms; logic deals with what is common to and can be abstracted from different sentences. (Goldfarb, 2001, p. 26)

The schematic conception specifies the old hylomorphic idea: sentences have

logical forms, and inferential moves among them are licensed by structural relations among these forms, regardless of their content. As it is evident, the problem lies in how the notion of *logical form* is defined. In the hylomorphic-schematic tradition, this is done by selecting a set of linguistic particles that are identified as invariant across subject matters on arguments. For instance, consider the following two arguments:

1. *Montevideo is a city. Cities are not countries. Thus, Montevideo is not a country*

2. *Fido is a dog. Dogs are not Birds. Thus, Fido is not a bird*

While (a) and (b) are about entirely different topics, they are structurally equivalent if we take "not," "and," and "thus" to be logical constants; "Fido" and "Montevideo" as names; and "City," "Country," "Dog" and "Bird" as predicates. The logical analysis of these arguments tells us that both (a) and (b) fit into the same logical scheme: $(Pa \wedge \forall x(Px \rightarrow \neg Qx)) \rightarrow \neg Qa$. From an inferential perspective, this means that deductive inferences *pivot* exclusively on logical constants, and that names and predicates are*inferentially inert*.

Topic-neutrality and truth-functionality

Specifying the exact set of logical constants is not an easy task (see, Bonnay, 2014; Gómez-Torrente, 2002). The traditional criterion for this says that logical constants have to be *topic-neutral*. Topic-neutrality is an expression coined by Gilbert Ryle (1954) for characterizing those linguistic items whose content emerges from the structural role they play in articulating and relating concepts and propositions. The peculiarity of logical constants —as concepts—, is that they are not representational, i.e., they do not stand for any object or class of objects in the world. As Ryle explains:

> Formal Logic, it might be said, maps the inference-powers of the topic-neutral expressions or logical constants on which our arguments pivot; philosophy has to do with the topical or subject-matter concepts which provide the fat and the lean, but not the joints or

> the tendons of discourse. The philosopher examines such notions as pleasure, colour, the future, and responsibility, while the Formal Logician examines such notions as all, some, not, if and or. (Ryle, 1954, p. 116)

Now, as Ryle himself recognizes, topic-neutrality is a slippery concept for doing this demarcation. The problem is that many different words that are often taken as predicates in classical logic exhibit some degree of topic-neutrality. For instance, relational concepts like *taller than* and *north of* refer to specific properties —they have a "topic"— but can be applied to several different domains —e.g., people, houses, trees, and so on. What is more, and as we will see later in this work, these expressions may allow for different (schematic) inferential patterns that are intuitively correct but formally invalid, like "$\forall x \forall y(Taller(x, y) \rightarrow Shorter(y, x))$."

One possible way of fixing this issue is to choose as logical constants only those terms which are *maximally general* (cf. MacFarlane, 2000, Chapter 3). In other words, lexical items that are not about anything in particular. Truth-functionality is the property that seems adequate to do this job. A term (*connective*) is truth-functional when the truth-value of any expression in which the term participates is a function of the truth-values of the compounds of that expression. Truth-functional connectives are topic-neutral; they cannot be *about* anything because they cannot have a truth-value by themselves. Instead they *articulate* the truth-values of others propositions.

This is the path that the schematic view has taken since it seems to be the only way a sharp distinction between *form* and *content* can be established within logical systems. In particular, they assume that only truth-functional lexical items can have inferential properties. Thus, they construe validity as formal —topic-neutral— because it only depends on the truth-functional structure of sentences (cf. S. Read, 1994).

Now, what is the contribution of *content* to the process of inference according to this approach? Truth-transmission is a bottom-up process that starts with atomic propositions being assigned a truth-value. The only semantic property

which matters here is that they are *truth-bearers*. All other semantic properties, e.g., those associated to the *topic* of the predicates of these propositions, are completely irrelevant. Since the truth-functional structure of arguments can be mirrored by syntactic features of language, and since this structure has nothing to do with *content* (predicate-meaning), the hylomorphic-schematic tradition explains deductive inference as the result of a sort of division of labor between the syntax and the semantics of language: we can make valid inferences about things without having to "look into" the content of extra-logical terms, because all what matters to validity is truth-transmission, and this is precisely mirrored in the syntax of sentences.

To sum up, the tradition in question characterizes deductive inference as a function of logical form, and logical form as completely topic-neutral. When logic is taken as a model of thinking, these ideas are given a psychological interpretation in which reasoning consists of decoding logical forms from natural language sentences, and then applying formal rules with truth-functional properties, but are content insensitive. In few words, inference is formal because the content of extra-logical terms does not play any role in deductive transitions.

In what follows, I will briefly explain how these ideas played a foundational role in the development of, arguably, the most influential philosophical framework in contemporary cognitive science, the Computational Theory of Mind (CTM), mainly due to the influence of Fodor's analysis of psychological explanation.

1.2 Formality in the foundations of cognitive science

1.2.1 Formality, causality, and rational thought

Broadly construed, *formality* played a central role in the foundations of cognitive science *via* the idea that all mental processes are computational in some sense (see, Piccinini & Scarantino, 2011). That idea is known as the *computer*

metaphor, and it is probably the most important analogy in the history of cognitive science (cf., Dennett, 1984; Searle, 1990). Its historical and philosophical minutiae are too many, and it is not among the aims of this thesis to cover them. Nevertheless, it is important to notice that logical formality is a particular case of computational formality within the computational paradigm in cognitive science. While the former exclusively concerns the computational structure of rational thought, the latter concerns any cognitive mechanism whatsoever — perception, learning, categorization, memory, etc. Furthermore, computational formality emphasizes that cognitive mechanisms can be specified by some algorithm, while logical formality plays an explanatory role in the relation between syntax and semantics in rational thinking. In what follows, we will focus on this last point.

As the story goes, one of the main challenges for psychology is to provide a scientific explanation of a phenomenon that is both intentional and physically grounded (Horst, 1999a). In particular, thoughts are intentional entities with semantic content —for instance, they are *about* something— articulated in non-arbitrary ways. And reasoning is a specific case of thought-transition that preserves (ideally) some semantic coherence. For those concerned by giving an empirically-grounded explanation of this process, the central question is then: how is rationality mechanically possible? (see Rescorla, 2012).

Jerry Fodor saw that one way of answering this methodological and foundational issue was through a psychological interpretation of logical formality (see Fodor, 1975; 1987; 2008; 2015). Roughly, he claimed that psychology must understand the mind as a syntax-driven machine that performs formal operations over language-like entities —thoughts— with both syntactic and semantic properties. Causal transitions between thoughts are possible thanks to their syntactic properties and since these properties *mimic* (Fodor, 1985, p. 93) the semantic content of thoughts, rational thought is also possible. As Fodor explains:

> ...you connect the causal properties of a symbol with its semantic

> properties via its syntax. The syntax of a symbol is one of its higher-order physical properties . To a metaphorical first approximation , we can think of the syntactic structure of a symbol as an abstract feature of its shape. Because, to all intents and purposes, syntax reduces to shape, and because the shape of a symbol is a potential determinant of its causal role, it is fairly easy to see how there could be environments in which the causal role of a symbol correlates with its syntax. It' s easy, that is to say, to imagine symbol tokens interacting causally in virtue of their syntactic structures . The syntax of a symbol might determine the causes and effects of its tokenings in much the way that the geometry of a key determines which locks it will open . (Fodor, 1987, pp. 18-19)

Fodor's view builds on logical formality because, as said before, this idea explains a sort of *division of labor* between syntactic and semantic properties of language-like structures. In particular, he takes inspiration from classical proof-theory, a formal system in which purely syntactic rules mirror truth-preserving transitions between propositions (see Fodor,1987, p. 19; Fodor, 1985, p. 93). [2]. In Barwise words:

> What has captured Fodor's imagination is that we logicians have developed formal proof procedures for certain formal languages, procedures that can be used to build inference engines, machines that can carry our formal proofs, even if not very well. Here, Fodor thinks, is hope for a mechanism underlying thinking. (Barwise, 1986, p. 331)

Thus, Fodor argues that the only way a *naturalistic* psychological theory can explain non-arbitrary causal transitions between thoughts, while preserving intentional notions like *meaning* and *truth*, is by specifying the computational

[2]For instance, Fodor and Pylyshyn claimed that classical cognitive science is "*an extended attempt to apply the methods of proof theory to the modeling of thought (and similarly, of whatever other mental processes are plausibly viewed as involving inferences; preeminently learning and perception.) Classical theory construction rests on the hope that syntactic analogues can be constructed for nondemonstrative inferences (or informal, common-sense reasoning) in something like the way that proof theory has provided syntactic analogues for validity.*" (Fodor & Pylyshyn, 1988, p. 30)

mechanisms that are exclusively syntactically-driven but that can mirror semantic and normative properties:

> Thinking can be rational because syntactically specified operations can be truth preserving insofar as they reconstruct relations of logical form; thinking can be mechanical because Turing machines are machines. ... [T]his really is a lovely idea and we should pause for a moment to admire it. Rationality is a normative property; that is, it's one that a mental process ought to have. This is the first time that there has ever been a remotely plausible mechanical theory of the causal powers of a normative property. The first time ever. (Fodor, 2001, p. 19)

The formalist approach, pioneered by Fodor, has had a tremendous influence on the philosophy of mind and cognitive science ever since. Many researchers saw it as establishing the foundations for a naturalistic theory of high-level cognition, which can co-exists with an intentional psychology. However, this view comes at a price. First, this approach assumes with little evidence that a *semantic engine* can supervene on a *syntactic engine*. In Haugeland's words (1989, p. 106), they believe that "*if you take care of the syntax of a representational system, its semantics will take care of itself.*" That is, at least, polemic. Formalists do not propose an explanation of how this semantic mirroring works in a psychological context. In general, they do not offer any criterion for demarcating syntactic properties from semantic ones, something that is problematic since their main claim is that computation has to be purely syntactic. (cf. Aydede, 2005; Peacocke, 1999; Rescorla, 2012).

Fodor's own criterion is far from systematic. Even if he claims that the *formality condition* is the main requirement for an empirical/intentional psychology, the only definition he gives of "formal" is "non-semantic" (see Fodor, 1980, p. 102). He then claims that formality depends only on the *shape* of mental symbols, implying that these shapes cannot include any kind of semantic information in them, even if they mirror semantic content.

Now, everything seems to depend on what "semantics" turn out to be. In this regard, Fodor follows -again- the logical tradition. The only semantic properties he conceives are denotational relations between extra-logical concepts and external things, and the truth conditions of belief states. As he states in his last book: "*reference is the only semantic property of mental or linguistic representations*" (Fodor & Pylyshyn, 2015, p. 10). This last claim is far from obvious. Meaning can be thought as having properties that go beyond reference. For instance, the sentence "the cat is on the mat" refers to a specific situation in the world, but also induces a representational state —in those who understand it— with rich conceptual information that can trigger inferences or other forms of thought transitions. I do not see how this information can be seen as non-semantic, and not having a causal role in rational thought-transitions. I will discuss the sources of this view in the following chapter, and some of its implications in Chapter 3.

To sum up, CTM is fully committed to the formality thesis. Reasoning is formal because its causal structure is fully determined by the syntactic properties of belief-like mental representations, not because it can be specified by some mathematical structure. Then, what matters here is not "general" computational formality, but logical formality:

> To say that an operation is formal isn't the same as saying that it is syntactic since we could have formal processes defined over representations which don't, in any obvious sense, have a syntax. Rotating an image would be a timely example. What makes syntactic operations a species of formal operations is that being syntactic is a way of not being semantic. Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference and meaning. (Fodor, 1980, p. 64)

I will go back to the CTM later, while discussing the notion of representation. In what follows, I explain the role of the formality thesis in some of the most influential theories of reasoning in cognitive psychology. It is impossible

to do justice to such a complex topic in the few following pages. It is not among my aims to offer a systematic historical analysis of this. Instead, I intend to illustrate how this thesis shaped the views that psychologists had about reasoning. For doing this, I will focus on Piaget's developmental view of reasoning skills, Mental Logic theory, and Mental Models theory.

1.3 Formality in the psychology of reasoning

1.3.1 Piaget's logicism

As said before, logic played a central role in the development of cognitive psychology, especially in early explanations of reasoning (Henle, 1962; Overton, 1990). One of the most influential figures in this process, the Swiss psychologist Jean Piaget, put together a theory of high-level cognition that understood logical reasoning as the zenith of human cognitive development (Piaget, 1956). Together with Inhelder, he famously claimed that "*reasoning is nothing more than the propositional calculus itself.*" (Inhelder & Piaget, 1958, p. 305). For various reasons, his theories have been mostly abandoned (see Braine, 1962), but no doubt they set the conditions for the future discussion in psychology about the relationship between logic and reasoning. In what follows, I will briefly discuss Paiget's ideas in this regard.

Logic was a central piece in Piaget's theoretical apparatus. The Swiss was deeply familiarized with the philosophical discussions about the relation between logic and psychology, rejecting any naive form of psychologism —i.e. the idea that logic must be founded in psychological facts. In his "Logic and Psychology" (Piaget, 1957), he discusses different forms in which logic has been related to psychology in philosophy. The central issue here is whether logical laws are related to the cognitive structures that organize experience —in the Kantian sense (see Wartofsky, 1983), whether they are mere empirical contingencies, or whether they are formal relations between the sentences of natural language (Black & Overton, 1990).

Among the various alternatives that he considers, he leans to *operationalism*, a popular methodological view at that time (Feest, 2005). Piaget's argument for this is straightforward. He seems to believe that since operations are what characterize cognition, an abstract theory of them must be central for the explanation of cognitive mechanisms. He writes:

> Operations (in spite of Couturat!) play an indispensable role in logic, since logic is based on an abstract algebra and made up of symbolic manipulations. On the other hand, operations are actual psychological activities, and all effective knowledge is based on such a system of operations. (Piaget, 1957, p. 7)

In this sense, Piaget believes that the operations studied by logic can shed light on the structural features of cognitive procedures:

> The algebra of logic can therefore help the psychologist, by giving him a precise method of specifying the structures which emerge in the analysis of the operational mechanisms of thought. (Piaget, 1957, p. xviii)

Just like in contemporary philosophy of logic, Piaget's notion of formality is based on the principle of invariance under substitutions:

> The content of an operational relation is constituted by the data, or by the terms that can be substituted within it, while the "form" is what stays unchanged during these substitutions. (Piaget, 1949, p. 38; *my translation*)

However, which specific elements assume the role of "form" and "content" depends on the kind of operation we analyze. Piaget distinguishes between operations working at the inter-propositional and at the intra-propositional levels. The former correspond to the structures studied in propositional logic. In that case, the *forms* are relations among propositions determined by logical constants: The formula $((p \to q) \wedge (q \to r)) \to (p \to r)$ has "$p \to q$", "$p \to r$", and "$p \to r$" as *content* and $(\phi \wedge \psi) \to \psi$ as *form*. The latter correspond to those

operations characterizing the internal structure of propositions or *thoughts*. In these cases, the *form* is specified by set-theoretical relations like membership and inclusion, while the *content* the very *objects* of these operations. According to Piaget, intra-propositional operations are behind basic cognitive mechanisms like categorization.

Piaget's developmental view of cognition led him to a *gradualist* view of formal operations. In early stages of development, children reason with domain-specific cognitive schemes. For instance, during the "concrete operational state" —ages 7 to 11—-, they are able to make transitive inferences regarding physical properties, like "An apple is bigger than a grape; and a watermelon is bigger than an apple. Thus, a watermelon is bigger than a grape". But they are not able to understand *transitivity* as an abstract relation, i.e., they do not grasp it as a domain-general principle. This type of understanding takes place when children see that transitive inferences are *reversible*, i.e., that they have "symmetric" operations that can reverse the original one. For instance, understanding that A is bigger than B implies that B is smaller than A allows them to grasp "smaller than" as the reverse of "bigger than." Children master this when they can make inferences like "A is bigger than B, and C is smaller than B; thus, A is bigger than C." This requires an abstract understanding of transitivity (Piaget, 1947). Thus, the passage from the concrete operational state to the formal operational state happens when children grasp the domain-general principles behind the concrete operations they used to perform.

In general, Piaget thinks that this transition depends on children's ability to come up with hypotheses about the logical relations underlying the concrete operations they perform (see Chapman, 1979, for a detailed explanation). The formal operational state is a high-order state because it includes the formal schemes of these logical relations, later used in hypothetico-deductive reasoning, which is, for Piaget, a form of thinking detached from immediate perception and based on conceiving possibilities (Inhelder & Piaget, 1958, p. 254).

In conclusion, Piaget was a logicist, since he believed that fully developed reasoning was completely topic-neutral and domain-general. However, his commitment to the formality thesis is relative. He avoids the sharp distinction

between syntax and semantics of the hylomorphic-schematic tradition; and the notion of logical validity seems to play no role in his view. Instead, Piaget seems to think about logic in "Boolean" terms, that is, prioritizing the study of the algebraic nature of logical operations over the notion of deductive validity. I believe this shows that he viewed logic more as a competence model of reasoning than as a model of performance.

As a final comment, there is an interesting plot-twist in Piaget's logicist view of reasoning, which is mostly overlooked in the literature. By the end of his career, Piaget abandoned logicism for a content-based view of reasoning (see Byrnes, 1992, for a detailed explanation). Influenced by Anderson's and Belnap's relevant logic (A. Anderson, Belnap Jr, & Dunn, 2017), Piaget convinced himself that truth-functional logic, and specially purely formal operations, were not adequate for describing reasoning. Consequently, he tried to develop a concept-based view of inference that he called "logic of meanings" (Piaget & Garcia, 1990). In a way, this thesis in an attempt to advance some basic ideas for developing a *logic of meanings*. That is, a theory of inferences that depend on the representational structures behind lexical-concepts, and not in logical form.

1.3.2 Mental Logic Theory

The theory that is most faithful to logic and the formalist thesis is Mental Logic theory (MLT). It directly builds on the idea that reasoning consists of operations over language-like mental representations, guided by logical rules which are "activated" by the syntactic properties of these "mental sentences" (Braine, 1990; Braine, Reiser, & Rumain, 1984; Rips, 1994).

The influence of classical logic in this theory is straightforward since its basic principles are a psychological interpretation of natural deduction. MLT finds its foundations in Fodor's version of the CTM, that —as we saw— conceives reasoning as operating over a language of thought that clearly separates syntactic and semantic properties. Based on this, MLT claims that we are

equipped with a set of abstract derivation rules that apply to information that is propositionally represented in the mind/brain.[3]

For instance, MLT assumes that we have an abstract *modus ponens* scheme like the following one:

> If A, then B
>
> A
>
> Therefore, B.

Whenever we are before propositional information that match the form of the rule, we use it to infer. For instance, if we hear the sentences "*if the car is not at home, then the house is empty*, and *the car is not at home*", the rule allows us to see the entailment and draw the conclusion "*the house is empty.*"

One prominent theory within this tradition is Rips' psychology of proof *1994). Like the other models (Braine, 1978; Osherson, 1975a), Rips' central claim is that human reasoning is essentially a logical proof-system, he calls this the "Deduction-System Hypothesis." The role of psychologists is to specify the set of psychologically plausible rules of inferences and the cognitive mechanisms that apply them in everyday reasoning. He writes:

> I assume that when people confront a problem that calls for deduction they attempt to solve it by generating in working memory a set of sentences linking the premises or givens of the problem to the conclusion or solution. Each link in this network embodies an inference rule. . . , which the individual recognizes as intuitively sound. (Rips, 1994, p. 104)

Rips' model (called "PSYCOP") is a psychological interpretation of standard natural deduction.[4] Roughly, it is constituted by a set of inference rules for logical constants —including quantifiers—, plus a device that extracts the syntactic form of premises as input for applying the rules (see Figure 1.1). Examples of rules are the introduction of the conjunction (*A, B. Thus, A & B.* and

[3] Another important theoretical framework for MLT is Macnamara (1986).

[4] PSYCOP was implemented in PROLOG and tested experimentally in various occasions (see Rips, 1994, Chapter 4 and 6).

the *modus ponens* (*If A, then B. A, Thus B.* The systems also deals with suppositions. Just like in natural deduction, suppositions used in the demonstration process have to be "discharged" —by using an introduction of the conditional or reaching a contradiction— in order to reach the conclusion.

Let's see a brief example of this. Consider the premises (a) "If Maria did not go to the library, Juan did not meet her", and (b) "Juan met Maria". The proof-system will first retrieve from (a) and (b) the logical forms (a') $\neg A \rightarrow \neg B$ and (b') B. Then, $\neg A$ is introduced as a supposition, and is used with (a') in a *modus ponens* to infer $\neg B$. $\neg B$ is in contradiction with premise (b'), and this allows for using the rule of *reductio ad adsurdum* to finally infer A ("Maria went to the library"), discharging also the supposition.

One of the main issues for MLT is the problem of the computational viability of rule application. Logical systems do not offer any constraints regarding the derivations one may draw from a given set of premises. For instance, given the premises p and q, one can derive $p \rightarrow q$, $q \rightarrow p$, $p \wedge q$,$p \wedge (q \wedge p)$, $p \wedge (q \wedge (q \wedge p))$, etc. As it is evident, arbitrary rule application would be considered completely irrational, since informational and contextual factors constrain human inference.

Rips' solution to this issue is to constrain the application of introduction rules exclusively for backward chains ——-from conclusions to premises— (cf. Braine, 1978). He designed PSYCOP for working with three types of rule-application: rules allowed for backward inferring, rules allowed for forward inferring, and rules that work both ways (see Figure 1.1).

The set of psychologically plausible rules only includes rules that are somehow intuitive for agents. However, Rips' experiments show that some of the rules above do not meet this criterion —the introduction of "or", for instance. Rips also assumes variability across individuals in the set of rules. People may learn new rules or use non-standard rules in reasoning (Rips, 1994, p. 104), violating rules of classical logic but also accounting for individual variability in reasoning.

Another important challenge for this theory is to deal with the potentially intractable computational complexity of proofs (Rips, 1994, pp. 63-68). Rips proposed a deterministic search protocol that implies to check all applicable

FIGURE 1.1: Forward, backward, and bidirectional rules in Rips' PSYCOP. From Rips (1994).

forward rules for finding a suitable conclusion, and if this process fails, then the system works backwards from the conclusion till it finds the inferential steps required to arrive to the premises. If the system fails to find a proof after applying this protocol, then the subject concludes that there is no valid conclusion to draw from the available set of premises.

1.3.3 The Wason selection task: troubles for logicism

Beyond its technical issues, this syntactic theory of reasoning had found some empirical validation in various studies (see, Braine et al., 1984; Ford, 1995; Galotti, Baron, & Sabini, 1986). However, experimental data were not always benevolent with logicist theories of reasoning. The Wason selection task (Wason, 1968), the most influential experiment in the psychology of reasoning, provided

robust evidence for the idea that our logical competence is rather bad, directly undermining any logicist model of reasoning.

Roughly, this experiment starts by showing participants four cards with two letters and two numbers inscribed on them. They are then told the following rule: *If a card has a vowel on one side, it has an even number on the other side.* The task consists of asking the subjects which of the cards are worth turning over to test the rule in question (see Figure 2.2)

Which cards do I need to turn over in order to see if the rule is true?

FIGURE 1.2: Classical setting of the Wason selection task.

This abstract form of the Wason selection task yielded the following distribution of answers (Wason & Shapiro, 1971): 45% of the participants pick the A card and the 4 card; 35% pick the A card alone; 7% pick the A card, and the 7 card; 4% pick the A card and the 7 card; 9% pick other combinations of cards.

Notice that the logical structure of the problem is quite simple. Participants have to test a conditional rule of the form $P \to Q$. The easiest way of doing this is by considering the equivalence $\neg Q \to \neg P$, and applying a *modus tollens*: $P \to Q$, $\neg Q$, thus $\neg P$. Thus, the correct answer is to turn over the A and 7 cards. Strikingly enough, this is the less chosen option. The quick conclusion is that people are rather bad at applying logical rules. This is serious trouble for formalist theories; why are we bad at reasoning with logical rules if they are supposed to be built-in into our cognitive structure? In this sense, Wason —criticizing Piaget— wrote:

> The results are . . . disquieting. If Piaget is right then the subjects in the present investigation should have reached the stage of formal operations. A person who is thinking in these terms will take account of the possible and the hypothetical by formulating propositions about them. He will be able to isolate variables . . . and subject them to combinatorial analysis. But this is exactly what

> subjects in the present experiment singularly fail to do. . . . Could it be that the stage of formal operations is not completely achieved at adolescence, even among intelligent individuals? (Wason, 1968, p.281)

Content and Context effects

In Wason and Shapiro (1971), it was shown that when the same task was done with familiar content the results changed dramatically. In the *thematic* version of the task, subjects evaluated the rule "*Every time I go to Manchester I travel by car*", using cards with cities on one side and modes of transport on the other. The experiment showed that 62% of the participants gave the correct answer, against a scarce 12%when the task used more abstract content in the same study. According to Wason and Shapiro, the improved performance on the thematic version was evidence that reasoning was mostly driven by content-related mechanisms, and not by syntactic ones. After this, psychologists started talking about "content-effects" on reasoning, a robust empirical phenomenon of changes at the level of the solution frequency of reasoning tasks of the same logical structure, according to the kind of *contents* used in them (see, Dominowski, 1995; Pollard & Evans, 1987).

Empirical studies on content-effects consistently showed that when subjects reason with rules which are similar to the ones used in everyday life, they have a good logical performance (Golding, 1981; Griggs & Cox, 1982; Manktelow & Evans, 1979). In particular, prior knowledge and past experience seem to make a big different in deductive reasoning (see Evans & Feeney, 2004, for a review).This led researchers to claim that agents do not typically use domain-general rules of inference, but they look for information and counterexamples in domain-specific memories in problem-solving contexts.

Furthermore, context also has a big influence in reasoning in general. Belief biases are very common in syllogistic reasoning (see, e.g. Klauer, Musch, & Naumer, 2000; Markovits & Nantel, 1989). For instance, take the following argument:

Mammals have fur.

Dogs have fur.

∴ Dogs are mammals.

This argument is invalid, but most people take it as valid since the conclusion coheres with their prior knowledge. Evans et. al. (1983) claim that there is a complex interaction between *validity* and *believability* in argument evaluation. There seems to be a strong tendency to accept or reject an argument based on how believable or unbelievable is the conclusion, disregarding logical structure. Also, Thompson (1996) found that premise-believability has an impact on the perception of argument strength.

Content and context effects had a strong impact on the psychology of reasoning. Most researchers, since the 1980s, started assuming that reasoning was content-dependent: "*dependent, that is, on content which evokes relevant knowledge from the memory*" (Manktelow & Over, 1990, p. 111). The explanatory limitations of the formalist approaches for accounting for these facts made them lose ground within the psychology of reasoning (Evans, Newstead, & Byrne, 1993; Johnson-Laird, 2010a; Johnson-Laird & Byrne, 1991).

However, this was not the end of the logicist view. There have been different attempts from the logicist side to accommodate these effects. A prominent proposal consists of incorporating content-specific rules to the theory or enriching the formal language with modal operators of different sorts. [5]. However, incorporating semantic content to the explanatory structure of the theory would contradict the basic tenet about the purely syntactic character of our inferential mechanisms. Also, as some psychologists have noticed, this would weaken the parsimony and testability of the theory (Manktelow & Over, 1991). Furthermore, in case this is done, the theory would have to explain how semantic content is articulated in reasoning, something that they do not do.

Other attempts two stays close to logicism consisted of weakening the formality assumption by claiming the inferential rules behind reasoning, are not

[5]This is the psychological equivalent of the meaning-postulates approach in semantics, which we will discuss in the following chapter.

topic-neutral and domain-general, but domain-specific. Two competing theories emerged in this line: One was the *pragmatic reasoning schemas* approach, developed by Cheng and Holyoak (1985; 1986); and the other was the social contract theory, developed mostly by Cosmides and Tooby (1989; 2010). Both theories attempt to explain content-effects on inference by analyzing deontological reasoning: problematic environments in which agents have to reason about permissions and obligations. For reasons of space, I will not analyze these theories here. However, it is worth saying that even if they gave content a constitutive role in inference, they focus on fairly narrow reasoning domains. In other words, they do not provide a general theory of content-based reasoning.

Content and context insensitivity is often considered a defeater of formalists approaches like MLT. This makes sense, since as we saw initially, the theory is explicitly founded on the formalist thesis, and there is no room for *content* within this explanatory framework. However, mental logicians still try to argue that we must distinguish between two different processes that constitute problem-solving (Bonatti, 1994b). The first one is *comprehension*, which is content-sensitive; and the second one is reasoning proper, which is purely syntactic. As Bonnatti explains:

> After a first processing roughly delivering a syntactic analysis of a linguistic signal, the identification of its logical form and a first semantic analysis retrieving literal meaning, pragmatics and general knowledge aid to select a particular logical form for the input signal. Afterwards, representations possibly sharply different from the first semantic analysis are passed onto a processor blind to content and pragmatics...So a theory of mental logic cannot, and does not intend to, explain the role of content in reasoning, though it may help to locate how and when content and pragmatics interact with reasoning proper. (Bonatti, 1994b, p. 20)

The distinction between *comprehension* and *reasoning proper* seems somewhat arbitrary, unless we take for granted the formalist thesis. Reasoning,

broadly construed, is just drawing conclusions from sets of premises or data. Everyday inductive generalizations are deeply related to memory (Feeney, Hayes, & Heit, 2015; B. Hayes, Fritz, & Heit, 2013) and do not follow deductive rules. Likewise, abductive and inductive inferences cannot be reduced to logical rules, and they rely heavily on background knowledge. From Bonnatti's perspective, these are not mechanisms within "reasoning proper."

Furthermore, there is robust evidence showing that we have trouble reasoning with abstract materials, that is, with premises that are very close to the "logical forms" that are supposed to be used by our built-in deductive rules. If prior to "reasoning proper," there are semantic-based mechanisms that "clarify" the logical forms of the input, then it should be expected that the processing times with abstract materials are lower than with contentful ones. But the situation is the opposite. Similarly, children learn to reason with generic statements like "Birds fly" or "Tigers are stripped" before they have a proper understanding of quantifiers (S. Gelman, Leslie, Was, & Koch, 2015; Leslie, 2008). If reasoning proper is purely syntactic, it is hard to explain how they recover logical form while reasoning with generics and cannot fully grasp quantified statements.

1.3.4 Mental Models Theory and the semantic turn

A fierce critic of MLT was Philip Johnson-Laird. The British psychologist found rule-based approaches — especially fully syntactic ones— ill-equipped for explaining human reasoning. In particular, he thought that their inability to explain how people retrieve logical form from natural language, which is inherently ambiguous and context-dependent, plus their failure to account for thematic-content effects in reasoning, made the proof-theoretic views seriously wrong (Johnson-Laird, 1983, 2010a; Johnson-Laird & Byrne, 1991).

The alternative that Johnson-Laird advanced implied changing syntax for semantics as the foundational notion for understanding human inference. His "Mental Models theory" (MMT) builds on the idea that the representational format underlying reasoning is not propositional but *iconic*. According to MMT, reasoning consists of representing propositional information in "meaningful"

mental models that the agent sequentially construct and analyze in order to draw conclusions.

For MMT, reasoning unfolds in three stages: *comprehension*, *description*, and *validation* (Johnson-Laird & Byrne, 1991). In the first stage, agents construct an initial mental model somehow analogous to the state of affairs —or information— described in the premises. For this, they use "*knowledge of the language and their general knowledge to understand the premises*" (Ibid., p. 35). This stage is constrained by a "principle of truth" that minimizes the load on working memory by avoiding the representation of mental models of "false" situations (Johnson-Laird, 2010b, p.14).[6] In the second stage, the mental model is *inspected*, and a putative conclusion that is compatible with the model is drawn. Finally, in the validation stage, the agents analyze other possible models —implicit in the first representation— compatible with the premises looking for counterexamples. If they find some, then the conclusion is falsified. If that is not the case, the conclusion is definitely drawn.

To see an example, consider that a reasoner is given the following rule: *if the car is near the house, then John is in the house*. The agent's first move is to represent an initial model with a situation in which both the antecedent and the consequent are true. If the agent is then told "the car is near the house", she will directly infer "John is in the house," because with this new information no other compatible models are conceivable. This correspond to the application of the *modus ponens*. Now, in a situation in which the new information is not "the car is near the house," but "John is not in the house," the agent will have to *unfold* all possible models —second column in Figure 1.3— and she will find only one compatible model with the information in the premises —(c). She will then conclude "the car is not near the house". This correspond to the application of the *modus tollens*. [7]

As illustrated above, reasoners draw a conclusion that holds in the models, but that is not present explicitly in the premises. A basic notion of *inferential*

[6]These ideas are similar to Barwise's situation semantics, in particular, to his notion of "partial situation"(Barwise, 1989).

[7]Note that the models in the second column of Figure 1.3 correspond to the truth-table of the conditional in the premise minus the case in which the conditional is false.

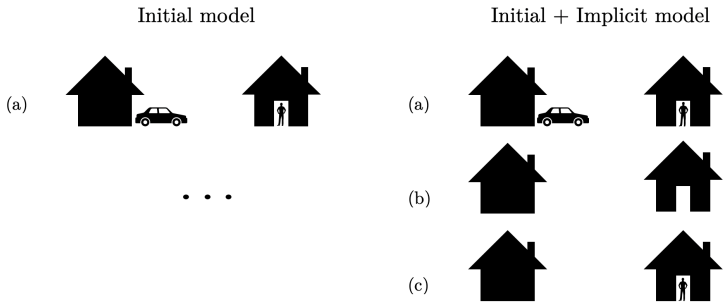


FIGURE 1.3: Mental model of the premise *if the car is in front of the house, then John is in the house.*

strength can be defined with this: a *necessary* conclusion is one that holds in all models compatible with the premises. A conclusion that holds in most compatible models is probable, but not necessary. And a conclusion that holds in at least one compatible model is possible according to the premises (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999).

MMT makes several predictions that have been empirically tested. One that is straightforward is that the number of situations represented in the implicit model is a positive function of the complexity of the inferential procedure, since the working memory load would be bigger. Consequently, the more situations an agent has to revise during reasoning, the longer the time the process will take. One way of testing this is by comparing agents' performance while reasoning with inclusive and exclusive disjunctions. MMT predicts that reasoning with statements of the form *A or B, but not both* should be easier than with statements of the form *A or B, or both*. Due to the principle of truth, the implicit model of the former includes only two sub-models ($\neg A \& B, or \neg B \& A$), while the implicit model of the latter includes three sub-models ($\neg A \& B, \neg B \& A, or A \& B$) (see Johnson-Laird & Byrne, 1991; Johnson-Laird, Byrne, & Schaeken, 1992).[8] Besides this apparent empirical success of the theory, several authors have noticed that it is not clear if the MMT does it better than the MLT accounting for the empirical literature in propositional and syllogistic reasoning (Evans, Over, & Handley, 2005; Oberauer, 2006)

[8]MLT fails to account for this. Within Rips' model, the exclusive disjunction has to be rephrased in terms of the conjunction and the inclusive or ($(A \vee B) \wedge \neg(A \wedge B)$). Thus, the number of rules used for proving statements with exclusive or is always longer than with similar statements with inclusive or.

Beyond its success, MMT has received many criticisms. The first problem that the theory has concerns the vagueness of its central notion: *model* (Braine, 1978; Evans, 1993a).According to Johnson-Laird, models are based on "*descriptions, on perception, and on knowledge*" (Johnson-Laird, 2010b, p.14). Taking Craik (Johnson-Laird et al., 1992) and Peirce (Johnson-Laird, 2002) as forerunners of the idea, Johnson-Laird has claimed that mental models are iconic, not sentential (Johnson-Laird, 2012, pp.135–136). That feature would make them a "semantically rich" format of representing information. However, most applications of the theory do not make use of this "rich" semantic content, but only represent logical information (propositional) in the same way as MLT does.

In other occasions, Johnson-Laird claimed that mental models are inspired in model-theoretic —Tarskian— semantics (Johnson-Laird, 1981). As some have argued (Hodges, 1993; Lowe, 1993), it is far from clear how a Tarskian model could be psychologically represented. These entities are set-theoretical structures (not image-like) that are used to provide extensional interpretations of propositions in a formal language. For any proposition, there are infinite models that make it true. It is hard to see how to conceive an economic cognitive procedure able to select from infinite possibilities relevant models for interpreting sentences.

Another problem of the MMT is that there is no explanation of how people construct these models in their minds, and how information is structured within them (O'Brien, Braine, & Yang, 1994). In other words, Johnson-Laird seems to allow mental models to represent many different kinds of information —conceptual relations, iconic and spatial information, and propositional information—, but there is no straightforward procedure of how semantic information (in the general sense) is structured within the models (For a critique in this line, see Evans, 1993b).

How *semantic* was the semantic turn?

As said earlier, MLT was outspokenly syntacticist: It assumed the Fodorian thesis that inference is formal manipulation of sentences in a language of thought.

In that way, it endorsed the hylomorphic tradition that draws a sharp line between form and content and assumes that all that matters to deduction concerns the former and not the latter. The robust evidence of content-effects in reasoning seriously troubled the idea that inference could be exclusively syntactic. Consequently, two theories (more or less compatible with the MLT) emerged as ad-hoc solutions to this issue: the pragmatic-arguments scheme theory and the domain-specific approach. However, as we explained before, these attempts lack systematicity: they do not offer any general explanation about how content participates in reasoning, and they focus exclusively on very narrow reasoning domains.

Before the limitations of the rule-based approaches, the MMT emerged as the alternative promising to include the right "*machinery to deal with meaning*" (Byrne, 1991), and there is no doubt that Johnson-Laird and his collaborators made a big contribution in this sense. The theory has a foundational dimension that provided an explanation of how semantic content is constitutive of inference, at the same time that it was able to explain empirical data and make new predictions. However, as I will next argue, the notion of semantic content that the MMT uses is still logicist, and even if it can explain some content effects in syllogistic reasoning, this is not enough to explain content-based inferences.

MMT builds on a procedural semantics that decodes logical form from ambiguous linguistic or perceptual inputs. [9] The semantic procedure in question translates a propositional representation into a mental representation preserving essentially an non-ambiguous logical form. For instance, two conditional rules with the same logical form but different content are going to generate mental models —initials and explicit— that are structurally equivalent.

Even if according to the MMT, the "units" of reasoning are iconic models in the mind of the agents instead of sentences in a language of thought, the structure of the sets of models in some instance of reasoning responds to logical formality, and the normative structure of the model to classical validity (for a similar critique, see Mercier & Sperber, 2017). But, instead on focusing on

[9] There is no clear explanation of how mental models are constructed from perception. The vast majority of the examples used on the literature are from linguistic inputs.

syntax, it focus on the truth-functional structure of the propositions. Johnson-Laird's theory is *semantic* in the same sense as truth-tables are semantic (Andrews, 1993): They represent possible distributions of truth values for a given proposition according to the distribution of logical constants on it. In principle, there is no room in this view for the content of extra-logical terms in the core process of reasoning: just like with MLT, the inferential power lies exclusively on logical constants and not on lexical concepts (For similar observations, see Bonatti, 1994a; Lowe, 1993; Rips, 1994).

This *logicist* conception of reasoning that characterizes MMT can be seen in the treatment that the theory gives to material implication. In its first version, the theory had no constraints for reasoning with conditional statements with completely unrelated antecedents and consequences. In particular, the theory accepted the classical conditional, with its respective paradoxes of the material implication (Johnson-Laird et al., 1992), which are widely considered as counter-intuitive inferences that theories of reasoning should reject (Bonatti, 1994a; R. C. Stalnaker, 1968). In this sense, Evans and Over wrote:

> For us, the original sin of JLB [Mental Model] theory is their endorsement of the logical validity of the paradoxes of interpreting a natural language conditional as truth functional. We hold that the paradoxes are logically invalid for natural language conditionals, including "basic" conditionals. (Evans & Over, 2004, p. 153)

Furthermore, as Bonnatti showed (1994a), another relevant dimension of semantic content that the MMT is not able to account for is context-sensitivity. The procedural framework used to recover the meaning from the propositional inputs in MMT is only capable of grasping literal meaning, neglecting in this way, the pragmatic contexts of sentences. Also, as Van der Henst points out (2000), the pragmatic dimension also constrains the kind of inferences that someone can draw given a set of premises. For instance, if we have the following propositions:

- Pedro is taller than Juan.
- Juan is taller than Paul.

You can infer "Pedro is taller than Paul" or "Paul is the shortest of the group" or "Pedro is taller than Juan and Paul" , etc. In general, as the relevance theory shows (Sperber & Wilson, 1986; D. Wilson & Sperber, 2012), the context of processing will have an important role in determining which inference is the most relevant to draw. There is no explanation in MMT of how inference could be context-sensitive in this sense.

In later work, Johnson-Laird and Byrne (2002) tried to fix these issues by distinguishing two kinds of "meanings" in model-construction: the "core meaning", which responds to the truth-functional structure of the premises, and the "modulated meaning" which is the result of a mechanism that constraints the models of the core meaning according to different content-based relations between the antecedent and the consequent. To see an example, consider the conditional statement (a) *If Pat is in Italy then she is not in Rome* (with logical form $A \rightarrow \neg B$). The mental models corresponding to the core meaning of (a) are:

- Pat is in Italy and is not in Rome. $< A, \neg B >$
- Pat is not in Italy and is not in Rome.$< \neg A, \neg B >$
- Pat is not in Italy and is in Rome.$< \neg A, B >$

However, the represented models are only the first two, since the third one is blocked by semantic modulation since that situation contradicts common knowledge —*if x is in Rome then x is in Italy.*

There is some empirical evidence about this sort of semantic modulation (see Quelhas, Johnson-Laird, & Juhos, 2010). Essentially, this update to the MMT takes care of some semantic relations between the clauses of conditionals in a better way than the original version. It goes beyond the truth-functional analysis of the *if...then...* relation in natural language, and recognizes a contribution of extra-logical content in reasoning (Byrne & Johnson-Laird, 2009).

I consider this an excellent improvement regarding Johnson-Laird original intentions of building a semantic-based theory of reasoning. The changes made to the original theory go in the direction of the ideas defended in this thesis:

a semantic-based theory of inference must build on account of conceptual relations, and not only on truth-functional considerations. However, Johnson-Laird and Byrne do not develop any systematic account of conceptual content for understanding this new kind of *modulation*. The semantic constraints on model construction seem to be based in our (or the experimenter's) intuitions about conceptual relations, neither on a theory of conceptual representation nor on a formal model of semantic relations.

This last point is the central concern of this thesis. As said in the introduction, I attempt to provide the basis for a model that can explicate the role that lexical concepts have in rational inference. This model will explain immediate inferences with different word classes. And in this sense, as it will be explained in Chapter 5, it could provide a formal framework for modeling the second type of semantic modulation that concerns MMT.

1.3.5 Bayesian models against formality

Logicism holds a highly idealized view of reasoning and rationality. As Cherniak has shown, this view makes unrealistic assumptions about our computational capacities, the structure of working memory, and our sensitivity towards the global coherence of our belief system (Cherniak, 1986). Maybe the biggest of these idealizations is the one at its core: we reason by considering truth-functional relations between logically-structured beliefs. Truth-functional relations in logic are *informationally conservative*, i.e., the information in the conclusion is implicit in the premises. In this sense, a deductive inference is a safe inference because it is absolutely certain. However, everyday reasoning cannot afford this. Navigating our environment requires us to cope with constant uncertainty and partial information (Oaksford & Chater, 1989). Truth-preservation is not the central concern here. The type of mechanisms needed are those allowing us to make risky predictions, and draw nonmonotonic inferences about new stimuli exploiting background knowledge.

The emergence of probabilistic approaches to reasoning and rationality —— nowadays mainstream— was, to a certain extent, a reaction to the limitation

above. (Oaksford & Chater, 2007, Chapters 1-3). In particular, Bayesian approaches claim that the normative framework provided by logicism is wrong and, consequently, much of the experimental data that we thought indicated that we were bad reasoners, when reinterpreted from a probabilistic perspective shows the opposite. One of the core assumptions of Bayesianism is that we do not reason from true premises to true conclusions. Instead, we have *degrees of beliefs* about sentences, and we reason from more or less plausible beliefs to more or less plausible conclusions. This basic assumption requires replacing the main computational story of logicism with one based on probabilistic assumptions.

While logicism offers a "bottom-up" explanation of reasoning *via* the formality thesis, Bayesianism rejects any assumption related to formality in cognition and develops a top-down strategy that takes reasoning as a situated activity, where the goals of the cognitive system and the informational structure of the environment are accounted for to derive an "optimal behavior function" (Oaksford & Chater, 2003) that will serve to evaluate our rational performance:

> So the idea...is to understand the problem that the cognitive system faces, and the environmental and processing constraints under which it operates. Behavioral predictions are derived from the assumption that the cognitive system is solving this problem, optimally (or, more plausibly, approximately), under these constraints. The core objective of rational analysis, then, is to understand the structure of the problem from the point of view of the cognitive system, that is, to understand what problem the brain is attempting to solve. (Oaksford & Chater, 2009, p. 72)

Reasoners are conceived here as "information optimizers." The point is not whether they arrive or not to valid conclusions, but if they manage to conclude the most probable statement from their degrees of belief about on the premises. Naturally, this requires to assume that instead of logical rules, agents implement some Bayesian probability rules while thinking.

Most of the criticism that Bayesian theories of reasoning make to "logicist cognitive science" converge with the ideas defended here (Oaksford & Chater,

1989, 1991, 2007). However, while anti-formalist, Bayesianism does not say much about the role of conceptual content in reasoning. Their framework does not even make use of —or need— notions like *lexical concept* or *concept-based inference*. Consequently, this approach will not play a role in the following chapters. We will only return to it briefly in chapter 7, while we discuss category-based induction.

1.4 Summary and conclusions

The formality thesis has played a central role in the development of cognitive science in general, and the psychology of reasoning in particular. In this chapter, the origins of this thesis, as well as its interpretations within theories of reasoning, were discussed. We saw that besides the traditional normative role that classical logic was supposed to have over reasoning, certain ideas about logical form and logical validity received different psychological interpretations and took part in the very explanatory —descriptive— structure of these theories.

The crisis of the formality thesis in psychology started with a series of empirical findings showing that logical performance of people was consistently poor. As a consequence, two questions undermined the idea that logic could work as a model of inferential competence: why are we bad at logical tasks if we reason following logical rules? And, how is that rational inference is affected by content and context if reasoning is syntactic-driven?

A prominent answer to these questions consisted of negating the core assumption of the formlity thesis. MMT emerged as an alternative proposal that did not explain reasoning as syntactic rule-following. Instead, reasoning was conceived as a semantic-based mechanism working by looking for counterexamples instead of by rule-application. However, we saw that even this theory was still responding to the formality thesis and, in this sense, it continued within the domains of logicism.

This chapter delved into the origins and development of the idea that reasoning is a formal —non-semantic— mechanism. But that is only one side of the story. In the next chapter, I will discuss what I consider the other source

of this idea: a "divorce" between inference and meaning that was promoted by philosophical semantics, also due to the great influence of logic in this discipline. In short, there was a view of inference as something not related to meaning. At the same time, there was a view of meaning as something totally independent to inference —or any other cognitive mechanism.

Chapter 2

Meaning and cognition

Summary

This chapter focuses on the relationship between meaning and inference within the framework of semantics. It discusses how and why mainstream analytic philosophy construed *meaning* as independent of *inference*. Conceptual and inferential role semantics are later analyzed as alternatives to this view. Finally, cognitive semantics is defended as the proper framework to account for this relationship.

2.1 The formalist turn in semantics

In the previous chapter, we have analyzed the tradition in philosophy and psychology that studies reasoning as independent of meaning. In what follows, we will discuss the other side of the coin: the tradition in analytical philosophy that understands the notion of meaning as completely independent of reasoning — and of cognition in general. I speak of a "formalist turn" in semantics because I consider these ideas to be the result of the efforts of some philosophers to provide a semantic framework that fits the requirements of classical logic.

2.1.1 From *intensions* to *extensions*

Both semantics and logic have to do, essentially, with relations between properties and objects in language. The issue of which of these two notions is more fundamental than the other has divided philosophers for centuries (Rescher, 1959; Swoyer, 1995). Those who believe that objects are explanatory prior to

properties are *extensionalists*; those believing otherwise endorse some form of *intensionalism* (see Bar-Am, 2008). The terms *intension* and *extension* comes from Carnap's work o semantics (Carnap, 1988), but the idea behind them has a long history. It concerns the basic distinction between what an expression *means* and the object(s) it *denotes*.[1] For instance, the term "dog" denotes objects in the world, but its meaning consists of ideas about what dogs are, i.e., the cluster of properties and concepts that characterize dogs according to some linguistic community. This distinction is relevant for multiple reasons; a particularly salient one is that expressions with different meanings can have the same denotation. For instance, "4+5" and "3x3" denote the same object but mean different things.

Maybe the most influential extensionalist, both in logic and semantic, was W.v.O. Quine. Quine famously promoted a "flight from intensions" (Quine, 2013, Ch. 6) since he considered this notion as obscure and theoretically ill-conceived (see Harman, 1967). His entire program consisted of reducing the notion of *predicate* to reference —sets of objects— and thus building a semantics that is free from any mentalistic, Platonistic or metaphysical view of meaning and predication. Quine's path to radical extensionalism was a reaction to a tradition starting from Frege assuming the intensions as central elements for analyzing meaning (see Parsons, 2016). As Jerrold Katz (1992; 2004) convincingly showed —and I will later explain— this discussion has a lot to do with developing the classical notion of logical form and the formalist view of inference that we explained in the previous chapter.

As the story goes, intensionalism in analytic philosophy started with Frege's idea that a proper analysis of meaning must distinguish between the *sense* of an expression and its *reference* (Frege, 1948). For instance, the sentences "Manchester United's top goalscorer of all time" and "The captain of England's national football team in 2016" refer to the same person —Wayne Rooney— , but are clearly different from an informational perspective. Frege famously claimed that the sense of an expression contains the "mode of presentation" of

[1] The Port-Royal school used the terms "comprehension" and "denotation" for this distinction. John Stuart Mill used "connotation" and "denotation." Frege famously used "sense" and "reference."

its denotation. Senses have different cognitive values, even when they refer to the same entity.

It is well known that Frege defended *senses* in his semantics because of their role in explaining *identity* and *analiticity*, two crucial elements for his logicist program (see MacFarlane, 2002). In particular, the notion of sense allowed Frege to abandon Kant's notion of the analytic as *concept-containment*, i.e., the idea that in analytic propositions, the concept in the predicate is *contained* in the concept in the subject. Against Kant, Frege proposed a logic-based notion of analyticity: a judgment is analytic if it can be proven by logical laws plus definitions (Frege, 1980, §3). Changing this notion allowed Frege to argue that arithmetic was analytic and not synthetic a priori, as Kant held (see MacFarlane, 2002, or Hanna 2004, Ch. 3.3).

An often overlooked aspect of Frege's semantics is the explanatory relation between senses and references. As Katz (1992) claims, Frege's intensionalism is quite weak, since *denotation* is explanatory prior to *sense*:

> ...our understanding of the notions of sense and reference must come from an account of reference, just as our understanding of the notions of employer and employee must come from an account of hiring. The sense/reference distinction is then a distinction within the theory of reference, between the instruments of reference determination (senses) and the objects which those instruments determine (referents). It is not lost on a Fregean reduction, but rather recast as a distinction within the reducing theory. (Katz, 2004, p. 13)

Katz' point is that even if Frege is considered the main figure of intensional semantics, his ideas were, in fact, the sources of extensionalism, since his theory of sense is reducible to its theory of denotation. That is mainly because he conceives senses as mere means for reference determination. Thus, all that is left to semantics is to explain the referential side of language, while logic keeps the monopoly of its inferential structure.[2] The upshot of this view is the

[2] A similar analysis can be found in Diego Marconi's book "Lexical Competence" (1997), where he shows how traditional semantics was focus on explaining our "referential competence" while taking inference as something pertaining only to logic.

reaffirmation of the hylomorphic tradition in logic: only logical constants have inferential properties.

Now, according to Katz's historical reconstruction, Frege's semantics encountered its first problem when Wittgenstein tried to apply it in the Tractatus. Broadly speaking, Frege's point of view typically asserted that logical relationships between sentences were a function of their logical structure. Thus, atomic sentences could not be related to other atomic sentences through their internal structure. In this sense, Wittgenstein writes that "*It is a sign of an elementary proposition, that no elementary proposition can contradict it*" (Wittgenstein, 2001 [1921], 4.211). In other words, Wittgenstein claimed that atomic sentences cannot have inferential properties, because inference is the result of truth-functional relations (see Rosenberg, 1968, for a detailed explanation). However, he realized that some atomic propositions seemed to be in obvious relations among to each other. For instance, the sentences "The spot x is blue" and "The spot x is red" are in contradiction even if they are both atomic. Wittgenstein tried to justify this by saying that it was due to "the logical structure of colour" (Wittgenstein, 2001 [1921], 6.375 and 6.3751). However, according to Katz, he later saw that this kind of relations between atomic propositions needed a more in-depth explanation, and he ended up by abandoning the Fregean program in semantics altogether and proposing a radically different view of meaning. [3]

The general problem here is that of explaining the inferential properties of lexical concepts in natural language. The issue directly connects semantics with logic because semantic relations between atomic propositions translate into inferential (or *entailment*) relations —e.g. "Fido is a dog" licences the inference "Fido is a mammal" because *dog* is an hyponym of *mammal*. Since logic is supposed to study inferential relations among propositions, it should account for these kinds of intuitive implication-relations.[4] However, these entailment relations are not captured in the classical notion of validity —as we saw in the previous chapter— for the sake of preserving formality. A particular case of

[3]For an analysis of Wittgenstein's color-exclusion problem, see (Sievert, 1989).

[4]For a discussion about this on purely logical grounds, see (S. Read, 1994; Sagi, 2018).

this problem —central to the semantic tradition we are considering now— is how to account for the class of sentences that seem to be true in virtue of their meaning and not in virtue of their form. This is the problem of *analyticity*, the backbone of the discussion about meaning in analytic philosophy from Frege to Quine (see Boghossian, 1996).

As the story goes, Carnap attempted to save Frege's intensionalism with his theory of meaning postulates. In particular, Carnap saw that Frege's notion of analiticity was problematic because its reliance on a vague notion of *definition* (Carnap, 1988). In face of this issue, Carnap proposed to *explicate* analytic entailments between atomic sentences in a formal system by introducing sets of *meaning postulates* as axioms of the system (Carnap, 1952). As an example, suppose that I want my formal system to capture the analytic —conceptual— relation between *bachelor* and *not married*. Then, I can introduce as an axiom in the system the following formula: $\forall x(Bachelor(x) \rightarrow \neg Married(x))$, and use it in proofs.

Carnap's meaning postulates theory attempts to capture the inferential powers of extra-logical terms in a language. Something that he considered epistemologically relevant since these relations were supposed to reflect the conditions that objects must meet in order to be denoted by an extra-logical term, thus playing a role in a possible explanation of our referential competence (Carnap, 1955, p. 34). However, Carnap's explication of analiticity was also committed to an extensionalist program, since he made clear that meaning postulates must be understood as restrictions on the extensions of predicates. For instance, a meaning postulate like $\forall x(Bachelor(x) \rightarrow \neg Married(x))$ stipulates that any individual within the extension of *bachelor* is not excluded from the extension of *married*. Consequently, even if Carnap tried to keep the distinction between *intension* and *extension*, his semantics was focused on *objects*, rather than on *properties*, and the notion of reference was considered —again— as explanatory prior to any other semantic notion.

Carnap's meaning postulate theory was a significant attempt to integrate inference with meaning. However, Carnap avoided any commitment to more in-depth interpretations of this idea and presented it mostly as a technical or

methodological solution to the problem of analyticity in logic. For instance, he leaves open the question about the origins of these postulates, particularly, to which extent they mirror forms of linguistic usage or other kinds of relations between concepts. Meaning postulates had much more impact in linguistics than in philosophy, through the development of formal semantics in the 1970s (see Partee, 2014; Zimmermann, 1999).[5]

The final attack to intensionalism —and to the idea that extra-logical terms can have inferential properties— came from Quine. He thought that Frege's and Carnap's efforts to define analyticity were pointless because the analytic/synthetic distinction was fundamentally wrong. One of his arguments consisted of showing that the notion of *analytic proposition* requires the notion of *synonymy* (Quine, 1951).According to him, this last notion cannot be systematically articulated without referring to *synthetic propositions* expressing matters of fact. Consequently, there is no possible clear-cut between analytic and synthetic propositions. In a few words, Quine thought that our philosophical intuitions about the intensional dimension of meaning were impossible to articulate in a systematic theory. The upshot was that the idea of atomic sentences establishing analytic relations was theoretically useless (For a detailed explanation, see Decock, 2010; Gaudet, 2006).

2.1.2 Extensionalism and logical form

By reducing meaning to reference, extensionalists block any possible way of connecting meaning to inference. They provide the hylomorphic tradition with a semantic framework reaffirming the idea that inference is a matter of *form*, and not of *content*. Quine's views on logical implication confirm this:

> Logical implication rests wholly on how the truth functions, quantifiers, and variables stack up. It rests wholly on what we may call,

[5]Formal semantics is an influential framework for analyzing sentential meaning based on its compositional structure. It directly builds on the extensional theory described above, where denotation and truth-conditions are enough for characterizing meaning. Since this framework left lexical meaning unanalyzed, it is not useful for tackling the problem of this thesis, i.e., how concept possession underlies inferential competence. In particular, formal semantics is not about *how we understand* meanings, but about *what is computed* during language processing (see Johnson-Laird, 1982, p. 15).

> in a word, the logical structure of the two sentences. (Quine, 1986, p. 48)

Katz claims that the whole issue rests on something he calls the *extensionalist dogma*:

> The article of faith is that there exists a justifiable distinction between the logical and nonlogical components of sentences, one that enables us to divide a theory of connectives and quantifiers from a theory of the meaning of nouns, verbs, adjectives, etc., that form the expressions and sentences they connect and quantify. (Katz, 1975, p. 77)

Again, the extensionalist story about linguistic meaning builds on the same principle than the one behind the hylomorphic tradition: inference is formal and it is about truth-transmission, while lexical meaning is about reference, and truth-determination.

Intensionalists, on the other hand, claim that inferential relations between extra-logical terms are crucial for understanding natural language inference . As a consequence, they promote a richer notion of *logical form*:

> lntensionalists understand the logical form of a sentence to consist of every property of the sentence that determines the role it plays in valid arguments. For the intensionalist, any feature of a sentence S that is part of S's grammatical structure and by virtue of which S occurs as an essential premise (or the conclusion) of a valid argument (in the standard sense of one whose conclusion must be true if its premises are)M is a feature of S's logical form. In Frege's terms: whatever "influences its possible consequences. Everything necessary for a correct inference ... " (Katz, 1975, p. 76)

However, the intenstional program in the philosophy of language has been overshadowed by the success of the referential tradition, which found a solid backup in the technical success of model-theory in logic and formal semantics in linguistics.

Extensionalism found an important ally in the anti-mentalist trends in philosophy of language, claiming that semantic properties are completely independent of mental properties. In what follows, I briefly discuss these ideas and their contributions to the "divorce" between meaning and inference.

2.2 Anti-mentalism in semantics and the divorce of meaning and cognition

Before the mathematical revolution in logic, semantics leaned toward mentalism. The prevailing view was that the meaning of a word was the idea regularly associated with it. According to this tradition, language was a tool for making these ideas public and communicating them to others.[6]

The mentalist tradition was mostly abandoned during the logical revolution of Boole and Frege. According to Stephen Land (1974), this was due to a gradual change in what philosophers considered to be the *units of meaning*. The change went from the Lockean *idea* to the notions of *proposition* and *sentence*. This change affected both the notion of grammatical structure and the notion of logical form. As Quine's similarly observed, the inflection point here can be traced back to Bentham's *Doctrine of paraphrasis*, who understood that sentences (and not terms) had to be "the primary vehicles of meaning." The central consequence of this change of focus, as Quine also claimed, was a new agenda for epistemology, from *concepts* to *truth* and *justification*. (Quine, 1981, p. 70).

[6]Locke was the central figure of this *ideational* view of meaning (see Hanna, 1991). In his "Essay Concerning Human Understanding" he says: "*The use, then, of words is to be sensible marks of ideas; and the ideas they stand for are their proper and immediate signification*" (Locke, 1979, III.2.1, 405). For Locke, successful communication implied that the hearer decodes the speaker's words' references into her associated ideas. One advantage of ideational semantics was that it could easily relate meanings and language with other cognitive faculties. For instance, Locke's theory of demonstrative inference made essential use of his semantic theory. Just like Descartes, Locke was not convinced by the Aristotelian formalist approach, and he explained reasoning as based on intuitions about the connection between ideas. Complex reasoning implied to construct chains of ideas in such a way that the connections between them were made explicit for intuition (see Owen, 1999, Chapter 3). As we will see, the *spirit* of the view defended in this work is very much in this line.

This was also the start of a strong trend of anti-psychologism in semantics (see Elffers, 2014). In other words, breaking the association between meaning and ideas was the first step towards taking meaning "out of the heads" and re-locating it in the abstract relation between language and the world. Frege, who was strongly anti-psychologistic in logic, assumed that meanings were completely mind-independent. Even with the intensional layer of "senses", meanings were to be found in the relation between thoughts —abstract propositions—, objects and truth-values; and not in the mind or any cognitive-related phenomenon.[7]

Again, one of the leading figures in this trend was Quine, who was a behaviorist regarding linguistics. He thought that mentalism in semantics was not scientific, and that any construal of *meaning* in terms of *ideas* or any intentional entity would "end up as grist for the behaviorist's mill." (Quine, 1969b, p. 26). His arguments against mentalist semantics —and linguistics in general— were based on a combination of methodological ideas about how to develop empirical linguistics and philosophical arguments against intensional notions —e.g., *ideas*, *thoughts*, *meanings*, or *propositions*.[8] To name one, he claimed that intensional notions could participate in opaque expressions that did not admit quantification, blocking the possibility of a truth-functional analysis. In other words, he thought that since the classical tools of logic were not able to explicate intensional expressions, they must be eliminated from our theories of meaning.

While Quine's concerns leaned towards the methodological and ontological issues of mentalism, Putnam offered a different argumentative strategy to prove a similar point. According to Putnam, the semantic properties of expressions are mostly determined by external factors to the speaker-hearer. Roughly, his argument shows that it could be possible for two subjects to have the same psychological states associated with two different expressions and yet these expressions could have different meanings if the extensions of the terms they use

[7] Similar ideas were dominant within the neo-positivist tradition, which, as Carnap explicitly suggested (Carnap, 1938-55, p. 46), avoided to use the term "concept" in the philosophical analysis of language for being "too psychological."

[8] Quine went as far as to say that intensions were "creatures of darkness" (Quine, 1956, p. 180) and that "obscurity is the breeding place of mentalistic semantics" (Quine, 1969b, p. 28).

differ. He then concludes that meanings are not determined by psychological facts —whatever they are— as the internalists claim, but by certain causal relations between language and the world (see Putnam, 1975; also Wikforss, 2008).[9]

Semantic externalism was extremely popular and contributed to overshadowing the epistemic and cognitive dimensions of meaning. However, by exclusively focusing on reference and truth, we risk ending up with a myopic semantics which lacks the tools for answering a series of crucial questions about the nature and use of meaning and language. As Michael Dummett argued in his influential paper "What is a Theory of Meaning?" (Dummett, 1993, Ch. 1), reducing meaning to reference and truth-conditions involves abandoning epistemological issues like how do we understand the meaning of expressions, or what is the relationship between meaning and public language usage. It seems then that if one of the goals of philosophical semantics is to explain linguistic competence, we have to look beyond truth-conditional theories.

2.3 Back to cognition: meaning, inference, and understanding

2.3.1 Conceptual and inferential role semantics

The explanatory limitations of classical theories of meaning led to the development of "conceptual role semantics" (CRS), a family of views claiming that meanings emerge from the role of lexical concepts —and propositional attitudes— within the complex cognitive ecology of agents (Block, 1986; Brandom, 1998b; Harman, 1982). There is no consensus on which specific cognitive

[9]Putnam's defense of semantic externalism is often understood as an attack to psychological semantics. However, Putnam's view is not as radical as Quine's. His concern regard the stability of reference across conceptual frameworks. He believes that if we assume that reference is determined by intensional entities or by epistemic properties, then reference will we subjected to constant variability. However, it seems that people with completely different background knowledge can still refer to the same objects, just like scientific theories change but yet the entities they talk about remain. Putnam's conclusion is that even if reference does not exhaust meaning, it has to be the essential notion for semantics. Anchoring meaning in reference would be the only way to protect semantics from relativism.

processes are meaning-constitutive, but attention has been mainly focused on reasoning, perception, and categorization (see Greenberg & Harman, 2005; Sellars, 1953).

CRS is considered a *use* theory of meaning; that is, a theory based on the assumption that the content of linguistic expressions depends on how agents use them within a linguistic community. In this sense, CRS downplays the notions of *truth* and *reference* in the explanatory structure of philosophical semantics. The central questions of semantic theory are no longer about how expressions refer to things in the world, but about their functional/causal use in agents' mental life within a language community. However, how to specify the notion of *use* is a major challenge for CRS. For instance, the lexical item "dog" participates in many different contexts of use. We use it to talk about food, categorize animals, or as a pejorative term in specific dialogical contexts. If we take into account every context of use as meaning-constitutive, as well as all the cognitive mechanisms they mobilize, we risk not being able to individuate the content of the term at all (cf. Fodor & Lepore, 1991). For dealing with this issue, CRS theorists restrict this notion to individual cognition:

> There are three broadly different ways in which symbols can be used—in communication, in speech acts like promising that go beyond mere communication, and in thinking. CRS takes the last of these uses, the use of symbols in thought, to be the most basic and important use for determining the content of symbols, where that use includes (at least) perceptual representation, recognition of implications, modeling, inference, labeling, categorization, theorizing, planning, and control of action. (Greenberg & Harman, 2005, p. 270)

Although this is an attempt to delimit the notion of *use*, it still seems too broad to specify the functional/causal role of concepts in the cognitive mechanisms listed above. What is more, if meaning is to be explained in terms of the above listed psychological abilities, then semantic theory must be explanatory dependent on cognitive psychology. However, CRS theorists did not make

any attempt to articulate this explanatory relation. The upshot is that CRS lacks the analytical tools for a fine-grained explanation of the relation between meaning and cognition. As Ned Block recognized (Block, 1998a), CRS looks more like a general framework for developing a theory of meaning than a proper semantic theory.

Maybe the most systematic effort to develop a CRS theory was made by Robert Brandom, building on Wilfrid Sellars' ideas. Instead of conceiving several cognitive mechanisms as meaning-constitutive, Brandom focuses on inference. "Inferential role semantics" (IRS) affirms that the meaning of a concept is determined by the set of inferential moves in which the concept participates in. For instance, "Fido is a dog" licenses inferences like "Fido is a mammal," "Fido is warm-blooded," "Fido barks," and can also be the consequence of sentences like "Fido is a German shepherd" or "Fido is a Chihuahua," among others. In this sense, the meaning of a lexical item can never be specified in isolation —for instance, via a relation between the concept and something in the world—, but it is a function of its role in a broader group of concepts (Brandom, 1998b, 2000).

The central strategy of Brandom's semantics is to explain meaning in terms of relationships of material validity and incompatibility. Again, truth-conditions and reference are marginal notions in this approach. As Brandom explains:

> The standard way [of traditional semantics] is to assume that one has a prior grip on the notion of truth, and use it to explain what good inference consists in....[IRS] reverses this order of explanation also. It starts with a practical distinction between good and bad inferences, understood as a distinction between appropriate and inappropriate doings, and goes on to understand talk about truth as talk about what is preserved by the good moves. (Brandom, 2000, p. 12)

For Brandom, meaning is based on a normative structure that is socially regulated. The context in which linguistic normativity takes place is discursive interaction. That implies that sentence meaning have priority over lexical

meaning because assertion is the first speech act that allows making explicit our conceptual commitments (Brandom, 1998b, p. 79). The content of declarative sentences unfolds according to the inferences we make from them, and this is normatively evaluated by other agents during interaction. A crucial point of this approach is that *pragmatics* become central to semantic theory (Brandom, 1998b, p. 83-84). As MacFarlane explains, within Brandom's framework "*the fundamental semantic concepts can be defined in purely pragmatic terms*" (MacFarlane, 2010, p. 89).

One of the most controversial implications of the IRS is meaning holism, since lexical and sentence meaning depend on large bodies of concepts —or propositional attitudes:

> ...inferentialist semantics is resolutely holist. On an inferentialist account of conceptual content, one cannot have any concepts unless one has many concepts. For the content of each concept is articulated by its inferential relations to other concepts. Concepts, then, must come in packages (though it does not follow that they must come in just one great big one). (Brandom, 2000, pp. 15-16)

Semantic holism has various problems (Fodor & Lepore, 1992). For instance, it is hard to explain concept acquisition and conceptual competence from a holistic perspective. How is that a subject learn its firsts concepts if possessing one requires having others previously? How is it possible to have full conceptual mastery considering that everyday cognition is seriously limited, and individual knowledge is never complete? (e.g., see Fodor, 1994; Jönsson, 2014).

Another important problem of IRS concerns concept individuation. Fodor and Lepore (Fodor & Lepore, 1992) gave strong arguments against the possibility of conceptual individuation for holistic views of meaning —i.e., the meaning of an expression depends on the entire body knowledge. One popular solution to this is local holism, which claims that only a closed set of inferential relations between expressions are meaning-constitutive (Block, 1998b; Weiskopf, 2009). For instance, the concept *bachelor* would be individuated though inferential relations with concepts like *man*, *married*, *civil status*, and *young*; while a logical

concept like *and* would be individuated through its typical introduction and elimination rules —i.e., $p \wedge q$, then p; $p \wedge q$, then q; p and q, then $p \wedge q$— (see Peacocke, 1992, pp. 6-8). Nevertheless, Fodor and Lepore still think that no demarcation criterion between meaning-constitutive and non-constitutive inferences would be adequate, due to Quine's arguments against analyticity and synonymy. This problem opened a discussion that continues until today. I will suggest a demarcation criterion in this sense in Chapter 5.

CRS and IRS are good attempts to build a bridge between meaning and inference. A bridge that was destroyed by the anti-mentalistic trends in the philosophy of language and by the success of truth-conditional semantics. However, I do not think that any of these theories allow us to build fine-grained explanations of this relation. As Machery claims, the discussion about concepts in philosophy is relatively isolated from the discussion in psychology and linguistics (Machery, 2009, p. 3). The questions that philosophical semantics try to answer are different from those that psychologists and linguists worry about. Philosophers wonder about concept identity, individuation and possession conditions. Psychologists, on the other hand, disregard foundational issues in favor of analyses that try to understand the role of concepts in cognitive functions like categorization, learning, and reasoning (Carey, 2000; G. Murphy, 2004).

Block's seminal paper, "Advertisement for a Semantics for Psychology" (1986), proposed CRS as a theory of meaning for psychological theorizing. I think that the strategy for semantics should be the other way around, a philosophical theory of meaning must build on psychological explanations. Especially if what we want to elucidate is the relation between meaning and reasoning.

I believe that inference and concepts are two sides of the same coin and that we cannot understand any of them in isolation. In this sense, I will look into psychologically informed semantics to find conceptual tools for analyzing reasoning with concepts. I will argue in favor of Cognitive Semantics as the best candidate for this task.

2.3.2 Cognitive and conceptual semantics

As it is well known, the cognitive revolution in linguistics came with Chomsky's generative grammar (Chomsky, 2014 [1965]). Chomsky's work was the main source behind the recovery of the notion of *representation* in cognitive science (Chomsky, 2005), which was almost forbidden due to the influence of behaviorism. Chomsky's review of B.F. Skinner's book "Verbal Behavior" (Chomsky, 1959) —a milestone in the history of cognitive science— showed that behavioral linguistics was implicitly based on intentional notions, and that assumed an intuitive recognition of linguistic structure, besides its anti-mentalistic rhetoric.

Chomsky was one of the leading promoters of the return to internalism. He was a fierce critic of semantic externalism (see Pietroski, 2017) and argued against the possibility of studying language as a structure independent of knowledge and cognition:

> We should, so it appears, think of knowledge of language as a certain state of mind/brain, a relatively stable element in transitory mental states once it is attained; furthermore as a state of some distinguishable faculty of the mind – the language faculty – with its specific properties, structure and organisation, one module of the mind. (Chomsky, 1986, pp. 12-13)

He provided a new methodology for studying language which replaced the frameworks of Bloomfield and Harris —Quine's central influences. It successfully combined empirical, theoretical and formal considerations; and was also deeply engaged with cognitive psychology. Within this new framework, understanding linguistic competence is required to study the nature of those cognitive structures that make possible language development and acquisition (Cummings, 2013, p. 53). Chomksy showed that it was possible to build a scientific theory of language that used intentional and representational notions, and which was at the same time situated at the core of empirical science: language was assumed to be a natural object, a component of the human mind, physically represented in the brain and part of the biological endowment of the species (Chomsky, 2002, Chapter 1).

Chomsky's work focused on syntax. However, his methodological program and theoretical commitments quickly reached semantics and linguistics in general. A new field of study emerged thanks to his influence: Cognitive Linguistics, a research program whose main objective was to develop an analysis of language as an information system that mediates our interaction with the world through the cooperation of several cognitive faculties such as perception, categorization, reasoning and memory. As Geeraerts and Cuyckens explain:

> ...where "cognitive" refers to the crucial role of intermediate informational structures in our encounters with the world. Cognitive Linguistics is cognitive in the same way that cognitive psychology is: by assuming that our interaction with the world is mediated through informational structures in the mind. It is more specific than cognitive psychology, however, by focusing on natural language as a means for organizing, processing, and conveying that information. Language, then, is seen as a repository of world knowledge, a structured collection of meaningful categories that help us deal with new experiences and store information about old ones. (Geeraerts & Cuyckens, 2007, p.5).

A subfield of cognitive linguistics, *Cognitive Semantics*, emerged in the 1970s carried out by George Lakoff, Ronald Langacker and Leonard Talmy, among others. Again, their central tenet was that *meaning* cannot be studied in isolation from the psychological structures involved in language processing and knowledge representation —memory, categorization, inference, perception, etc.— (see Geeraerts, 2010, Chapter 5).

Contrary to mainstream truth-conditional semantics, this view takes lexical meaning as explanatory prior to sentential meaning. And the notions of *reference* and *truth* play a relatively marginal role in its theoretical structure. The main idea of cognitive semantics can be summarized in the following motto: "meaning is conceptualization." In other words, semantic processing involves the constant mobilization of knowledge structures to decode lexical and sentential meaning. In words on Langacker: "*Semantic structure is conceptualization*

tailored to the specifics of linguistic convention. Semantic analysis therefore requires the explicit characterization of conceptual structure" (Langacker, 1987, p. 99). The building blocks of these knowledge structures —or as Quine would say, the "vehicles of meaning "— are not propositions, but notions like prototypes, frames, or image schemes, depending on the specific theory.

To see a quick example of how cognitive semantics approaches meaning, let us briefly see Lakoff's account of *polysemy*. According to Lakoff, lexical terms' meanings are constructed as complex radial structures organized around a composite prototype.[10] The members of this structure establish different kinds of relations to the prototype according to linguistic conventions, and remain in the *mental lexicon* for being used in thought and language processing (Brugman & Lakoff, 1988; Lakkof, 2017).

Semantic phenomena like *polysemy* can be easily explained within this framework. Consider, for instance, the lexical concept *fruit*. The sentences (a) "Apples are fruits" and (b) "My salary is the fruit of my work" display different senses of *fruit*. In Lakoff's view, all the different senses that a word can have depend on the conceptual (radial) structure of the category. Moreover, the different senses emerge as specific semantic relations with the prototype, like metaphoric meaning or generalization. In the cases above, (a) express a prototypical sense of fruit while (b) a metaphorical one (see Figure 2.1).

Polysemy is here a deep semantic phenomenon that reflects conceptual organization at the level of mental representation; and not a superficial linguistic phenomenon associated to usage.

Another theory of meaning in this line is Jackendoff's *conceptual semantics* (Jackendoff, 1992, 2002). Jackendoff's influential approach shares many theoretical commitments to cognitive semantics, with the difference that he remains committed to the generative program. Jackendoff supports the idea of

[10]Lakoff follows Rosch's ideas about the prototypical structure of concepts (Rosch, 1983). Categories are not homogeneous, since some of their members are more typical than others and, consequently, more representative of the category in question (e.g., *robin* is a prototypical bird while ostrich is rather atypical). We will come back to this idea in Chapter 4 since it is central to our work.

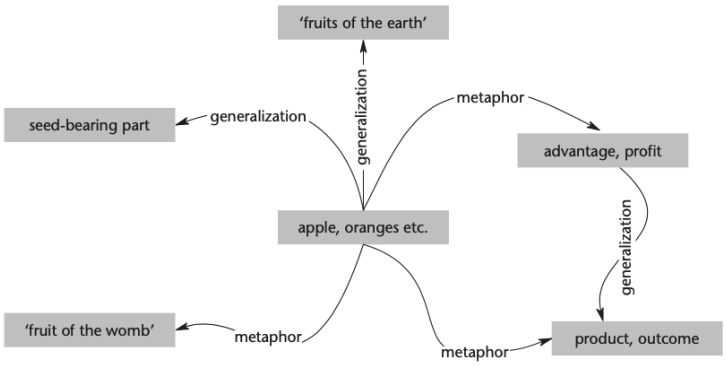


FIGURE 2.1: Radial structure of *fruit*. The links among senses represent semantic relations. From (Geeraerts, 2010, p. 195).

a language faculty with a modular architecture. Different structures —such as syntax and phonology— cooperate in semantic processing (Figure 2.2).

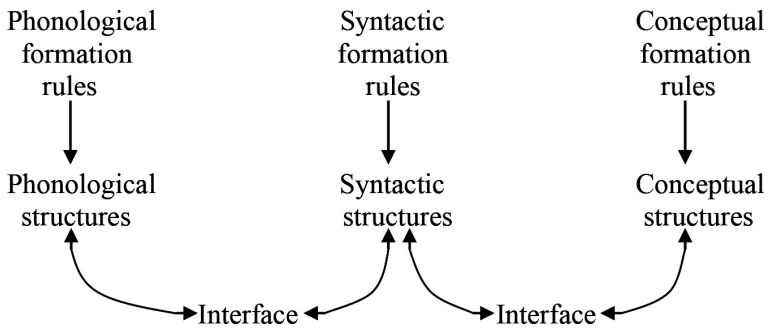


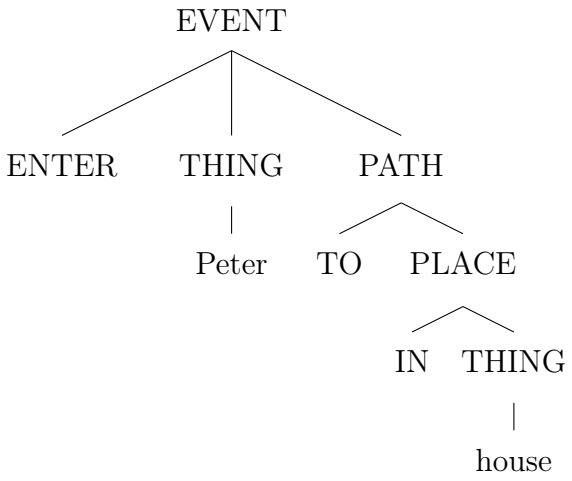
FIGURE 2.2: The structure of the language faculty according to Jackednoff. From (Jackendoff, 2017).

Regarding meaning, the crucial module is the conceptual structure: an autonomous level of cognitive representation in which concepts are interpreted in terms of sets of semantic primitives. Conceptual structure is an interface between other linguistic modules such as phonology and syntax, and non-linguistic structures representing perceptual information —such as vision:

> Conceptual structure is not a part of language per se – it is a part of thought. It is the locus for the understanding of linguistic utterances in context, incorporating pragmatic considerations and "world knowledge"; it is cognitive structure in terms of which reasoning and planning take place.(Jackendoff, 2002, p. 123)

Jackendoff's theory is decompositional: it assumes that complex concepts are analyzable in terms of conceptual primitives, i.e., a set of basic building blocks of meanings combined through innate rules that constitute a *conceptual grammar*. Conceptual primitives are concepts like "thing", "place", "path", "property", "event" and "action". Sentence-structure has a parallel architecture to phonological, syntactic and semantic structure. For instance, the sentence "Peter entered into the room" has the following semantic structure:

$$[_{Event}ENTER([_{Thing}Peter]_i,]_{Path}TO([_{Place}IN([_{Thing}ROOM]_m)])]_k)]_j$$



In general, Jackednoff proposes that we can uncover the underlying semantic structure of every meaningful sentence with this kind of strategy.

2.3.3 Inference and meaning structure

Both cognitive and conceptual semantics offer interesting –and psychologically informed— ways of relating inference with meaning. George Lakoff, for instance, claims that:

> The heart of metaphor is inference. Conceptual metaphor allows inferences in sensory-motor domains (e.g., domains of space and objects) to be used to draw inferences about other domains (e.g., domains of subjective judgment, with concepts like intimacy, emotions, justice, and so on). Because we reason in terms of metaphor,

> the metaphors we use determine a great deal about how we live our lives. (Lakoff & Johnson, 2008, p. 245)

In this account, metaphors play a crucial role in reasoning by allowing to exploit inferential patterns from one domain to another domain *via* conceptual mappings. However, Lakoff did not developed formal tools for explicating these mechanisms.

Jackendoff, on the other hand, uses his formal analysis of meaning structures to explicate semantic-based inferences. This problem has a central role in his theory. He writes that "*one of the requisites of an adequate semantic theory is that it provides an account of entailment between sentences*" (Jackendoff, 1976, p. 110). In the same spirit as Katz's idea that deductive inference is related to grammatical form (broadly construed), Jackendoff claims that lexical types have associated inference patters. These patters can be generalized by analyzing the structure shared by the decompositional analysis of sentences with the same type.

Let's see an example of this from Jackendoff (1992, p. 39). Consider the following causal inferences:

> X killed Y → Y died.
>
> X lifted Y → Y rose.
>
> X gave Z to Y → Y received Z.

The semantic structure of these sentences is:

> X kill Y: $X_{cause}[Ydie]$
>
> X lift Y: $X_{cause}[Yrose]$
>
> X gave Z to Y: $X_{cause}[YtoreceiveZ]$

And the generalized inference pattern for causal sentences is the following:

> $X_{cause}[Etooccur] \rightarrow Eoccurs$

This kind of analysis that uncovers inferential patterns associated with meaning structures can be carried out systematically with other lexical types in this framework.

That is very much in line with the approach to concept-based inference developed later in this work. However, instead of using Jackendoff's generative grammar, I will use Gärdenfors' conceptual spaces. This latter theory is better equipped, both formally and theoretically, to explicate these kinds of inferences. Furthermore, Jackendoff's theory makes a central use of the notion of *rule of inference* (Jackendoff, 1976, Sec. 4), while the approach defended in this thesis avoids this notion since it could be problematic from a philosophical perspective (see Hlobil, 2014; Valaris, 2017).

2.4 Summary and conclusions

In this chapter, I discussed some of the reasons underlying the explanatory gap between meaning and inference in traditional philosophical semantics. Reducing semantic properties to reference and truth-conditions might be functional to the logicist program —which claims that the inferential structure of natural language is a matter of logical form— but it lacks the tools for explaining some evident entailment-relations between atomic statements —which are ultimately relations between concepts. This approach might help to shed some light into the mechanisms behind our referential competence (see Calzavarini, 2020, Chapter 3), but it has serious limitations at the moment of accounting for our *inferential competence* (cf. Marconi, 1997).

I argued that addressing the relationship between meaning and inference requires recourse to cognitive semantics. This framework takes the psychological dimension of linguistic meaning as the central feature of its explanatory structure. And as we will see later, it provides tools for analyzing the role of language in reasoning. I will return to it in Chapters 3 and 5. Next, we leave meaning aside for a moment and move on to the relationship between the more general notion of *representation* and its relation to *reasoning*.

Chapter 3

Inference and Representation

Summary

This chapter advance an analysis of the relation between representation and inference. It is claimed that inferential mechanisms exploit properties of the structures encoding conceptual information; and that the computational efficiency of these mechanisms depends on the "format" or organization of these structures. An important part of the chapter develops criticism to *representational conservativism*, i.e., the idea that all personal-level reasoning takes place in a language-like representational system like Fodor's *language of thought*. Finally, an pluralist framework for understanding inference is defended and some examples of different inference-types are discussed.

3.1 Introduction

Inference is a cardinal notion in several research fields. Philosophy, Cognitive Science, Computer Science, AI, and Statistics are some of the disciplines that have it in their conceptual repertoire. Inference is such a fundamental notion, that it often goes undefined.

In Cognitive Science, we can see a broad and narrow use of that notion. In the broad usage, inference is equated with computation, and it covers both personal and sub-personal mechanisms.[1] For instance, some researchers consider

[1] *Personal* mechanisms are those whose explanation requires the postulation intentional attitudes as part of their causal structure. On the other hand, *sub-personal* mechanisms can

low-level cognitive mechanisms —like perception— as inferential even if they have a completely different computational structure than high-level cognition (e.g. Aggelopoulos, 2015; Hatfield, 2002);[2] while others use inference only for referring to personal-level reasoning like deduction and induction (R. A. Wilson & Keil, 2001, p. 404).

In philosophy, the notion has been central to epistemology and logic, and it has evolved hand by hand with other notions like *judgment*, *justification* and *proof*. Philosophers have mostly thought about inference as a personal-level mechanism with three main features: it is language-based, it follows rules, and it can be normatively evaluated. Deductive reasoning has monopolized the attention since Aristotle. However, the discussion about non-demonstrative inference gained protagonist with Locke's and Hume's discussions on probable reasoning (see Owen, 1999). Nowadays, the standard view in philosophy accepts the existence of three kinds of inferential mechanisms that are computationally different: *abduction*, *deduction* and *induction*.

AI and Computer Science show a more homogeneous use of the notion, in general, anchored in the notion of *information*. Roughly, inference refers to processes carried out by programs that extract "new" information from a set of "facts" or suppositions represented in some programming language in the database of the program (e.g., Henschen, 1987). Researchers in this area developed a large amount of *inference engines* specialized in different kinds of inferential procedures, most notably, deductive, inductive, nonmonotonic inference. (e.g., Singh & Karwayun, 2010).

Besides broadly used, *inference* does not seem to have a stable meaning, neither within each particular discipline nor cross-disciplinary. In the following section, I inquire into the foundations of this notion to advance a possible definition. I put the emphasis on the relation between the notion of *representation* and *inference*; and I will argue in favor of a pluralist view of inference that

be explained without reference to intentional states as causes. The distinction comes from Dennett (1969).

[2] In the famous essay "Some consequences of four incapacities", Peirce (1868) proposes a notion of inference according to which every "operation of the mind" is an inference, including emotions and sensations.

builds on the idea that thinking can make use of different formats of representing information.

3.2 What inference is not

Inferences are transitions between mental states that take part in what it could be called —using William James' terminology— "the stream of thought".[3] Obviously, there is more to the *stream of thought* than just inferential transitions. We are all familiarized with different kinds of mental associations —between beliefs, perceptions, memories, and so on— that we would not consider *inferential*.For instance, I can have a personal disposition to think about guitars whenever I see a boat; or remember my childhood's home when I think about cats. These —idiosyncratic— mental transitions cannot be normatively evaluated because they do not answer any specific rationale, and they do not seem to follow any informational criterion. [4] Inferential transitions, on the other hand, are supposed to satisfy these last two points.

Mainstream philosophy thinks about inferences as ruled-based transitions between judgments. Frege, a crucial figure here (see Mezzadri, 2018), illustrates this in the following passage:

> The connections that constitute the essence of thinking differ in a distinctive way from associations of ideas. The difference does not lie in a mere ancillary thought [Nebengedanke] that is also present and that adds the justification for the connection. (quoted in Hlobil, 2019, p. 10)

Inferential transitions are not mere associations because the causal relationship between premise(s) and the conclusion has a non-idiosyncratic (normative) justification.

As we said earlier in this work, deductive logic was largely considered the normative model of this justificatory relation. Again, the reason for this can

[3] We could talk about *informational states* if we wanted to include inferences in artificial systems.

[4] This does not mean they cannot be explained from a psychological perspective.

be found in Frege's definition of inference as making "*a judgment because we are cognisant of other truths as providing a justification is known as inferring*" (Frege, 1979, p. 3). Frege's conceived logic as the set of laws governing truth transmission (see N. J. Smith, 2009). In this sense, the *epistemic anchor* of inference is truth: An agent infers ϕ from ψ if she takes ϕ to be true and recognizes a truth-functional relation between them. Thus, our inferential competence depends on our ability to grasp truths and truth-functional relations among propositions *via* our knowledge of logical laws.

As explained in Chapter 1, this is the core of the formality thesis. Logical rules apply to the *form* of the propositions expressed in natural language sentences because their syntax mirrors their semantic properties —understanding *semantics* as truth-conditions plus reference. We can see the endurance of this view within philosophy in the following passage of John Broome's influential book "Rationality through reasoning":

> In...reasoning, you operate on the marked contents of your conscious attitudes, following a rule. The marked contents are complex. They have a syntactic structure, and the rules you apply in operating on them depend on their structure. In operating on them, you have to hold them in your consciousness, maintaining an awareness of their syntactic structure. Language is well suited to doing that. It has a meaning that can represent the semantic elements of the marked contents, and it has a syntax that can represent their syntactic structure. It is plausible that, without the help of language, you could not keep the marked contents properly organized in your consciousness. (Broome, 2013, p. 267)

In recent years, Paul Boghossian took up the Fregean approach to inference emphasizing its intentional nature —that was already suggested by Frege— (Boghossian, 2014, 2018). Roughly, he argued that for a thought-transition to count as inferential —and not merely *associational*— the agent must *take* the premises as support for the conclusion. More precisely, the *taking condition* claims that any theory of personal-level inference must account for the fact

that inferential moves are carried out by the agent acknowledging a justificatory relation between the premises and the conclusion.

The Frege-Boghossian approach makes two interesting points about when a thought transition qualifies as inferential: it must follow a systematic criterion, and it must be an *action* that is the result of the agent's understanding of a justifying relationship between the premise(s) and the conclusion. However, although this seems like a good starting point, I believe that it is also clearly insufficient.

First, inferences indeed have to be the output of some stable information-processing mechanism. However, as it is widely known, deductive logic is not enough as normative criterion —as we saw earlier in this work. Most of our inferential activity is under uncertainty and does not follow logical rules in any classical sense (see Oaksford & Chater, 1989). Second, the *directionality* of reasoning is not exclusively from premise(s) to conclusion. For instance, belief revision and personal justification in dialogical contexts use different heuristics to finds reasons (premises) for the agent's claims or ideas (cf. Harman, 1986; Mercier, 2012). Third, our grasp of truth-functional relations between statements is hardly central to everyday thinking. We do not need to fully believe a proposition to use it in reasoning (see Staffel, 2013). We also make constant use of hypothetical situations while reasoning (Evans, 2019). And in many cases, our intuitions about premise(s)-conclusions relations are better explained by probabilistic models than by logical ones (e.g., see Elqayam & Over, 2012; Oaksford & Chater, 2007). The same problem affects the classical notion of logical validity based on truth-conditions. Recent Bayesian models of deductive reasoning have shown that some inferential patterns traditionally classified as formal fallacies are in fact reasonable inferential moves if they are evaluated from a probabilistic perspective (Eva & Hartmann, 2018; Hahn, 2020).

Another problem of the Frege-Boghossian approach is that it does not include any informational constraint for inference production. Transitions like "2+2=4. Thus 2+2=4 or Munich is in Germany" are genuine inferences if the agent takes the premise as reasons for the conclusions. Something that would be correct because the transition is logically valid. There is an infinite set of

logically valid implications like the one above that can be considered completely artificial and counter-intuitive because they lack any relevant cognitive use. A theory of reasoning must include constraints regarding the informational fruitfulness of inferences. Inferential moves are not produced in the void, but they respond to specific linguistic or physical stimuli within specific informational contexts (cf. Barwise, 1989; Gardenfors, 1994). We will come back to informational constraints in at the end of this chapter.

These are some of the problems of the Frege-Boghossian view of inference. Most of them converge with the critiques to the formalist thesis discussed in Chapter 1. In what follows, I discuss another important assumption of this tradition playing a central role in the mainstream computational view of thinking: The idea that thought transitions operate over propositionally-structured mental states; i.e., in a *language of thought*.

3.3 Representation

If, while arriving home, I see the door of the house open, I will infer that someone is inside. This thought transition has as premise a mental state that is itself a product of a complex computational process. It starts when light hits the retina and triggers a cascade of electrical signals that travel through the visual cortex, processing first low-level visual features, like orientation and spatial frequency, and then high-level features like shape and semantic category (see Bullier, 2001). The output of this process is a belief that I somehow relate to background information, to give rise to a new belief with the content "someone is at home." The whole process is a matter of information-processing. The key issue here is how the informational structure of the inputs are decoded and processed. When we focus on the process' outputs, the "information" jargon is often changed for a "representation" jargon.

Perceptions, beliefs, images, and memories are all structured representations that occur in our "mind's eye" and constitute high-level psychological processes like categorization, imagination, and reasoning. That idea has been at the core

of cognitive science since its very beginnings (Von Eckardt, 1995). From philosophy to artificial intelligence, every region of contemporary cognitive science has used —or at least discussed— a notion of representation in their explanations of cognitive phenomena. Due to this, the notion has been intensively discussed by philosophers in relation to its role on psychological explanation (e.g, Chemero, 2000; Dennett, 1979), its ontological status (Scheutz, 1999); and for its possible connection to more fundamental explanatory frameworks coming from neuroscience (see Shea, 2018).

The present work follows the mainstream tradition of cognitive science assuming that mental representation is a viable scientific hypothesis with important explanatory power, despite all the philosophical and methodological problems it brings. For reasons of extensions, I cannot cover the many facets of the issue. Instead, I will focus on one particular problem concerning the *format* of mental representation, which I believe it is deeply related to the problem of inference.

3.3.1 Representational conservativism and the translational approach

Analytic philosophy has been loyal to the following explanatory scheme for thought and reasoning: no matter the topic, the units of thinking are beliefs, and beliefs are language-like entities with logical properties —semantic and syntactic. Reasoning, *modulo* formality thesis, consists of transitions between beliefs generated by mechanisms that exploit their syntactic structure. As it is obvious, this is the fundamental idea behind the computational theory of mind (CTM) discussed in the first chapter. We will now revisit it, focusing on its use of the notion of representation.

The underlying hypothesis of the CTM is that the question about the format in which information is represented is directly connected to the computational structure of cognitive mechanisms. Computationalists, specially Fodor and Pylyshyn (Fodor & Pylyshyn, 2015), defend a "conservative" view of representational format: all psychological representation is based on an amodal

language-like system, the *language of thought* (LOT) or *mentalase* (Fodor, 1975, 2008). According to Fodor, LOT is "the only game in town" (Fodor, 1975, p.55), i.e., the only plausible hypothesis for building a scientific psychology, because the LOT has the right properties for explaining the *productivity* and *systematicity* of rational thought.

This idea requires a *translational* view of mental representation (see Clark, 2006; Landy, Allen, & Zednik, 2014). Roughly, all the different modalities of information —visual, auditory, tactile, and so on— that the brain processes to feed cognitive mechanism must be translated into the LOT in order to be used by high-level cognition. As Landy et al. (2014, p. 3) explains:

> Computationalist and semantic processing accounts of symbolic reasoning are equally translational because they both assume that problem representations are passed from a perceptual apparatus to an internal processing system in a form that is no simpler than the external —notational or linguistic— problem representation. That is, they assume that all transformations that involve changes in semantic structure take place "internally," over Mentalese expressions.

Since the 1970s, there has been an ongoing debate between representational conservatives and representational pluralists (cf. Dove, 2009; Fodor, 1975; Pylyshyn, 1981; Simon, 1978). The central issue here is how can representational conservatism explain that thinking seems to includes formats of representation that are not propositional. Everyday cognition makes constant use different sorts of —external and internal— representational structures that are not propositional, like images, maps, diagrams and formulas —just to name few. There is also an overwhelming amount of evidence coming from psychology and neuroscience about the central use of non-propositional representations in cognition (e.g., see Dove, 2009; Kosslyn, Thompson, & Ganis, 2006; Parsons, 2016; Pylyshyn, 2002; Shepard & Metzler, 1971), as well as various robust theories of visual and diagrammatic inference coming from logic and AI (Barwise & Etchemendy, 1996).

Fodor's take on this issue is that despite all this psychological evidence, mental imagery —or any other non-propositional mental representation— cannot be truly representational because only LOT sentences can be intrinsically meaningful, i.e., having proper intentional properties (Burnston, 2020; Fodor, 1985). Thus, if we want to accept in the ontology of our psychological theory alternative representational types, they must be "anchored" in the LOT. To put an example, Fodor claims that mental imagery would only make sense if each mental images-token comes *labeled* with a description "written" in a mentalese script. However, as various authors have noticed (Cummins, 1992; Horst, 1999b), the whole idea is deeply problematic. First, because there is no clear translational strategy that can explain how our use of mental images can be reduced to the LOT; second, the issues of how the LOT has intentional properties —and a proper syntax— are far from being settled. As a consequence, any translation strategy would inherit all the theoretical problems of the LOT hypothesis.

3.3.2 Representation and the organization of information

How representational structures organize information has an impact on the cognitive processes that exploit them. For instance, a map of Paris' metro system could be informationally equivalent to a set of sentences describing it. However, it is often much easier for people to extract relevant information from the map than from its linguistic description (see Lloyd, 1993). Non-discursive —external— representations are pervasive in all our cognitive practices (see Paivio, 2013). We use them in all kinds of contexts, and they seem to have a positive impact in several cognitive procedures, from learning to everyday reasoning (see Horowitz, 1967). This is hardly an intrinsic property of the representational structures themselves. Instead, as Larkin and Simon (1987) famously argued, it probably depends on the relation between the way they organize information and the information-processing mechanism acting on them.

Computationalists like Fodor seem to underestimate this point. In general, those supporting the LOT as the universal medium of representing information

have understood the issues of knowledge organization and computational efficiency as marginal in their explanations of high-level cognition. AI researches, on the other hand, have quickly seen that this problem was central for modeling processes like reasoning in artificial systems (Lakemeyer & Nebel, 1994; Woods, 1987). The propositional view, widely popular at the beginnings of AI thanks to the prevalence of logical models, lost much popularity for a simple reason: classical logic has several computational disadvantages when used as a framework for knowledge representation (Minsky, 1991; Sowa, 1999).[5] New representational frameworks were developed for organizing knowledge structures more efficiently. Frames (Minsky, 1974), semantic networks (Quillian, 1967), and conceptual graphs (Sowa, 1991) were all non-propositional alternatives to classical logic that proved themselves as useful ways for representing information in inference engines. Even if, in most cases, these methods do not have psychologically realistic foundations, and they do not explain where this semantic organization comes from (Brachman, 1977), their analysis seems relevant for the other disciplines studying reasoning.

The discussion about knowledge organization in AI is deeply related to the *frame problem* (see P. Hayes, 1988; Lormand, 1990). Roughly, the frame problem is how to retrieve relevant knowledge from a rich database in problem-solving contexts. For instance, let us say that I have to decide how to go from my house to the cinema. Most of the information in my background knowledge is completely irrelevant for answering that question; so there must be a mechanism of information retrieval that select from this diverse corpus those pieces that are useful for my task —for instance, from the category *means of transportation*, concepts like *train*, *bus*, *car*, or *bike*.

It seems evident that the conceptual organization of knowledge in semantic memory plays a crucial role in solving the frame problem. If we assume that beliefs are unorganized abstract entities floating in a *belief box* — as some classical epistemologists do—, then the retrieval mechanism must go through all the corpus of information to find the possible answers, something that will have

[5] See (P. Hayes, 1977; Lifschitz, Morgenstern, & Plaisted, 2008) for a discussion on this matter.

an extremely high computational cost. On the other hand, if we assume that knowledge is organized in different conceptual domains, the retrieval mechanism would only require to make a localized search in some closed body of information.

According to Fodor, the frame problem is the biggest challenge to the CTM (Fodor, 1988, 2001). This is a consequence of its reliance on the formality thesis. If reasoning is a purely syntactic mechanism operating over an unstructured set of sentence-like beliefs (LOT), [6] there is no way of explaining why when we reason —when we look for reasons for justifying a claim, for instance— we consider only information that is relevant for the problems' domain.

One way of avoiding the frame problem is to reject the LOT theory and assume that representation is organized in structured domains.[7] In philosophy, there have been few attempts to advance alternative theories of representation that are not language-based. The first wave of ideas in this direction came from Lewis *map theory of belief* (D. Lewis, 1994), which vaguely proposes that beliefs represent content through structures that are isomorphic to what is represented; consequently, knowledge is somehow organized mimicking the structure of the things represented (Hendricks, 2006; Shea, 2014).[8] Similar views have been proposed by Colin McGinn (1989) and Frank Jackson (1997).[9] More recently, a second wave of anti-LOT theories was developed in a very similar direction by Rescorla, Camps and Prinz (Camp, 2007; Prinz, 2004; Rescorla, 2018). The efforts were focused on discussing how map-like representations could meet certain properties that, in Fodor's analysis, are central to representational systems and are required for any account of mental causation —systematicity, productivity, logical structure, etc.

[6] In Fodor's view, the lack of structure is due to the fact that concepts —the *building blocks* of these mental sentences— are *atomic* and totally disconnected from each other. For reasons of space, I will not discuss this topic here. But it is deeply related to Fodor's support to purely referentialist semantics, and his rejection of any kind of meaning holism, such as that advocated by Brandom and Sellars.

[7] Another solution to the frame problem comes from massively modular approaches to high-level cognition (e.g., see Sperber, 2001)

[8] This idea has a long history in philosophy. One of the main defender of a structural view of representation and reasoning was Leibniz. See Swoyer (1991) for an explanation.

[9] See Braddon-Mitchell and Jackson (2006) for a review of this approach to belief.

Map-like theories of belief are interesting alternatives to the LOT because they offer a way of explaining the strong connectivity of our *web of beliefs* and, as Haugeland (1987) observed, this is an important step towards dealing with the frame problem . However, the available theories have several limitations. First, they do not offer any systematic model of the structure of these non-linguistic representations. Second, they do not show how their structural features affect crucial cognitive mechanisms like reasoning [10] and categorization. [11] Third, they are still conservative —and thus probably translational— regarding representational formats: they do not seem to accept different forms of representation of information in cognitive processes. As I will show later, representational pluralism is an interesting hypothesis for explaining reasoning and other important cognitive mechanisms.

3.3.3 Knowledge structures and the centrality of belief

Compared to classic computationalism, the views mentioned above try to give a richer idea of our knowledge system's connectivity. However, they also seem to buy the assumption that everything concerning mental causation starts and ends with the notion of *belief* —or *propositional attitude*. In this sense, they are aligned with the CTM since they seem to support some from of representational conservatism. Fodor —following Peirce— used to talk about reasoning as *belief fixation*; and this makes sense, since inferences are supposed to depart from beliefs and have as output other belief —or at least some doxastic attitude towards a proposition, as Harman (1986) observed. But, are *beliefs* enough for explaining inference and knowledge?

Suppose I am told that Fido is a dog, and from this, I immediately infer that he is also a mammal. The formalist will tell me that this piece of reasoning is an *enthymeme* —i.e., a syllogism with implicit premises— and that what there is in-between the premise and the conclusion is the implicit belief "All dogs are mammals" and some logical rules that exploit logical form. However, which

[10] It is fair to mention that the psychological theory that builds on similar ideas, Mental Models, does not give us a detailed description of what these structures are.

[11] Rescorla does show how his map-theory of belief could illuminate some problems of spatial navigation (Rescorla, 2018).

is the retrieval mechanism that brings this implicit belief to mind? And, how this retrieval mechanism choose from the *belief box* the adequate proposition if reasoning is a fully syntactic-based process —i.e., insensitive to content?

Computationalists like Fodor cannot answer this. As said earlier, this retrieval mechanism must be content-sensitive, and it operates over what is usually called *background knowledge*. It associates lexical concepts like "dog" with related concepts that may be relevant for an instance of reasoning —in this case, "mammal." As far as I can tell, there is no need to postulate that background knowledge is structured as a belief-box or as a belief-network because what is needed are not relations between propositional attitudes, but relations between lexical concepts.

Fodor saw propositional attitudes as so central to knowledge and cognition that he claimed that people have "innate beliefs about linguistic structure" (Fodor, 2001, p. 95). However, this is misleading. Most people are competent language users without having *beliefs* about the grammatical rules they use. In other words, having a linguistic ability does not require to have an attitude towards the truth of a proposition (see Jackendoff, 2002, p. 124-129).

Cognitive scientists took this direction when they started studying how background knowledge is structured in semantic memory. Semantic memory is a crucial notion in the explanation of cognition. It is supposed to be the *reservoir* of all the —experience-independent— information we have about concepts, properties and things in the world (Yee, Jones, & McRae, 2018). It has a central role in reasoning and other high-level cognitive mechanisms (Benedek et al., 2017; De Neys, Schaeken, & D'ydewalle, 2002). Several models of semantic memory have been developed over the years. Among the most influential, we find network models like those developed by Quillian's or Anderson's (2014; 1966); feature-based models like Smith et al. (1974) or McRae's (2004); and more recently, probabilistic models like the one developed by Kemp and Tenembaum (2008).

None of these models requires the use of a notion like *belief* —or any other propositional attitude— in its explanatory structure. Instead, they often assume a level of representation of semantic information that is not language-like,

and whose structural features are determinant in the interaction with other processing systems. This is a crucial assumption for this thesis. In particular, I will defend that there is a family of inferential mechanisms that exploit semantic information at a sub-symbolic level and that a proper model of the structural features of this information can shed light on the computational features of them.

The idea that there is a sub-symbolic level of representation of information is not new at all. For several decades, AI has been polarized between symbolic and sub-symbolic approaches (Eliasmith & Bechtel, 2006; Smolensky, 2012). The former propose that a language-like medium of representation of information, plus a logic-like system of rules, is enough to model high-level cognition. The latter tries to show that a connectionist model of representation can explain cognition without reference to sentence-like notions, like beliefs (cf. Von Eckardt, 2005).

I do not think that the two models need to be understood as rivals. I follow Gärdenfors (2000) in claiming that there can be various levels of representation of information that interact in different ways during cognitive processing. In particular, I think that reasoning requires the interaction of information that is explicitly represented in language —or other external format of representation— , with information that is implicit and codified in some sub-symbolic representational structure within semantic memory. As Gärdenfors has showed (1997), [12] assuming the existence of an intermediate level of representation of information that bridges the symbolic —linguistic— level with the connectionist level has many explanatory advantages. In particular, it can help explain and model several cognitive phenomena associated with conceptual representation manifested in linguistic behavior but not accountable from the symbolic. To name just three, typicality effects in categorization and reasoning (Rosch, 1983), non-monotonic reasoning (Osta-Vélez & Gärdenfors, n.d.), and some important features of the dynamics of language acquisition (Gärdenfors, 2014).

As mentioned before, in the upcoming chapters I will explicate some of these

[12] Also see Lieto (2017) for a development of Gärdenfors' argument.

ideas by using conceptual spaces to model various kinds of concept-based inference. The thesis I defend is that a big deal of the structure of background knowledge can be explained with conceptual spaces, and that some of these structural features can shed light on semantic-based inferential mechanisms. I will show how different word-classes with specific semantic structures allow for different patterns of inferences. Notice that is similar to Jackendoff's idea discussed in the previous chapter. The difference is that conceptual spaces offer a much more powerful formalism to account for these inferences, and this translates in both a better mathematical model and a more promising explanatory framework.

Before introducing conceptual spaces in the next chapter, I will discuss some more general ideas about how to understand inference and its relation to representation.

3.4 From representational pluralism to inferential pluralism

This chapter opened by claiming that *inference* was a widely used yet underdefined notion. Besides the recent efforts of some philosophers (e.g., Boghossian, 2014; Hlobil, 2019; Valaris, 2017), I think it is still not entirely clear which analytical categories are best suited for theorizing about inference. In what follows, I will propose a broad definition that tries to capture some common ideas of what inference is and that is capable of encompassing the many different mechanisms that are described in the literature under this cover term:

Definition:

An inference is a transition from an informational (mental) state I_1 to a new informational (mental) state I_2 that satisfies the following points:

(i) It is informationaly fruitful;

(ii) It follows a systematic criterion which exploits properties of the representational structures that underlay I_1 and I_2.

Regarding (i), it is not among the aims (or possibilities) of this work to develop a systematic criterion of informational fruitfulness. However, I think that Johnson-Laird and Byrne's three constraints on inference could be a good starting point (Johnson-Laird et al., 1992, Chapter 3). The first one states that inferences should not "throw away" semantic information. For instance, people do not infer from a premise p, a disjunctive conclusion $p \vee q$ because, even if logically valid, the conclusion is less informative —it is compatible with more possibilities— than the premise. Second, rational inference is *parsimonious*: We do not conclude, from a set of premises a conjunction of each of these premises. Again, this is logically valid yet uninformative. Finally, inferences should add something new to what it is already stated in the premises.

As mentioned earlier, the lack of criteria of informational fruitfulness is an important problem of logical models of reasoning. Classical logic, taken as a model of deductive competence, allows any inferential movement regardless of the informative relationship between premises and conclusions —e.g., "Tigers are reptiles. Thus it is raining or it is not raining" is a legitimate inference. The principle of explosion (*ex falso quodlibet*) is an extreme symptom of this issue.

Furthermore, (i) would require to introduce principles accounting for the role of the context of production in which inference is made. In the sense of Barwise's situation logic (Barwise, 1989) or, if the context is dialogical, some *Gricean* constraints like those proposed by relevance theory (Sperber & Wilson, 1986). Finally, the new Bayesian paradigm of rationality can also contribute to the specification of (i). As Eva and Hartmann have recently shown, argument schemes that were traditionally considered as fallacious (and thus informationally useless) like *affirming the consequent* and *denying the antecedent* , when analyzed from a probabilistic perspective can be seen as informationally fruitful since they provide considerable support for a conclusion (Eva & Hartmann, 2018).

(ii) captures what is mainstream in contemporary cognitive science: cognitive processes operate on specific systems of representation of information. In particular, the ideas I defend follow Mercier and Sperber's approach to inference (Mercier & Sperber, 2017), which makes a crucial use of the notion of *representation*. According to them, inferences are mechanisms that exploit empirical regularities of the environment for understanding, prediction and —fundamentally— action. As they write: "*No regularities, no inference. No inference, no action.*" (ibid, p. 85). These regularities are encoded in systems for representing information that can be private or public. Natural language is the most important public structure for representing regularities, but also diagrams, scientific notations, and mental and external imagery do this job. [13] Since these systems of representation have different structural properties, it follows that the inferential mechanisms they allow are computationally different. This last point is the focus on this work, and it is behind of what I call *inferential pluralism*: A view that assumes the existence of a plethora of inferential mechanisms that are computationally different because they are based on different systems of representing information. Inferential pluralism requires the rejection of the formality thesis, and of any logicist and translational approach trying to describe any inferential mechanism with the same set of inference rules.

3.4.1 The varieties of inference

Perceptual Inference

The definition of inference presented above builds on a general notion of representation. I do not take representation to be only a personal-level phenomenon. Any information-processing cognitive mechanism can be analyzed as building on a way of representing information. An interesting case of this, is visual perception. In particular, object recognition. Hoffman and Richards (Hoffman & Richards, 1984) showed that object recognition exploits geometrical properties in the retinal representation of objects that mirror regularities

[13] This idea has a long history in philosophy. Leibniz conceived reasoning as a *surrogative* mechanism, that is, a process that manipulates information about a target phenomenon through its symbolic representation in the human mind (see Swoyer, 1995).

in nature. More specifically, the visual system parses object contours at *extrema* of concave curvature —this is called "minima rule." This mechanism works by exploiting a topological regularity in objects —"transversality": distinct parts of objects intersect in a contour of concave discontinuity of their tangent planes (Figure 3.1). At any point around this intersection, a tangent to one of the surface's part forms a concave vertex with the tangent of the surface of the other part. In other words, the transversality rule implies that the different parts of images of complex shapes are segmented by recognizing the "concavities" of their figures marking the divisions between the contours of their constituent parts. This mechanism is crucial for object recognition but also for 3-dimensional vision and optical illusions (Hoffman, 2005). [14]

This way of analyzing perceptual inferences is consistent with the definition presented above and with Mercier and Sperber's idea about the relation between inferences and empirical regularities. For matters of space, I will not attempt an explanation of the informational criterion on visual perception. However, it is generally accepted that perceptual inferences follow informationally efficient heuristics (Hoffman, Singh, & Prakash, 2015).

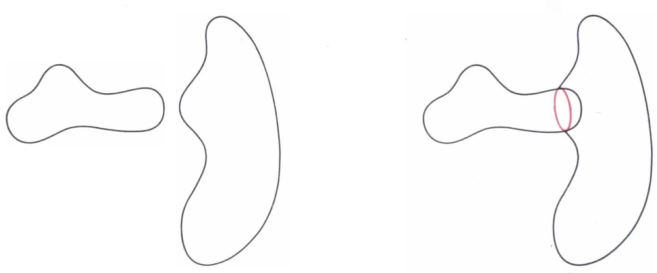


FIGURE 3.1: The rule of transversality states that when two surfaces penetrate each other at any random point, they always meet at a concave discontinuity —in red—. From (Hoffman, 1983, p. 157)

[14] It is interesting to compare visual inferences with other forms of perceptual inferences that exploit environmental information in different ways but accomplishing similar results. A particularly illustrative example is fish's electroreception. There is robust evidence showing that some species of fish measure distances, and recognize shapes and other properties of objects not by visual cues, but by interpreting information about the environment coming from self-produced electric signals with epidermal electroreceptors (von der Emde, 2004).

Concept-based inference

Inferences that exploit properties of conceptual structure are the main focus of this work. As said earlier, the idea is that information about lexical concepts is encoded in a sub-symbolic representational system that supports several cognitive mechanisms among which is reasoning. For instance, inferences like "$Tiger(x) \rightarrow Mammal(x)$" use core —definitional— knowledge of the concept *tiger* (we will discuss this in Chapter 5). A nonmonotonic inference like "$Robin(x) \rightarrow Fly(x)$" builds on the prototypical structure of the concept *bird*. As we will see in the following chapters, most categories have prototypes that "concentrate" those features which we consider typical (*regular*) of objects falling under these categories (Rosch, 1983). Chapter 6 explicates how prototypicality is exploited in nonmonotonic reasoning with conceptual spaces; while Chapter 7 will extends this idea to category-based induction. In general, this model will exemplify how semantic structure is exploited in inference instead of logical —syntactic— structure.

Diagrammatic and model-based inference

Let's briefly talk about inferences based on external —artificial— systems of representations. I will use an example given by Mercier and Sperber to elaborate this point (Mercier & Sperber, 2017, pp. 101-102). Numeral systems are symbolic structures that represent mathematical information in different ways. For instance, the number nine and three are represented as "9" and "3" in the decimal system, and as "3" and "10" in a base-nine system. For those familiarized with decimal numerals, it seems evident that the multiplication "300 x 3" equals "900", since 9 is three times 3 and we know that adding zeros to both numbers preserves the multiplicative proportion. But doing the same thing with a base-nine system is much harder since this latter system does not display these proportions with rounds numbers so easily (the above operation is represented as "3x10=1210"). Thus, reasoning about the same problem is different according to the system used, since they have other representational properties (see also, Buijsman, 2018).

Similarly, diagrams play a central role in inference. In particular, mathematical reasoning, both in teaching and professional contexts, makes a substantial use of figures and diagrams, so that they are hardly replaceable by equivalent linguistic descriptions. In recent years, this issue has gained relevance in psychology and philosophy. Research indicates that diagrammatic systems have their own inferential properties, which with significant epistemic advantages compared to natural and logic-like languages (see Barwise & Etchemendy, 1996; Moktefi & Shin, 2013).

In general, scientists use hybrid representational structures for reasoning about phenomena. Philosophers of science have been discussing this for decades now. In particular, they are interested in how scientific models make possible forms of reasoning that were not feasible with only natural language (see Nersessian, 1999; Vorms, 2011). Chapter 8 analyzes this particular problem by looking into the development of the notion of instantaneous speed from geometrical physics to analytical mechanics. This will exemplify how conceptual information can also be distributed across external representational structures, for later being used in inference.

Logical inference

The view defending here is incompatible with the formalist thesis and with any translational approach like the one proposed by CTM. However, this does not mean that it cannot account for formal inferences like the ones described by classical logic. Experimental evidence from (Falmagne, 1990) suggests that adults untutored in logic, when told the sentences with blank —unknown— predicates "If Sarah fibbles, then she thabbles" and "Sarah fibbles," consistently deduced that Sarah thabbles. Falmagne then claimed that, even if it is clear that reasoning is content-related, we are able to deal with abstract schemes like the ones proposed by logic (see also Falmagne & Gonsalves, 1995).

Formal properties like the ones described by classical logic are indeed properties of natural language —which is a representational system. It is possible that when people reason with blank predicates, they focus on other cues like the syntax and properties of logical form. In these cases, just like in the cases

described above, subjects are exploiting properties of a representational system to make inferences.

3.5 Conclusions

The few examples discussed above intend to illustrate how reasoning is anchored in representation and how the format of the representational system in use determine the "form" of the inferential mechanism that exploit it. From the existence of multiple representational structures, it follows that inference must be studied from a pluralist —not translational— perspective.

This chapter closes the theoretical discussion on inference and meaning within the formalist tradition dominant in both psychology and philosophical semantics. The following chapters of this dissertation try to lay the foundations of a formal model attempting to explicate three types of semantic-based reasoning: material inferences, non-monotonous inferences, and category-based induction. The model builds on the theory of conceptual spaces (Gärdenfors, 2000, 2014). In particular, this framework is used to show that a rich and systematic approach to the structure of concepts can shed light into the nature of the relationship between concepts, representation and inference.

Chapter 4

Introducing Conceptual Spaces

Summary

This chapter contains a standard introduction to Conceptual Spaces, a crucial framework for the analysis of semantic-based inferences in the following chapters. Its basic theoretical and formal aspects will be explained, as well as some of its possible applications.

As stated in the introduction, one of the fundamental ideas behind this thesis is that reasoning requires understanding, and understanding requires concept-possession. Therefore, a theory of how we reason cannot be detached from a theory of how we represent concepts. In previous chapters, we discussed the disadvantages of assuming otherwise. The endorsement of the formality thesis faces us to the frame problem and blocks any possible explanation of content-effects in reasoning —among other issues. It was argued that a way of avoiding these issues was to assume that reasoning builds on features of representational structures that organize conceptual knowledge in the mind/brain.

If we focus on language-based inference, the main representational structure is our conceptual system, which encodes all sort of information associated with lexical concepts. (e.g, see Jackendoff, 1992) In this sense, a theory of semantic-based inference needs to build on a theory of the structure of concepts. I believe that Conceptual Spaces (CS) is the best-suited theory for this task. It provides a rich explanatory framework for analyzing the structure of concepts and their interrelations and offers powerful formal tools to model several concept-related phenomena. In what follows, I will explain the basics of this theory for latter use in modeling different forms of semantic-based reasoning.

4.1 Defining conceptual spaces

Conceptual Spaces (Gärdenfors, 2000, 2014) is a research program in cognitive science for modeling several cognitive phenomena involving concepts and conceptual structures —e.g., semantic processing, learning, reasoning, categorization, concept formation, etc. Unlike the dominant computational tradition in philosophy and cognitive science, CS does not assume that language —or some language-like structure like the LOT— is the unique representational system supporting high-level cognition. Instead, as explained in the previous chapter, CS builds upon the fundamental hypothesis that there exists an intermediate representational system that encodes semantic information with spatial structure.

This theory is an heir to the geometrical models of conceptual representations inaugurated by Shepard (1987) in psychology, and the development of the notions of "quality spaces" in Quine (1969a; 2013), “attributes spaces” in Carnap (1971), and “logical spaces” in Stalnaker (1981). Just like in the other geometrical models in psychology, the fundamental idea behind CS is that concept formation and representation takes place in some psychological space in which similarity can be represented in terms of distances determined from some metric. [1]

CS builds on two paramount notions: *quality dimensions* and *domains*. Quality dimensions are the "building blocks" of concepts. They represent different *qualities* of objects that are used as a basis for judging the similarities among different stimuli (Gärdenfors, 2000, p. 6). For example, *pitch* is a quality dimension of auditory stimuli; by focusing on pitch we can compare and classify different sounds. Quality dimensions are diverse, they can be innate, culturally acquired, phenomenal, or abstract depending on the concept.

A central point is that quality dimensions can be represented by different geometrical structures (see Gärdenfors, 2000, Chapter 1). For instance, weight and pitch can be both represented by a line isomorphic to the non-negative

[1] For an analysis of the notion of *psychological space*, see (Eliot, 1987).

real numbers (Figure 4.1). Other dimensions have a discrete structure and correspond to qualities that are represented as disjoint sets.

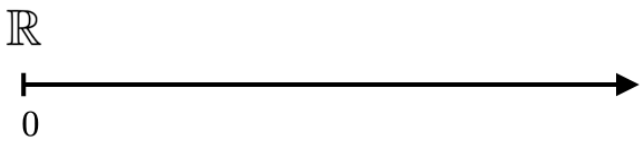


FIGURE 4.1: Geometric representation of the weight and pitch dimensions

Dimensions can be *integral* or *separable*. They are integral when it is impossible to assign to an object a value in one dimension without assigning another value in another dimension (see Maddox, 1992). For instance, we cannot represent a tone with a specific pitch but without a value for its loudness. In contrast, some dimensions can be represented independently from each other, like *height* and *wealth* when thinking about people. In these cases we talk about *separable* dimensions. Integral dimensions are often modeled with an Euclidian metric, while separable dimensions with a City Block Metric.[2]

A set of integral dimensions that are separable from all other dimensions is called a *domain*. The classic example of a domain is the color spindle. It is composed of three integral dimensions: *hue*, *saturation*, and *brightness*. The geometrical representation of hue is the color circle. Saturation or intensity is represented as an interval of the real line, while brightness varies from white to black and is thus a linear dimension with endpoints. Together, these three integral dimensions, one with a circular structure and two with a linear structure, make up the color space (see Figure 4.2).[3]

Domains serve to represent different qualities of objects through their geometrical and topological properties. A central notion in this sense is *distance*, which serves as a measure of similarity among properties in the domain: The closer they are in space, the more similar they are.[4] For instance, within the

[2]Johannesson (2002) showed that, in some cases, a Minkowski metric can be useful too.

[3]It is worth mentioning that the figures in this chapter have only one illustrative purpose. They do not come from real data about the conceptual spaces they are supposed to represent.

[4]Not all spaces have a metric. For example, some dimensions only have an ordering structure.

color space, predicates like *red*, *blue* or *orange* correspond to regions of the domain. The relationships among them can be analyzed as a function of their relative positions in the color domain. For instance, the distances in the color domain allow us to see why *orange* and *red* are more similar than *red* and *green*.

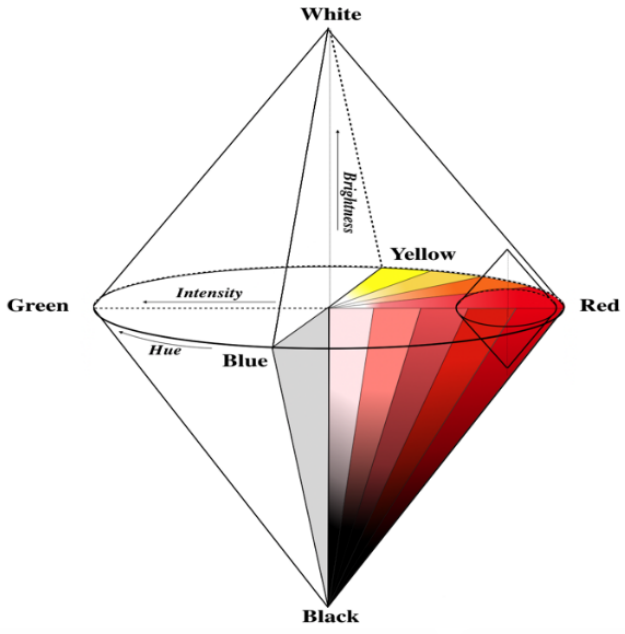


FIGURE 4.2: Color spindle. *Red* is a sub-region of the color domain.

The domains of a conceptual space are related in several ways since the properties of the objects modeled in the spaces will co-vary. For example, in the "fruit space", the dimensions of *ripeness* and *color* will co-vary, as well as *size* and *weight*. These co-variations are support different inferential procedures that exploit conceptual properties, as we will see in Chapter 6.

A *conceptual space* is defined as a collection of one or more domains with a distance function —a *metric*— which represents properties, concepts, and their —similarity— relationships. Similarity among concepts and can then be easily estimated since it is a monotonically decreasing function of their distance within the space (Shepard, 1987).

Within this framework, concepts are understood as a region of some conceptual space. Gärdenfors makes an important distinction between *properties* and *concepts*: when the space correspond to a single domain, we talk about

properties instead of concepts. He then claims that *natural* properties are characterized by the following criterio (2000, Chapter 3):

Criterion P: A natural property is a convex region in some domain.

Convexity exploits the geometric features of quality dimensions. A region is convex when for every pair of points x and y in it, all points between them are also in the region. Convexity is a crucial feature for categorization and concept comparison.

Gärdenfors (2000) claims that color terms, being natural properties, are constrained by the structure of the conceptual space in which they are grounded across different languages. In other words, it should not be possible for any language to have one single word for two colors like *red* and *green*, since they are represented as disjoint sub-regions of the color conceptual space —which is perceptually grounded. This conjecture has been confirmed for an important number of different languages by Jäger (2007).

Within this framework, concepts are represented as regions of some set of interconnected domains. Gärdenfors (2000, p. 105) defines concepts according to the following criterion:

Criterion C: A concept is represented as a set of convex regions in a number of domains, together with information about how regions from different domains are correlated.

Fruit categories are often used as examples of (natural) concepts. For instance, consider a toy model of an *apple-space* that is a subset of the Cartesian product of the color, taste, shape, ripeness, and texture domains. This apple-space would extend itself thorough certain regions of each of these domains —those representing the common properties of apples—, while leaving other regions "untouched" —for instance, we do not represent apples with pyramidal shape, or being black, so these properties are not covered in the conceptual space. The concept *apple* has several correlations between its properties: the

degree of sweetness and sourness, as well as the texture, are correlated to the ripeness level. These correlations can be also represented in the conceptual space though different mathematical tools (see Figure 4.3).

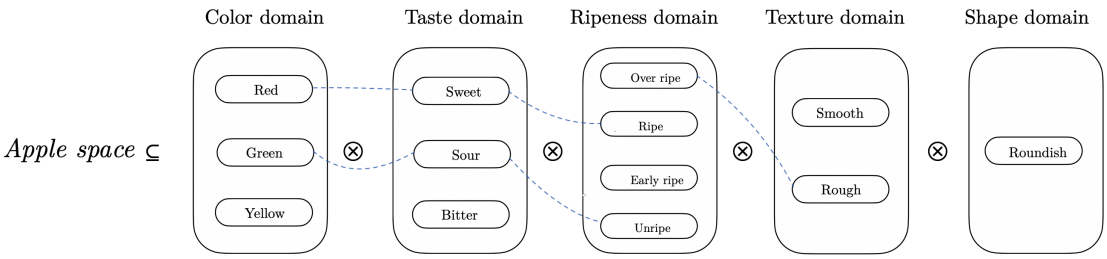


FIGURE 4.3: Illustrative diagram of an apple-space. The correlations among properties are represented by dotted lines.

Object representation in conceptual spaces

Object representation is a central point for this work. As just said, CS understands concepts as convex regions within the space, i.e., convex sets of points. Individual objects are then seen as instances of concepts and are mapped into points of the space. In formal terms, the conceptual space of concept M — written $\mathscr{C}(M)$— can be seen as a subset of the Cartesian product of n domains:

$$\mathscr{C}(M) \subseteq D_1 \times D_2 \times \cdots \times D_n$$

An object x falling under M is represented as a n-dimensional point $x =< x_1, x_2, \ldots, x_n >\in \mathscr{C}(M)$. Each x_i in x represents the coordinates of the point in the domain D_i , which will typically fall under some sub-region $R_i \subseteq D_1$that represent a sub-ordinate category of D_i.

Just as in the case of concepts, the similarity among two objects in the space can be estimated by using the built-in distance function. Similarity among objects is generally easier to compute than similarity among concepts, since the former requires to measure the distance among two points in the space and the latter the distance between sets of points in the space. We will come back to this point in Chapters 6 and 7.

An important assumption underlying this work is that for each of the domains that constitute a concept, there is a corresponding distance measure. Furthermore, it is assumed that these domain-specific measures can be weighted together to create an overall distance measure for the space. As we will see in Chapter 7, this weighting is, in general, context dependent.

4.2 Prototypicality

Most everyday concepts have prototypes. That is, some members of the concept that are considered as more *central* or more representative than the others. One important advantage of the conceptual spaces framework is that it can represent prototypes of concepts. In that sense, it fits very well with the prototype theory of categorization (Gärdenfors, 2000; Lakkof, 2017; Mervis & Rosch, 1981; Rosch, 1975). In particular, Criteria P and C allow for a natural way of representing the prototype effects (see Chapter 7). Within convex regions, one can take some specific point —or set of points— as the prototype of a category.[5] As a result, and using the built-in metric of the space, one can measure the degree of typicality of any member of a category by estimating its distance to the prototype. For example, focal colors are often considered in cognitive science and linguistics as prototypes of the color space (Douven, 2019; Rosch, 1971).

Assuming the prototypical structure of concepts does not require an actual object representing the prototype. Conceptual spaces can represent every possible object falling under a concept. Gärdenfors (2000) claims that a prototype can correspond to a partial vector containing only information about the values of the most relevant dimensions for the concept.

Gärdenfors (2000, pp. 87-89) showed that it is possible to argue in the converse direction. If we assume that concepts have a prototypical structure, then it is expected that they are represented as convex regions. To obtain a prototypically structured conceptual space, we start from a set of prototypes $p_1, ..., p_n$ of the categories to be represented —for example, different bird species— that

[5]It should be noted that this does not necessarily lead to being central in the regions they are assigned.

will be the central points in the categories they represent. If we assume then that for every point p in the space, its distance from each p_i can be measured, and stipulate that p belongs to the same category as the closest prototype p_i, then a partitioning of the space into convex regions can be generated. This is called a Voronoi tessellation —a two-dimensional example of this is illustrated in Figure 4.4. Thus, assuming that a metric is defined on the subspace that is subject to categorization, a set of prototypes will generate a unique partitioning of the subspace into convex regions.

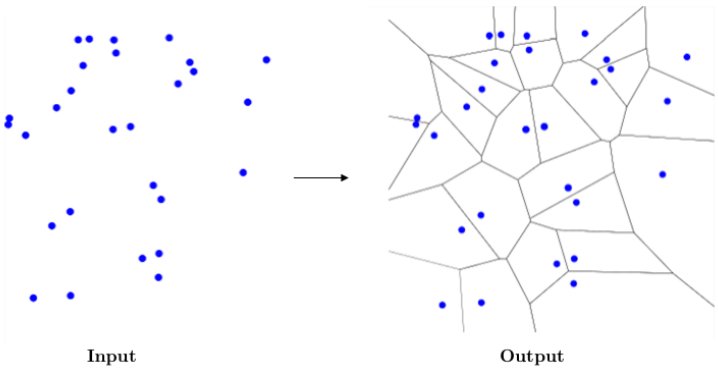


FIGURE 4.4: Voronoi partitioning of a space from a set of points.

Because of the role that prototypes have within this theoretical framework, typicality is an independent variable. As Gärdenfors shows, this particular spatial configuration of the space has several advantages in terms of the economy of cognitive processing (Gärdenfors, 2000, pp. 123–126).

4.3 Context, domain salience, and dynamic conceptual spaces

An important phenomenon that any theory of concepts must account for is that psychological similarity is a variable measure that is dependent on the context (Goodman, 1972). In particular, as noticed by Robert Nosofsky (1986), conceptual similarity is modulated by attention to specific domains of the compared concepts. For instance, apples are often seen as more similar to tomatoes than to dates. However, in the context of choosing dessert —in which "sweetness" is a salient feature— it is expected for this similarity judgment to change.

The contexts in which concepts are used are crucial in the modulation of the similarity measure. Context-effects have been extensively studied in the psychological literature (Goldstone, Medin, & Halberstadt, 1997; Keßler, Raubal, & Janowicz, 2007), and geometrical models of similarity have been often criticized because of their limitations at the moment of accounting for them (see Decock & Douven, 2011; Tversky, 1977, for a review).

The conceptual spaces model, however, does not suffer these problems (cf. Johannesson, 2002). In this theory, the context-sensitive character of psychological similarity is accounted for in terms of a weighted distance measure. For instance, within the context of a Euclidean metric, the distance measure will include attention-weights w_i that modify the salience of dimension i in the conceptual space:

$$d(x, y) = \sqrt{\sum_i w_i (x_i - y_i)^2}$$

When a larger value is given to a weight w_i, the conceptual space is "magnified" along that dimension. That means that this dimension i will become more important when determining the similarity between categories (Gärdenfors, 2000, p. 20). As we will see in Chapter 7, this weighted-distance function will have a central role for accounting for the role of context in category-based induction.

In summary, thanks to their geometrical and topological structures, conceptual spaces allows us to represent concepts and their interrelations in a well defined similarity space. It gives us then a powerful formal and explanatory framework for analyzing concept-related cognitive phenomenon. Concept formation and concept composition, categorization (Gärdenfors, 2000), semantic vagueness (Douven, Decock, Dietz, & Égré, 2013), and word learning (Gärdenfors, 2014) have been some of the cognitive or conceptual phenomena already modeled with this theory. In what follows, we will focus on their role in inference and reasoning.

4.4 Inference and conceptual spaces

As we mentioned earlier, the formalist thesis claims that reasoning is based on propositions and can be described by some set of topic-neutral and domain-general rules. As a consequence, conceptual structures are immediately dismissed as inferentially irrelevant. However, as explained in Chapter 1, that view has been criticized because of its psychological implausibility. Nowadays, mainstream cognitive science assumes —in one way or another— that concepts play a constitutive role in inferential processes. (see e.g, Carey, 1985, 2000; Evans, 1989).

CS can offer interesting insights into the role of conceptual knowledge in reasoning. In this framework, inference is not conceived as a process that takes place —exclusively— at the propositional level, but one that supposes the interaction between the conceptual and the symbolic levels (Gärdenfors, 1997). In particular, and considering what was explained in the previous chapter, we will use CS for showing how different forms of semantic-based inference can be explicated as mechanisms exploiting properties of a representational system encoding conceptual information.[6] In this case, since conceptual structure are geometrical and set-theoretical structures, inference can be understood as exploiting different geometrical and set-theoretical properties.

As a simple example, consider the inference "the car is red; thus, the car is not green". This inference is intuitively valid —yet logically invalid— for any subject who grasps the basic color concepts. In conceptual spaces, understanding the notion *red* involves being able to represent an object in the *red* region of the color spindle, something which immediately implies that the object is not located in the other regions of the spindle —*green*, *yellow*, etc.

Furthermore, Gärdenfors (2000; 2008) showed some ways in which CS can model features of nonmonotonic and metaphorical inference. Recently, also Schockaert and Prade (2013) used CS to model interpolative reasoning. In the following three chapters, CS will be used for modeling three different inferential

[6]Notice that, unlike what happens with the belief-centered tradition discusses in the previous chapter, CS does not require us to assume that all thinking consists of relations between propositional attitudes.

mechanisms: material inferences —also known as *semantic entailment*—, non-monotonic reasoning, and category-based induction.

Chapter 5

Explicating material inferences *via* Conceptual Spaces

Summary

This chapter discusses Wilfrid Sellars' notion of *material inference*, a semantic-based inferential mechanism playing an important role within inferentialist theories of meaning. I critically discuss this idea within the Sellars-Brandom tradition and point out some relations to current views in the psychology of reasoning. The main claim is that Sellars' approach lacks the analytic tools for developing a fine-grained —and psychologically grounded— account of material inferring. I propose an alternative framework based on cognitive semantics and conceptual spaces that can do this job.

5.1 Beyond logical forms

In Chapter 1, we saw how the classical notion of validity construes predicates —i.e., concepts and properties— as *inferentially inert* lexical items. Roughly, an inference scheme is valid in virtue of its formal structure, which depends exclusively on the truth-functional properties of logical constants and not in the content of the predicates involved.[1]

[1]As explained in Chapter 1, classical validity is formal because classical logic is truth-functional. That is, it only accepts as logical constants —terms with inferential properties— truth-functional lexical items (see S. Read, 1988, Ch.2, for a thorough examination of this issue).

This view has —at least— one important counter-intuitive upshot; it classifies as valid, inferences like (a),

(a) *If my dog barks at sunset, then Chet Baker plays the trumpet or Chet Baker does not play the trumpet.*

And as invalid, inferences like (b),

(b) *Munich is south of Berlin. Thus, Berlin is north of Munich.*

The issue is that (a) is intuitively absurd and (b) intuitively obvious for any competent language user; however, (a) is approved by logic, while (b) disapproved.

Argument schemes like the one instantiated by (a), are a result of the truth-functional interpretation of the "if...then..." connector in classical logic, and take part on what is called the "paradoxes of material conditional" (cf., Brandom, 1981). There is a long tradition in philosophical logic that criticizes this point, arguing that classical implication does not reflect the right features of our use of conditionals in natural language (Edgington, 2020; R. Stalnaker, 1981). In particular, any formal interpretation of conditionals that abstract away the thematic relation between premise and conclusion seems to be condemned to failure if it aims to capture the intuitions behind our everyday use of them.

Maybe the most prominent attempt to fix this problem in logic was the development of *Relevant logic* (A. Anderson et al., 2017; S. Read, 1988), a formal system that avoids the alleged paradoxes by introducing a content-based constraint on the interpretation of conditional expressions —roughly, it demands that antecedent and consequent of a "good" conditional are relevantly (topically) related. The central idea behind it is that truth-functionality is not enough to define a good conditional. We need to account for the intuitive fact that any claim of the form "P implies Q" suggests that the truth of P gives us a reason to believe in the truth of Q. Since the focus of this work is not logic, but inference from a cognitive perspective, we are not going to dig into these ideas further. However, I share the spirit behind relevant logic: that an

accurate characterization of "inference" requires to go beyond truth-functional structures and look into content-based relations among predicates.[2]

Maybe the most important attempt to build a theory of inference in the aforementioned sense was made by Wilfrid Sellars and Robert Brandom, through the development of the notion of *material inference* and *material validity*. Succinctly, an inferential move is *material* when it is based on a conceptual relation between the predicates in the premise and the conclusion. Material validity has nothing to do with logical form. It is not truth-functional, but it is related to how concepts are articulated within a normatively structured inferential practice. In what follows, we will take a closer look at Sellars' ideas about this notion.

5.2 Sellars on material inferences

Sellars was one of the first philosophers in the analytic tradition to tackle conceptual content through an analysis of inference. [3] His concerns were not about logical or meta-mathematical matters, but about the relations between language and thought. In his influential paper, "Inference and Meaning" (1953), he uses the terms "material inference" and "material validity" to refer to inferences like (b), which are not valid in virtue of their logical forms, but are "materially valid," that is, valid in *virtue of their meaning*. Roughly, Sellars believes that our inferential practice is mostly "material," because of the role that the rules behind these inferences —"material rules"— play in the construction and use of concepts in language and thought.

[2]Notice that this is also what guides Katz' program —as it was explained in Chapter 2.The difference is, however, that the approach presented here is entirely cognitivist, while Katz believes that meanings are "autonomous" entities (cf. Jackendoff, 1981).

[3]It may be worth mentioning that Sellars' discussion also revolves around the Kant-Carnap-Quine riddle of analiticity (seee, Westphal, 2015). In particular, Carnap anticipated the idea of meaning-constitutive inferences when claiming that the meaning of extra-logical terms is fixed by the set of *deduciblility* rations between expressions containing these terms and other expressions. For instance, the meaning of *arthropod* is determined by inferences from "$Athropod(x)$" to "$SegmentedBody(x)$," "$JointedLegs(x)$," and "$Animal(x)$" (Carnap, 1959, pp. 62-63).

Before explaining his views, it may be worth commenting on the framework within which they were developed. Sellars' point of departure is a critical revisions of Carnap's ideas on the syntactic structure of language as characterized by a set of *formation* and *transformation* rules.[4] According to Carnap, formation rules determine how symbols of the language can be combined to form complex sentences; while transformation rules specify conditions under which other sentences can replace sentences of the language (Carnap, 2000). For instance, the introduction of the conjunction ($P, Q \vdash P \wedge Q$) is a formation rule, while *modus ponens* ($P \rightarrow Q, P \vdash Q$) is a transformation rule.

Carnap thought that these rules exhausted the syntactic dimension of language. Sellars, however, pushed the rule-based view further and conceived the entire linguistic system as rule-governed (Sellars, 1950). In particular, he developed a *functional* view of meaning —as said in Chapter 2, he was one of the founders of IRS— in which the meaning of expressions is determined by their "role" in a language game. Thus, understanding the rules of the game will give us an insight into the conditions of meaningfulness of type-expressions in the language —i.e., into the semantic structure of the language.

Sellars identifies various kinds of rules governing language. There are *language-entry rules*, which specify how to verbally react to an environmental —non-linguistic— stimulus. For example, when I see red, I think or say "red". And there are *language-exit rules*, which concern how my actions are consistent with what I express about my intentions —e.g., if I say "I will open the window," then I open the window. For Sellars, language-entry transitions make an important contribution to the content of lexical —extra-logical— concepts. But they are far from been enough.

The rules that are crucial for this job are *language-language rules*. These are intra-linguistic transitions that determine our verbal behavior before linguistic inputs —and govern reasoning and understanding, according to Sellars (1974, pp.423-424). Carnap's formation and transformation rules are examples of this kind of language-language rules. But Sellars claims that Carnap's cannot say much about the meaning of lexical concepts in the language. He then proposes

[4]For a discussion on Carnap's influence on Sellars, see (Carus, 2004).

a different kind of transformation rules called *material*, that are supposed to do this explanatory job.

Material rules are transformation rules that govern our non-logical inferential practice, that is, all the inferences we make by exploiting our knowledge of conceptual relations and not our knowledge of logical constants. As said before, (b) is an example of a material inference that follows a material rule associated to the concepts *north-south*. But also inferences like "$Dog(x) \rightarrow Mammal(x)$," or the Wittgenstenian color-exclusion inferences like "$AllGreen(x) \rightarrow \neg AllRed(x)$" follow material rules. The difference between formal and material rules is stated in the following passage:

> The rules on which I wish to focus attention are rules of inference. Of these there are two kinds, logical and extra-logical (or 'material'). I can best indicate the difference between them by saying that a logical rule of inference is one which authorizes a logically valid argument, that is to say, an argument in which the set of descriptive terms involved occurs vacuously (to use Quine's happy phrase), in other words, can be replaced by any other set of descriptive terms of appropriate type, to obtain another valid argument. On the other hand, descriptive terms occur essentially in valid arguments authorized by extra-logical rules. Let me now put my thesis by saying that the conceptual meaning of a descriptive term is constituted by what can be inferred from it in accordance with the logical and extra-logical rules of inference of the language (conceptual frame) to which it belongs. (A technically more adequate formulation would put this in terms of the inferences that can be drawn from sentences in which the term appears). (Sellars, 1953, p. 136)

According to Sellars, without material rules accounting for how predicates are related, no logical or philosophical analysis of language would be accurate. [5] The reason for this is that the empirical content of natural language is codified in

[5] This is Sellar's main point of disagreement with Carnap. Carnap thought that P-rules i.e. material rules, were dispensable for the construction of the language (Carnap, 2000, §§51-52). Sellars', on the other hand, saw these rules as essential (Sellars, 1950, p. 268).

its predicates and their interrelations. It is impossible to explain how language can express, for instance, causal regularities about the world without accounting for rules regulating our use of empirical concepts. In this view, possessing a concept is knowing what inferences are afforded by it; which is basically to know how this concept is related to other concepts (Sellars, 1948). For instance, there is no way of possessing the (everyday) concept *lightening* without also knowing other concepts like *thunder*, *sky*, or *cloud*, and in particular, without knowing how they are inferentially articulated. As Brandom says:

> On an inferentialist account of conceptual content, one cannot have any concept unless one has many concepts. For the content of each concept is articulated by its inferential relations to other concepts. Concepts, then, must come in packages (though it does not yet follow that they must come in just one great big one). (Brandom, 2000, pp. 15-16)

Sellars (1958) discusses the relation between formal inferences and material inferences. The central question here is if material inferences are in fact *enthymemes*, and as such, inherit their validity from formal rules. For instance, in the *enthymematic* view, the inference "$AllGreen(a) \rightarrow \neg AllRed(a)$" would be a formal inference which has "$\forall(x)(AllGreen(x) \rightarrow \neg AllRed(x)$" as hidden premise; and whose validity depends on a formal transformation rule. If this is the case for all material inferences, they will not play any substantial role in describing the rules governing language.

However, Sellars' analysis concludes that material rules cannot be accounted for in terms of formal transformation rules. His main argument for this is that there is no way of making sense of subjunctive conditionals without ultimately relying on material rules (see Sellars 1958, or Brandom 1998b, pp. 102-104). As a consequence, material inferences cannot be enthymemes and their validity do not rely on formal rules but in material rules. I will come back to this point later.

5.2.1 Inference, laws, and regularities

The notion of inference plays a crucial role in Sellars' entire philosophical system, and it is closely related to the idea of *law*. Sellars can be seen as taking part in a tradition that understands laws *inferentially*, together with Gilbert Ryle, Stephen Toulmin, and Moritz Schlick (see Lange, 2000, p. 188-191). This tradition defends the idea that what better defines lawlike statements is not their logical structure, but a pragmatic feature: they *entitle* us to make inferences about causal relations. As deVries explains:

> Causal statements and other lawlike statements differ significantly from accidental generalizations in that causal statements perform a different function within our linguistic system, one that is not purely descriptive but is importantly prescriptive. They express our recognition of a standing permission to make certain inferences. (deVries, 2005, p. 146)

For Sellars, lawlike statements play no role in the object language. Instead, they are useful in a meta-language that we can use to make explicit our conceptual commitments. For instance, "Lightening causes thunder" is a generalization that expresses a conceptual relation between *lightening* and *thunder*, and which explains an inferential disposition: the disposition to infer "A thunder will soon come" after seeing a lightening. In this sense, lawlike statements —as sentences— have no direct role in our first-order linguistic practice. Instead, they are used for justifying or explaining our claims in discursive contexts, that is, in a second-order linguistic practice:

> To make first hand use of [lawlike] expressions is to be about the business of explaining a state of affairs, or justifying an assertion. Thus, even if to state that p entails q is, in a legitimate sense, to state that something is the case, the primary use of "p entails q" is not to state that something is the case, but to explain why q, or justify the assertion that q. (Sellars, 1958, p. 283)

Now, according to Sellars, the correct way of understanding the role of lawlike statements in thought and language is as modal —subjunctive— expressions that represent "inference tickets":

> If anything were A, it would be B...is actually an inference ticket, and not, so to speak, a letter of credit certifying that one has a major premise and a formal inference ticket at home. (Sellars, 1958, p. 286)

In this regard, Sellars claims that we should not think about induction as producing universally quantified statements that we use as "major premises from which we reason," but as "establishing principles in accordance with which we reason." And these principles have a subjunctive structure of the form "If anything were A, it would be B." (Sellars, 1958, pp. 286-287).

To sum up, Sellars saw our inferential practice as completely regulated by material rules capturing empirical generalizations and giving empirical content to natural language. Formal rules of inference, like those of classical logic, have a relatively marginal role in everyday reasoning, and in particular, they are not meaning-constitutive. On the other hand, material rules are essential for articulating the content of lexical concepts in the language, and they operate at an implicit level. We make them explicit in the form of lawlike statements only when a second-order justificatory practice requires it —typically in discursive contexts.

What remains unexplained in Sellars' view is where these material rules of inference come from, and how the agents apply them in personal-level reasoning. In general, a fine-grained analysis of material validity lacks within the inferentialist framework. Instead, Sellars (and Brandom) assume that material validity is just a matter of rule-following. Still, they do not discuss how we follow or represent these rules from a cognitive perspective. I will come back to this issue later in this chapter. In what follows, I will briefly discuss some relations between Sellars' characterization of our inferential practice with some

ideas coming from cognitive psychology.[6]

5.2.2 Connections to the psychology of reasoning

The previous sections just scratch the surface of Sellars' ideas on language and thought. However, they are (hopefully) enough to support a small point I want to make next. I believe that Sellars' framework is in line with some contemporary ideas in cognitive psychology about the roles of intuition and conceptual knowledge in reasoning.

The first connection I want to make concerns Mercier and Sperbers' argumentative theory of reasoning (Mercier & Sperber, 2011, 2017). As briefly explained in Chapter 4, this theory departs from the assumption that our inferential mechanisms exploit empirical regularities codified in different —private and public— representational structures. According to them, most of these mechanisms operate at the sub-personal level, while their outputs "come to mind" as intuitions, i.e., a kind of mental representation with a peculiar "meta-cognitive profile": Regardless of the opacity of their context of production, we confidently deem their content veracious (Mercier & Sperber, 2017, p. 66). For example, the "mindreading module" provides us with intuitions about others' mental states —their beliefs, desires, emotions, and so on—, such as "John thinks that that Maria is lying." We do not arrive at these ideas by a deliberate and effortful process; but they just pop up to consciousness as the upshot of an intuitive inferential mechanism in which we blindly trust. Now, in discursive or argumentative contexts, when we need to justify or explain our claims, we must make explicit the ideas supporting them. For explaining how do we do this, Mercier and Sperber postulates a meta-representational module —the "reason module"— that takes representations as inputs and give as back other representations that would work as premises or justifications for the first claims (Mercier, 2012; Mercier & Sperber, 2009).

[6]Beyond philosophy of language, Sellars' notion of material inferences has been recently used in philosophy of science for tackling induction (Norton, 2003) and conceptual change (Brigandt, 2010).

The connection I want to make is the following. For Sellars, our inferential practice is mostly *material* in the sense that it directly uses intuitions about regularities expressed in conceptual relations. When we need to explain or justify our claims —while arguing with others, for instance—, we come up with lawlike statement that expresses these intuitions about our conceptual commitments. In Brandom's terminology, in this way, we "make them explicit" (Brandom, 1998b). Thus, both in Sellars and in the argumentative theory, reasoning unfolds at two different levels: one is the first-order —intuitive— practice governed by material inferential rules; and the other one is the second-order, justificatory practice, in which we give reasons for our claims in a meta-linguistic or meta-representation mode. That being said, neither Sellars nor Mercier and Sperber explains how these intuitions about conceptual relationships work. This explanation is part of what this chapter attempts to do. [7]

The second point concerns a tradition in cognitive psychology that studies reasoning building on Tversky and Kahneman distinction between intensional thinking and extensional thinking (Tversky & Kahneman, 1983) . Intensional thinking is a kind of intuitive inferential mechanisms that exploit conceptual relations stored in semantic memory, disregarding extensional or probabilistic criteria(Hampton, 2012). On the other hand, "extensional thinking" refers to reasoning about the *extensions* of categories, within quantificational or probabilistic contexts. For instance, an inference like "Some football players run fast. Thus, at least one football player run fast" is based on extensional, rather than intensional, considerations.

In the psychological literature, intensional relations are often explained as material relations between concepts (e.g., see Hampton, 2012; Sutherland & Cimpian, 2017). For instance, the *intension* of the concept *bird* links it to other concepts like *have feathers*, *fly*, or *lay eggs*. Intensional reasoning consists on using these relations in generic statements like "ducks lay eggs," or "apples are sweet," to make inferences about categories or individuals. There is a growing

[7] It is fair to say that, beyond this coincidence, there is a fundamental difference between Sellar's and Brandom's inferentialism and the argumentative theory. The former prioritize inference over representation, while for the latter representations seems to come before inferring.

experimental literature showing that this kind of reasoning is highly prevalent in everyday cognition (Cimpian, Brandone, & Gelman, 2010; Hampton, 2012; Leslie, 2008; Leslie & Gelman, 2012). As Hampton explains:

> Whereas extensional knowledge provides for clear reasoning about the world, intensional knowledge is full of the richness and vagueness that turns out to characterize much of our everyday intuitive thought and language, for better or worse. (Hampton, 2012, p. 399)

Notice that this distinction mirrors —to some extent— the discussion between extensional and intensional views in semantics described in Chapter 2, as well as Sellars' and Brandom's idea that material relations among concepts are the crucial feature to account for when characterizing our everyday inferential practice, while logical —and extensional— relations are mostly marginal in this sense.

5.2.3 Limitations of the inferentialist approach

A salient feature of Sellars' philosophy is that it is strongly language-centered. This is particularly the case in his philosophical psychology, where he seems to subsume the analysis of thought and concepts to the study of linguistic activity. According to Sellars, mental activity is ontologically —and causally— prior to linguistic behavior, but the latter is prior to the former in the order of explanation (Sellars, 1991, pp. 161-164). As a consequence, a theory of concepts as mental entities must come from a theory of linguistic activity. In other words, we can only rationally reconstruct conceptual content and conceptual structure in terms of material relations of sentences within the game-like structure of natural language (cf. Marras, 1973).

The point that I want to bring out in this section is that the Sellars-Brandom overtly linguistic approach to conceptual content could have some limitations when explaining our conceptual activity. In particular, material rules and their roles in argumentative reasoning dynamics do not seem to be enough for telling the whole story about our conceptual systems and their structural features.

The main issue I see here is the centrality of the notion of language-based inference in their view of meaning. Sellars' seems to understand this as the only meaning-constitutive mechanism. We can see this in his discussion about the differences between "sentient" and "sapient" organisms (Brandom, 2015; Brown, 1986). We, humans, share with other organisms the ability to respond differently to external stimuli. Systematized differential response to external inputs can be seen as the crucial feature behind categorization, and consequently, to concept possession. But, can we say that animals, for instance, possess concepts because they have this ability? Sellars' answer is no. A parrot can be trained to utter the word "red" every time it sees a red object. Still, this does not mean that he possesses the concept *red*. According to Sellars, it only means that the parrot is in a causal state mediated by an *habit*; but not in a *cognitive* or *epistemic* state (Sellars, 1991, pp. 90, 131). For being in an epistemic state regarding the property *red* requires the agent to be able to *understand* the implications of classifying something as red, like for example, that the thing will also be colored, and that it will not be green or yellow. Sellars thinks that having the right verbal behavior in the appropriate circumstances is, in fact, a prerequisite of concept acquisition, but it is not equal to concept possession. This last capacity involves being able to use the concept to "articulate reasons" within a justificatory context —-a language game of giving and asking for reasons—, and this is something that only *sapient* creatures (and not merely *sentient* ones) has. As Brandom explains:

> What the parrot lacks is a conceptual understanding of its response. That is why it is just making noise. Its response means nothing to the parrot—though it may mean something to us, who can make inferences from it, in the way we do from changes in the states of measuring instruments. The parrot does not treat red as entailing colored, as entailed by scarlet, as incompatible with green, and so on. And because it does not, uttering the noise 'red' is not, for the parrot, the adopting of a stance that can serve as a reason committing or entitling it to adopt other stances, and potentially in need

> of reasons that might be supplied by still further such stances. By contrast, the [human] observer's utterance of 'That's red', is making a move, adopting a position, in a game of giving and asking for reasons. And the observer's grasp of the conceptual content expressed by her utterance consists in her practical mastery of its significance in that game: her knowing (in the sense of being able practically to discriminate, a kind of knowing how) what follows from her claim and what it follows from, what would be evidence for it and what is incompatible with it. (Brandom, 2015, p. 102)

Two things emerge from here. One is that concept possession depends exclusively on inferential articulation. And the other is that this articulation depends on the public practice of giving and asking for reasons, which characterizes language games. Brandom took this pragmatic character of inferentialism further (Brandom, 1998b, 2000), by claiming that semantics must be completely subsumed within pragmatics (see MacFarlane, 2010, for an analysis of this):

> Pragmatism in this sense is the view that what attributions of semantic contentfulness are for is explaining the normative significance of intentional states such as beliefs and of speech acts such as assertions. Thus the criteria of adequacy to which semantic theory's concept of content must answer are to be set by the pragmatic theory, which deals with contentful intentional states and the sentences used to express them in speech acts. (Brandom, 1998b, p. 143)

The "bridge" that allows to reduce semantics to pragmatics is *inference*, understood as a "*kind of doing*" (Brandom, 1998b, p. 91) that is regulated by a normatively structured linguistic practice. In this sense, inference appears to be prior to *representation* in the order of explanation (cf. Kremer, 2010). As such, any systematic explanation of concept representation and conceptual activity must emerge from a theory of material inferring understood as a practice embedded in a socially regulated language game.

It seems unlikely that a pragmatic framework can offer a complete explanation of semantic representation, and in particular, of those features of conceptual structure that are related to perception, which are cross-cultural and pre-linguistic (see, Barsalou & Wiemer-Hastings, 2005; Prinz, 2004). I want to suggest that there is no full explanation of material inferring, understood as a personal-level mechanism, without accounting for those fundamental cognitive mechanisms that are behind concept formation, and which provide the "skeletal principles" that give structure to our conceptual system.

As Susan Carey (2015) argues, a common problem with philosophical theories of concepts is that they systematically overlook the issue of concept acquisition. Inferentialism in not an exception to this. If concept possession depends entirely on the mastering of a set of material inferences that are determined by a public practice, then it seems that only competent language users can genuinely have concepts. The question about how cognitive agents come to acquire these concepts remains unanswered. What is more, since semantic representation depends ultimately on pragmatics, it seems that any possible inferentialist explanation of concept acquisition must be developed in the theoretical vocabulary of that discipline.

However, this has not been the case for psychological explanations of concepts. In order to explain the highly complex phenomenon of concept acquisition, most psychological theories make use of a conceptual toolbox that goes well beyond pragmatics. The first one to mention is the notion of an innate *quality space* that would give a basic structure the identification of perceptual features. Quine (1969a) and others have claimed that this is a necessary assumption for explaining concept formation to some extent. Within psychology, similar ideas are generally assumed in order to explain judged similarity among objects as a fundamental learning mechanisms behind concept formation (See for example, Decock & Douven, 2011; Goldstone et al., 1997; Medin, Goldstone, & Gentner, 1993; Rips, 1989). Perceived similarity, as well as typicality effects discussed in the previous chapter, can be certainly modulated by social and cultural factors. However, in early cognitive development the mechanisms behind perceived similairty and categorization seem to be innate, or at least, pre-lingusitic (Harnad,

1987; Mahon & Caramazza, 2011).

Statistical learning theories are quite successful in explaining the innate mechanisms behind categorization and language acquisition (see French, Mareschal, Mermillod, & Quinn, 2004; Romberg & Saffran, 2010). In a different line, some psychological theories of concepts distribute the load of concept formation between innate learning mechanisms and sets of primitive concepts that give some basic structure to experience at the beginnings of learning (cf. R. Gelman, 1990). Following Chomsky's *poverty of stimulus* argument (Chomsky, 2005), these approaches assume that the structure of physical inputs is too unstable for allowing the recovery of basic perceptual features, regardless how powerful the underlying learning mechanism are. Thus, they propose the existence of skeletal principles of knowledge, such as "core domains", "innate theories", and/or "core concepts" that would work as the *ground* from which complex conceptual knowledge emerges (Carey, 2000, 2015; Keil, 1979; Mandler, Bauer, & McDonough, 1991; Spelke & Kinzler, 2007).

In general, the existence of cognitive constraints and/or innate learning mechanisms behind concept formation seems to be an unavoidable assumption in most psychological theories of concepts. I believe this suggests that an explanation of semantic content cannot be developed in purely pragmatic terms, but has to account for those "hardwired" cognitive processes underlying conceptual acquisition. I do not see how this can be done from Sellars-Brandom inferentialism, where concept acquisition is reduced to socially regulated rule-following. This might be one of the reasons behind their inability to produce a fine-grained explanation of material inferring.

In what follows, I propose some guidelines for doing this. My approach is different to classical inferentialism. I do not consider inference to be prior to representation; neither I see rule-following as an explanatory useful notion. Instead, my analysis of material inference builds on ideas from cognitive semantics and uses conceptual spaces as modeling framework.

5.3 Word classes and types of material inferences

Philosophical semantics often prioritize coarse-grained analyses of meaning-related phenomena. It is not common among philosophers to discuss notions like *meaning* , *reference*, or *concepts* by distinguishing lexical categories, semantic types, or by classifying kinds of pragmatic contexts modulating sentential meaning. This is also the case, I think, for Sellars-Brandom inferentialism, where *concept* is a cover term for any kind of extra-logical term that can take part in a material inference.[8] That represents, I believe, a substantial limitation for any attempt to put together a systematic theory of material inference.

It seems quite evident that there are different kinds of concepts (Medin, Lynch, & Solomon, 2000). Abstract concepts like *freedom* or *causality*, motion verbs like *run* or *swimm*, spatial prepositions like *below* or *near*, artificial concepts like *bachelor*, and natural kinds categories like *dog* or *tree*, are all different from a cognitive perspective. They are learned at different stages of development and they probably require the availability of different cognitive resources, representational structures, and learning contexts (see, S. Gelman, 2009; Nagy & Gentner, 1990; Waxman & Leddon, 2011).

There are different ways of classifying lexicalized concepts.[9] For reasons of space, I am not going to discuss this issue here. Instead, I will build in the standard classification of *lexical categories* used by linguists that distinguish between *nouns*, *adjectives*, *verbs*, and *prepositions* —among others— (Baker, 2003). I will follow Gärdenfors analysis showing that different lexical categories have relatively different representational structures in conceptual spaces (Gärdenfors, 2014).

In the same way that there are different kinds of lexicalized concepts, there are different kinds of relations among them. For instance, *dog* and *table* are

[8] Just as classical logicians use the term *predicate* as a cover term for any non-logical concept. As Gärdenfors explained (2000), this notion is too coarse-grained for capturing all the nuances of conceptual representation.

[9] A *lexicalized concept* is one that has a corresponding word in the public lexicon, like *dog* or *festival*. We obviously do not have lexical concepts for every class of object. For instance, there are no English words for the class of animals that were "drawn with a very fine camel hair brush", or for "those that tremble as if they were mad", for using Borges' famous examples in his story "The Analytical Language of John Wilkins." In general, the availability of lexical concepts within a language seems to be, to an important extent, culture-specific.

subordinate concepts of *mammal* and *furniture* respectively. Linguists called this semantic relation *hyponymy* (L. Murphy, 2003, pp. 216-230). Hyponyms are in a *type-of* relation with their *hypernyms* —or "superordinates". At the same time, two lexical concepts that have the same hyperym are called *co-hyponyms* (see Figure 5.1).

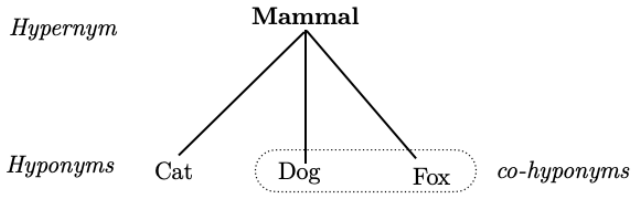


FIGURE 5.1: Hyponomy scheme.

These kinds of relations are pervasive in natural language since they give structure to conceptual taxonomies. Taxonomical structures are behind a specific kind of inference-pattern called "property inheritance" (see Etherington & Reiter, 1983; Sloman, 1998a), which allows the agent to attribute properties of the hypernym to the hyponym. For instance, if the only thing I know about platypuses is that they are mammals, I can use my knowledge of mammals to infer that they are also vertebrates, that they breathe and they have fur — among other things. Likewise, if I get to know that *platypus* is a co-hyponym of *cat*, then I can materially infer "$Platypus(x) \rightarrow \neg Cat(x)$", since co-hyonymy often implies semantic incompatibility (Cruse, 2002). [10]

In general, different kinds of semantic relations allow for different patterns of material inferences among lexical concepts. Thus, a systematic theory of material inferences must build on the classification of these relations among lexical categories. Consider, for example, the following inferences:

(c) *The car is red. Thus, the car is colored.*

(d) *Fido is a dog. Thus, Fido is a mammal.*

[10] Cognitive linguists studying lexical organization call these relations *sense relations* (see, Cann, 2011; Storjohann, 2016). Hyponomy, antonomy, synonymy, and meronomy are some well known examples of them. However, they rarely study them from the perspective of their inferential affordances —an exception of this might be Cummins analysis of the pragmatics of inference in Cummins (2013, Chapter 3).

(e) *Paris is north of Marseille. Thus, Marseille is south of Paris.*

(f) *The barbecue is behind the house. Thus, the house in front of the barbecue.*

(g) *Maria is Marco's wife. Thus, Marco is Maria's husband.*

(h) *Felix is a cat. Thus, Felix is not a bird.*

For Sellars-Brandom inferentialism, all the sentences above fall into the category of *material inference*, and their framework offers no tools for classifying them. However, there are clear differences and similarities between them. (c) and (d) are about different topics, but both exploit properties of the hyponymy-hypernymy relation. (e) and (d) are also topically different but they both build on properties of how spatial prepositions are cognitively represented. Finally, (g) is an inference based on the fact that the pair *wife*:*husband* are complementary terms, just like *teacher*:*student*; while (h) exploits the semantic incompatibility of co-hyponymy. In the following section, I will use conceptual spaces for explicating these semantic relations behind different patterns of material inferences according to how they are cognitively represented.

5.4 Explicating material inferences *via* conceptual spaces

5.4.1 Preliminary remarks: *core* meaning and attention shifts

There is no clear criterion for defining the set of implications that can be materially inferred from a concept-type in the scarce literature on material inferences. As said before, a systematic theory of this kind of inferences requires one. It is beyond this chapter's aims to build such a theory. Nevertheless, I believe that some of the principles presented in what follows could help this endeavour. Before presenting the basics of the model, I will briefly comment on two important issues on how to understand material inferences.

The first point concerns how to delimit which conceptual information an inference must exploit to be called "material." For instance, the inferences "$Apple(x) \rightarrow Fruit(x)$" and "$Apple(x) \rightarrow Red(x)$" are different in the sense that the former uses information from the basic definition of *apple* while the latter uses information that is probable, but not necessary for all apples. In other words the first inference seems to be truly "valid in virtue of meaning" while the other is under uncertainty, and thus nonmonotonic. I believe that Sellars' approach to material inferences was mostly concerned by the first kind of inferences, because of its relation to the notion of analytic entailment. Robert Brandom, on the other hand, thinks that material inferences use any kind of conceptual information, and that are mostly nonmonotonic (Brandom, 1998a; Schaefer, 2016).

This issue connects directly with the problem of establishing the "amount" of conceptual content characterizing concept possession. For instance, having the concept *dog* seems to require to also have the concepts *animal*, *four-legged*, and *domestic*; but not necessarily to have concepts like *sesamoid bone* or *vomeronasal organ*. Kiefer (1988) and Bierwisch and Kiefer (1969) distinguished between three different "layers" of lexico-conceptual knowledge. First, there is knowledge which constitutes the "core meaning" of a lexical item. Second, there is general "conceptual knowledge", which "*concerns predictable modifications of the core meaning in various contexts*" (Kiefer, 1988, p. 2); and finally, "encyclopedic knowledge," which goes beyond linguistic competence and is related to expertise.

As Marconi has remarked (1997, Chapter 2), drawing a sharp line dividing these three layers is a big philosophical challenge that remains unanswered (see also Paradis, 2003) . My analysis of material inference will focus on inferences that can be made with the "core meaning" of concepts. Something very similar to what *definitional* meaning is. I consider material inferences to include no uncertainty —or at least to be *minimally uncertain*. It seems to me that, from a personal-level perspective, someone performing an inference based on core meaning is entirely sure of its correctness. On that depends her being a competent user of the concept. If that person realizes that the inference was

wrong for some reason, this will lead to conceptual revision. In Chapters 6 and 7, the conceptual-space model will be extended to other kinds of concept-based inference that include uncertainty.

The second point concerns a theoretical hypothesis about the relation between material inference and attention. As said in Chapter 4, one of the main theses here defended says that material inferences are transitions between informational states that exploit properties of the semantic structures representing the lexical concepts in the premise and the conclusion. Now, what kind of cognitive mechanism *drives* the act of inferring? I think that the answer to this question can be found in the relation between attention and conceptualization, as it is construed in some theories within cognitive semantics (cf. Marchetti, 2015; Talmy, 2007). In particular, a central mechanism behind material inferences is "re-profiling" (Langacker, 1987), an attention-based cognitive ability allowing to change focus within conceptual configurations.

As explained in Chapter 2 .3, the central tenet of cognitive semantics is that meaning is conceptualization. That is, lexical concepts and sentences evoke different representational structures in the mind of the speaker/listener; structures which are a condition of possibility of language understanding (Fillmore, 2006; Langacker, 2000). To break down this idea, cognitive semanticists often build on the notions of *figure* and *ground*, a famous distinction introduced by gestalt psychologists for characterizing the organization of perceptual experience. Talmy (1975) brought it to linguistics for accounting for the meaning of sentences expressing spatial relations.

Location and motion relations are expressed by sentences that specify the position of one object —the *figure*— in relation to another object —the *ground*—, as in the sentence "the pencil [figure] is on the desk [ground]." Langacker took up this idea under the terminology *profile-base alignment* and generalized it as the main trait of sentential meaning. Roughly, Langacker stressed that lexical items and sentences always specify their content in association to organized clusters of concepts. For instance, the word "Tuesday" can only be understood against a base composed by the concepts *day* and *weekend*. The semantic structure of sentences typically consist in a *designatum* that "stands out" against

a background —or *base*— that is composed by other concepts and conceptual relations:

> A predication always has a certain scope, and within that scope it selects a particular substructure for designation. To suggest the special prominence of the designated element, I refer to the scope of a predication and its designatum as base and profile respectively. Perceived intuitively, the profile.... "stands out in the bas-relief" against the base. The semantic value of an expression resides in neither the base nor in the profile alone, but only in their combination; it derives from the designation of a specific entity identified and characterized by its position within a larger configuration. (Langacker, 1987, p. 183)

Attention is the fundamental cognitive ability behind profile-base alignments (see, Lampert, 2009; Talmy, 2007). Cognitive linguists often use visual perception as an analogy for clarifying the role of attention in conceptualization (see, for example, Langacker, 1987, p. 116 or Gärdenfors 2014, Sec. 1.3.3). For instance, navigating our environment requires constant visual scanning to identify possible obstacles and estimate distances. For that, we use procedures of figure-ground segregation which build on attentional mechanisms.[11]

Linguistic structures show similar features. The semantic content of sentences is specified by focusing on particular elements within rich conceptual frameworks. Using Langacker's terminology, sentences impose a profile over a conceptual base. For instance, the sentence "John is painting the door green" profiles an event in which an action performed by an actor is taking place. The same event can be re-profiled by changing the designatum, like in "the door is being painted green by John." In this last sentence, the same event is described with a different profile, one in which the object that was part of the base now is standing out.

[11] Some theories of concepts also assume that attentional mechanisms are behind the variety of usages that concepts have in everyday cognition. For instance, selective attention seems to play a crucial role in choosing, from a concept, the relevant information —or features— used in different cognitive procedures, like inference or categorization (Barsalou, 2003; Schyns, Goldstone, & Thibaut, 1998).

Re-profiling —or "alternate profiling", in Langacker's words— is an attentional shift within a conceptual base that produces a minimal semantic transformations. I submit that material inferences are cases of re-profiling propositions within their correspondent conceptual structures. For instance, an inference like $Dog(x) \rightarrow Mammal(x)$ consists on re-profiling the object x within the conceptual space of *dog* towards the conceptual space of *mammal*, which is implicit in the representation of dog because the former is a sub-region of the latter.

My explication of material inferences builds on two assumptions that I will develop in upcoming sections. The first concerns object-representation in conceptual spaces. A judgment expressing that entity x falls under concept C, is represented by an arbitrary point within the conceptual space of C. In high-order predication —when the subject of the proposition is not an entity but a concept—-, instead of representing an arbitrary point, sets of points (regions) of some conceptual space are represented. [12] The second assumption is that such inferences' validity does not depend on logical form, but on the formal properties of the conceptual structures that work as bases of the premises' sentences. The remainder of this chapter consists of an analysis of these properties in some word classes.

5.4.2 Nouns

We start by discussing material inferences with nouns. According to the conceptual space-model, nouns correspond to concepts and, as such, they are a convex region in a conceptual space, i.e., a subset of the product-space of the set of dimensions that constitute the space. Starting from a concept M, $\mathscr{C}(M)$ corresponds to a subset of the Cartesian product of n domains. As explained in the previous chapter, an object x falling under M corresponds to a n-dimensional point $x =< x_1, x_2, \ldots, x_n > \in \mathscr{C}(M)$ with the coordinates of the point in each dimension.

[12] In the next chapter, this assumption will be adapted for explaining inferences under uncertainty.

In most cases, x is a partial vector since the information available —the specific values on each of the dimensions— is sparse.[13]This lack of information translates into uncertainty when reasoning with these points. In the following chapters, it is argued that everyday reasoning deals with these information-gaps by "filling out" the unknown elements of the vector with the values of the prototype of the concept. The resulting inferences are based on *expectations* (Osta-Vélez & Gärdenfors, n.d., 2020a). For instance, if I am told that x is an apple, then I will feel entitled to make inferences under different degrees of uncertainty regarding the prototypical properties of x, like being round, red, and/or sweet.

Since expectation-based inferences build directly on semantic knowledge, they could also be called *material*. However, they differ from core material inferences in an important respect. The former are inferences under uncertainty that, if defeated, leave our background knowledge structure unchanged. On the other hand, the latter are not perceived by the agent as defeasible, since they reflect basic conceptual understanding. A defeated core material inference leads to conceptual revision, that is, to restructuring parts of the agent's conceptual system. Consider, for instance, an inference like $Eggplant(x) \rightarrow Vegetable(x)$. This inference is taken as correct in many everyday contexts of use, however, it is wrong, since eggplants are botanically classified as berries. If, while believing that eggplants are vegetables, I make that inference and someone who knows better corrects me, I will revise my concept of *eggplant* relocating it in a different part of the plant conceptual space —a case of conceptual change—, and I will avoid to make that inference again.

Going back to nouns, I will distinguish two kinds of core material inferences with nouns: *top-down* and *bottom-up* . In bottom-up inferences, the concept in the premise is a *hyponym* —a subordinate— of the concept on the conclusion, like in $Cat(x) \rightarrow Mammal(x)$. From the conceptual-space perspective, the material validity of this kind of inference lies in a straightforward set-theoretical fact. As we said above, when an entity x is categorized as being N, it is

[13]Here I am using "vector" as *sequence of coordinates*. Talking about *vectors* in a more technical sense could lead to confusion since the dimensions that compose a conceptual space are not necessarily isomorphic to the real numbers, some could be discrete.

represented as a point in $\mathscr{C}(N)$. Thus, for any concept M such that $\mathscr{C}(N) \subseteq \mathscr{C}(M)$, if $x \in \mathscr{C}(N)$ then $x \in \mathscr{C}(M)$. Furthermore, since the inclusion relation is transitive, $\mathscr{C}(N)$ will be included in every superordinate concepts of M. Thus, categorizing x as N will make it (by default) a member of every superordinate concept of N.

Let us now turn to top-down material inferences. These are transitions that exploit properties of the internal structure of the concept in the premise. In the case of nouns, top-down inferences are directed towards the properties that are subregions of the dimensions in the noun's conceptual space. I will introduce two different top-down inferences through an example. Assume that N is a noun spanning across 3 discrete dimensions, D_1, D_2, D_3. We will have that for any $x \in \mathscr{C}(N)$, $x =< x_1, x_2, x_3 >$ with $x_i \in D_i$. Now, let's further assume that D_1 and D_2 are partitioned into two disjoint subregions R_1^1, R_1^2, and R_2^1, R_2^2 respectively; and that that D_3 is not partitioned at all in $\mathscr{C}(N)$ —i.e., objects falling under D_3 have always the property $R_3 \subseteq D_3$.

We can now identify two different kinds of top-down material inferences from N. The first kind, which is hardly informative, consists on going from "$N(x)$" to "$D_i(x)$" ($i \in \{1, 2, 3\}$), where "D_i" is the name of the dimension. These inferences only require to acknowledge that an object falling under a concept will also have a "value" in each dimension that constitute the conceptual space of the concept. Inferences like this are often used as examples of material inferences in the philosophical literature. For instance, "x is a car. Thus x has a color" or "x is a man. Thus, x has a height".

The second type concerns inferences from N to the properties in the dimensions of $\mathscr{C}(N)$. In our toy model, these are the disjunctions $N(x) \rightarrow R_1^1 \vee R_1^2$ and $N(x) \rightarrow R_2^1 \vee R_2^2$ for domains D_1 and D_2; and the inference $N(x) \rightarrow R_3$ for D_1. I take this last inference to be the most informative of top-down material inferences, but it depends on the concept spanning exclusively across only sub-region of the dimension. Examples of this could be $Mammal(x) \rightarrow Vertebrate(x)$, or $Bechelor(x) \rightarrow Young(x) \wedge Male(x)$.

5.4.3 Co-hyponymy and material inferences with negation

As said earlier, co-hyponyms are concepts at the same conceptual level that share one or more common super-ordinate category —the *hypernym*. Classic examples are sets of categories like *cat-dog-rat-cow* or *blue-green-yellow-red*. Co-hyponymy is often considered as an incompatibility relation (see Cruse, 2000, Chapter 9): If an entity x falls under category N, then it cannot fall into N's co-hyponyms. For this reason, hyponymy is inferentially rich: representing x as falling under N, with $M_1, \ldots, M_n$ co-hyponyms of N, typically entails $\neg M_i(x)$.

Explicating this kind of material inference with negation in CS is straightforward. The set of categories $M_1, \ldots, M_n$, co-hyponyms of N, will be represented as disjoint subregions of the $C(N)$. This means that $M_i \cap M_k = \emptyset$ for all $i \neq k$. Since the sentence "$M_i(x)$" is represented by an object $x \in M_i \subseteq \mathscr{C}(N)$, then $x \notin M_k \subseteq \mathscr{C}(N)$ since M_i and M_k are disjoints subregions of $\mathscr{C}(N)$. This simple set-theoretical fact justifies all the material inferences of the form "$M_i(x) \rightarrow \neg M_k(x)$" for every category M_k co-hyponym of M_i. Normally, biological kinds are good examples of co-hyponyms in this sense. For instance, representing the sentence "$Dog(a)$" materially implies "$\neg Cat(a)$", "$\neg Seal(a)$", $\neg Whale(a)$, and so on, for any animal category, co-hyponyms of *dog* and at the same conceptual level. [14]

[14] One problem with this view of co-hyponymy is the existence of *compatible* co-hyponyms, like *queen* and *mother*, whose hypernym is *woman*; or *bachelor* and *actor* which are hyponyms of *man* (L. Murphy, 2003, pp. 217-219). As it is evident, the intersections of the sets that represent these hyponyms are not empty. For instance, some actors are bachelors, and some mothers are queens. The literature in cognitive linguistics is not clear on how to explain this ambiguity. I think that conceptual spaces can shed light into this issue. As explained before, a conceptual space is a similarity space partitioned into different sub-regions representing categories with the same dimensionality. For example, the conceptual space of *bird* is partitioned into several subregions for the categories *robin*, *penguin*, *pigeon*, and so on. Those categories are clear cases of incompatible co-hyponyms. However, there are other categories that also target sub-regions of the *bird-space* but are not part of the "natural partitioning" of the space. For instance, *seabird* spans across those sub-regions of $\mathscr{C}(BIRD)$ corresponding to categories having "sea" as fixed value in the habitat dimension. Similarly, *wader* refers to categories that occupy a specific subregion of the shape domain: birds with long necks and long legs. *Wader* and *seabird* are compatible co-hyponyms of *bird* because they "block" different properties in different domains. I suggest that this idea can be generalized as part of a definition of compatible co-hyponymy. Furthermore, it seems to me that the compatibility of co-hyponyms like *bachelor/actor* or *queen/mother* is also due to the fact that they are not the result of the same partitioning of the space —a "person-space" in this case—, but they are ad-hoc categories that *profiles* a person with salience in one or more particular dimensions. If two co-hyponyms profile an object from their hypernym 's conceptual space with salience in different dimensions, then they should be compatible.

5.4.4 Spatial prepositions and relational concepts

Prepositions form a class of words that serves multiple cognitive and communicational functions (see Lindstromberg, 2010; Tyler & Evans, 2003). *Spatial prepositions* are a sub-stet of them used for specifying locations, directions, motions and other space-related states of affairs (see Cuyckens, 1997; Jackendoff & Landau, 1993); and that seem to be grounded on some primitive intuitions about our representation of space and spatial relations among objects (Levinson, 2003; Talmy, 1983).

Prepositions are often overlooked in the philosophical literature on concepts and inference. However, some classic examples of material inferences include them. In what follows, I propose an explication of material inferences with some typical cases of spatial prepositions.

Let's start by defining some of the fundamental notions used in cognitive linguistics in the analysis of spatial prepositions. Consider the following sentences:

(i) Buenos Aires is *west of* Montevideo.

(ii) John *went to* the supermarket.

(iii) The cat is *on* the mat.

The three sentences above include entities that are being described in some spatial relation to some other entity. The described entity is called "trajector". It can be static, like "Buenos Aires" in (i) or "the cat" in (iii), or dynamic like "John" in (ii). The other entity is called "landmark," and it is the object according to which the trajector's location —-or trajectory of the motion— is specified (Langacker, 1987).

It is important to point out that spatial terms do not tell the exact locations of entities in space. Instead, they give information about a region in which an object might be within a particular spatial configuration. For instance, the sentence "The pen is *on* the desk" would be true if the pen is on the center of the desk, on one corner, or in any other possible position within the table's surface. We will come back to this point later.

Now, the guiding idea behind my analysis of material inferences with spatial prepositions is analogous to the one presented in the previous subsection. It builds on the assumption that we conceptualize sentences with those terms through representational structures that are exploited in inference in different ways. Again, I assume that these representational structures are conceptual spaces, and I will directly build on Zwarts and Gärdenfors (2016) and Gärdenfors (2015) for my analysis.

Unlike what happened with nouns and adjectives, spatial prepositions seem to be more naturally represented in conceptual spaces by using spherical coordinate systems (Gärdenfors, 2014). Within this kind of systems, a point p in the three-dimensional space is determined via three quantities $< r, \theta, \phi >$. r is the distance between the point and the origin (O) of the system, θ is the angle between p and the x-axis —with $0^o \leq \theta \leq 360^o$—, and ϕ is the angle between p and the z-axis (with $0^o \leq \phi \leq 180^o$) (see Figure 5.2).

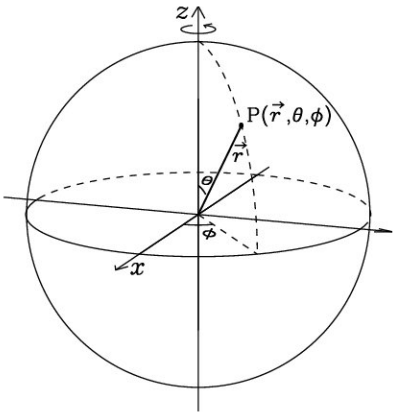


FIGURE 5.2: Polar coordinate system.

Using this coordinate system, Gärdenfors defines the notion of *polar betweenness*, which is used to show that, just as with nouns and adjectives, spatial prepositions could correspond to convex regions in conceptual spaces.

Now, following Zwartz and Gärdenfors (2016) we will take the center of mass of the landmark (L) —a point— in a spatial sentence as the origin of the coordinate system.[15] We will also assume that landmarks are circular/spherical,

[15] As Zwarts and Gärdenfors (2016, p. 118) explain, this is a strong idealization. Landmarks can have different forms and this can have an impact in the representation of the spatial relations.

with and extension r (radius). Then, a the preposition in a spatial sentences will determine a region of the space in which the trajector, represented as an arbitrary point, might be.

Let us illustrate this by analyzing the conceptual space of the preposition "CloseTo(L)". Most uses of "close" can be represented only with the horizontal plane. Thus, we can say that a trajector P is close to a landmark L iff $P \in \{< r, \theta >: c > r > r_L\}$. Notice that c has to be contextually defined. For instance, the c in the sentence "Munich is close to Augsburg" will be much larger than the c in "Maria is close to Juan" —we are only considering literal, and not metaphorical, meaning.

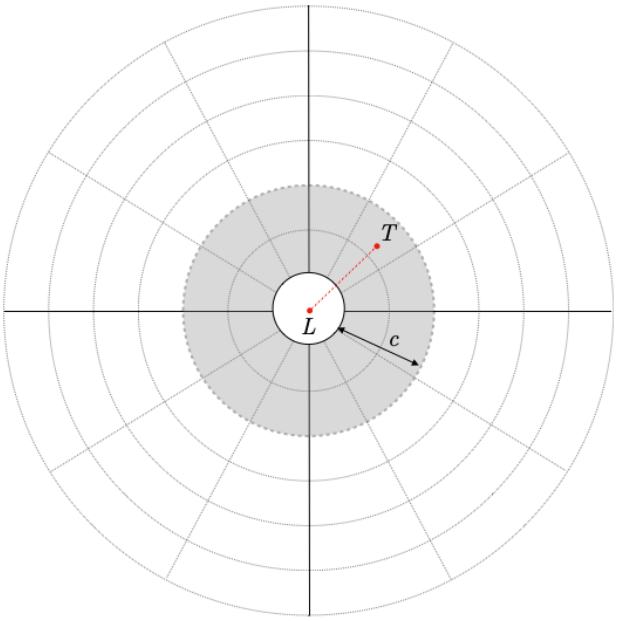


FIGURE 5.3: Representational structure for the meaning of "T is close to L".

This geometrical representation of "CloseTo" immediately shows that this preposition is inferentially symmetric. That is, if I am told that (a) "T is close to L", I can materially infer (b) "L is close to T." The representation of (a), with L as landmark, implies that $P \in \{< r, \theta >: c > r > r_L\}$. Giving that $d(L, T) = d(T, L)$, and considering that the c remains the same since (a) and (b) share context, if (a) is true then (b) has to be true since (b) "means" that $L \in \{< r, \theta >: c > r > r_T\}$.

Another material inference from this preposition is: "L is close to T. Thus, T is not far from L." Notice that *FarFrom* is an antonym of *CloseTo*. The former can be defined as occupying the complementary region of CS of the latter. Thus, considering that "T is close to L" means that $T \in \{< r, \theta >: c > r > r_L\}$, this imply that $T \notin \{< r, \theta >: r > c\}$, which means "T is not far from L." Again, this inference is explained as an (attention-based) re-profiling of the landmark and the trajector within a common conceptual base.

Cardinal Terms

The same rationale can be applied for explaining material inferences with cardinal terms. It is often claimed that knowing the meaning of cardinal terms entitle the agent to materially infer a a sentence like "Munich is south of Berlin" from "Berlin is north of Munich." These spatial prepositions can be represented in the same horizontal plane of a spherical coordinate system divided into convex regions with a common central landmark. Cardinal terms will be represented in a CS —approximately— by the following sets:

$North_L$: $\{< r, \theta >: r > r_L$ and $330^o > \theta > 30^o\}$

$East_L$: $\{< r, \theta >: r > r_L$ and $60^o > \theta > 120^o\}$

$South_L$: $\{< r, \theta >: r > r_L$ and $210^o > \theta > 150^o\}$

$West_L$: $\{< r, \theta >: r > r_L$ and $240^o > \theta > 300^o\}$

The inferential transition "T is south of L. Thus, L is north of T" is materially valid in virtue of the geometrical properties of the conceptual space in which the two cardinal terms are represented. Notice that "T is south of L" means that $T \in \{< r, \theta >: c > r_L$ and $150^o > \theta > 210^o\}$. Let's assume that T is the point $< r', \phi > \in South_L$. The inferential move from premise to conclusion implies to re-focus and take T as landmark and L as a trajector. In that case, the new coordinate system, with origin in T, will be aligned with the previous one since cardinal terms have absolute frames. We will have then that L is a point with

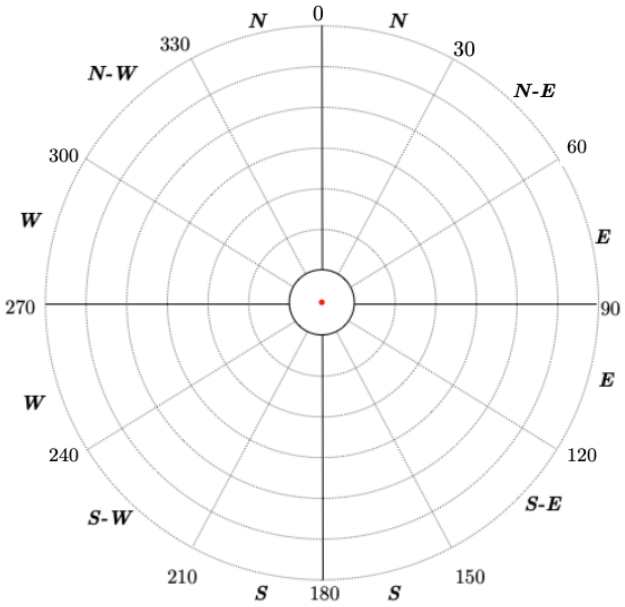


FIGURE 5.4: Conceptual space of cardinal terms.

coordinates $< r', \beta >$, and that $\beta = \phi + 180^o$. Given that $210^o > \phi > 150^o$, we have that $210^o + 180^o > \beta > 150^o + 180^o$, which is the same as $30^o > \beta > 330^o$. As it is obvious, this means that whenever $T \in South_L$, then $L \in North_T$.

Material inferences with negation, like "Munich is south of Berlin. Thus, Berlin is not east from Munich" follow the same rationale. This is also the case for many other forms of inferences with spatial prepositions having molecular sentences in the conclusion.

Spatial prepositions are relational predicates. However, they are hardly representative of this entire class of concepts because their conceptual spaces are rather peculiar. In general, the representational structures of relational concepts are diverse. Exhausitivity is beyond the aims and possibilities of this work, but to furnish my proposal with some more examples, I will briefly comment on the inferential affordances of a typical case of relational —dyadic— predicates: kin relationships.

Kin relationships can be represented in a product space of three discrete dimensions: a *gender dimension* with two possible values; a dimension representing "vertical" degrees of offspring isomorphic to the integers —son/daughter, father/mother, grandfather/grandmother, and so on— ; and a dimension representing "horizontal" degrees of kinship —brother/sister, cousin, second-cousin,

and so on— isomorphic to the natural numbers.

A relational expression like "x is the father of y" must be interpreted in the same way as locational prepositions, i.e., as a function that "locates" a *trajector* within the kinship conceptual space by using a *landmark* as reference (see Figure 5.5).

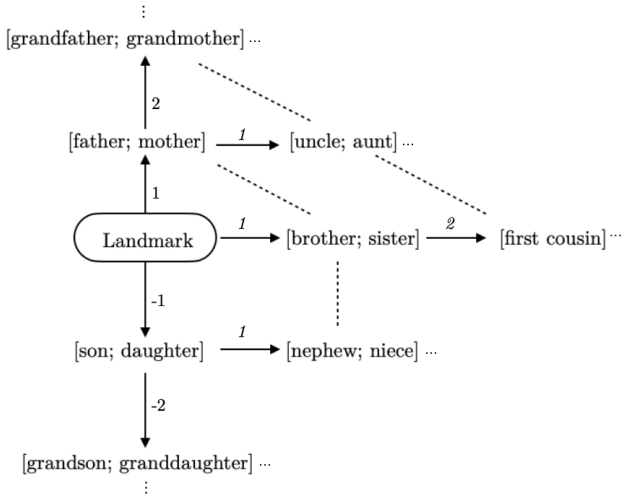


FIGURE 5.5: Kinship conceptual space.

To see some examples, in the first columns there are sentences with kin relationship terms and in the second their corespondent vectors in the CS representation.

(a) *x is the son of y.* (a') $< male, -1, 0 >$

(b) *z is the sister of y.* (b') $< female, 0, 1 >$

(c) *w is the grandfather of y* (c') $< male, 2, 0 >$

In general, these terms appear in expressions with proper names that — most of the time— indicate the gender of the landmark and trajector and allow for more precise inferences. In particular, names indicating gender allow to infer the lexical converses of the kinship relation in question.[16] A sentence like "Maria is John's mother" takes John as landmark and represents Maria with

[16] *Lexical converseness* is a sense relation defined by Lyons (1996, p. 129) and Cruse (1986, pp. 230-241) denoting pairs of terms expressing relations that have a directionality and that can be reversed. Like *teacher:student*, *buy:sell* or *employer:employee*.

the vector $< female, 1, 0 >$ in the kinship CS. Then, if by an attention shift I change the roles and take Maria as landmark, I can materially infer "John is Maria's son" since John —now trajector— will appear in the CS represented by the vector $< male, -1, 0 >$ (the symmetrical point with respect to the "horizontal" dimension), which means "son" in the CS —notice that Kin relations are represented within an absolute frame of reference.

If we do not have names, the material inferences that can be drawn are less precise. For instance, from (a) one can only materially infer "$Male(x)$," and "$Father(y, x) \vee Mother(y, x)$."

Kinship terms are —as most concepts— culturally grounded and have some degree of cross-linguistic variability (see, D. Read, Fischer, & Lehman, 2014). For instance, Spanish does not have a lexical concept for gender-neutral brother/sister relations, like "sibling" in English. Languages with kinship structures different than English also induce their users to represent and think about these relations in relatively different ways (cf., D. Read, 2013).

This tells us something about the normative structure of material inferences. Deductive validity in classical logic has always been thought of as something stable across languages since logic is not supposed to be culture-specific. On the other hand, material validity, due to its dependence on the representational structures underlying linguistic practices, is clearly culture-specific. This coheres with Brandom's idea that this kind of inferential validity emerges from socially-regulated linguistic practices. Nevertheless, in line with what was said in Section 5.2.3, the extent to which material validity is culture-specific has to be determined by considering not only pragmatic aspects but also those basic —and universally shared— cognitive constraints behind conceptual representation. [17] For instance, it has recently been argued that kin terminologies, even if subject to consistent cultural variations, are also the product of some innate constraints on conceptual representation (D. Jones, 2010). In any case,

[17] A possible way of explaining material validity can follow from (Gärdenfors, 1993), where the normativity of social meaning is explained in terms of a (social) "semantic power structure" that emerge from the interaction of individual meaning. Nevertheless, for matters of space I will not explore this possibility here.

the degree of cooperation between innate mechanisms behind conceptual representation and cultural and pragmatic factors is something that has to be empirically determined.

5.5 Summary and conclusions

In this chapter, I have discussed the notion of "material inference," a case of semantic-based inference commonly studied in philosophy. I have argued that Sellars' and Brandom's approach to this notion lacks the tools for explaining the cognitive mechanisms behind this kind of inferring. And I have proposed conceptual spaces as an alternative framework for doing this.

My proposal builds on two assumptions: (1) material inferences are transitions between mental states that exploit properties of the semantic structures representing —at a sub-symbolic level— the lexical concepts in the premise-conclusion. (2), they can be understood as cases of re-profiling a designatum within a conceptual base driven by an attention-based mechanism.

Through different examples, I provided guidelines on how to use CS to explicate this kind of inference. Nevertheless, most of the work, in this sense, remains to be done. A systematic theory of material inferences should explain the inferential affordances of each word class, according to its typical underlying representational structures. Furthermore, it is expected that such a theory can shed some light on the nature of sense-relations —like hyponymy, antonomy, and meronymy— that give structure to the mental lexicon (see Cann, 2011).

The value of the analysis presented here lies in showing how a relatively vague philosophical concept such as "material inference" — that logical models fail to explain— can be formally explicated using a psychologically informed theory of concepts and conceptual relations. In what follows, this model will be expanded to tackle inferences under uncertainty. As we will see, these inferential mechanisms deal with partial information by exploiting properties of conceptual representation that are not used by material inferring. In particular, they focus on relations of similarity and prototypicality among concepts.

Chapter 6

Nonmonotonic inference and expectation orderings

Summary

In Gärdenfors and Makinson (1994) and Gärdenfors (1992) it was shown that nonmonotonic inference can be modeled using a classical consequence relation plus an expectation-based ordering of formulas. In this chapter, it is argued that this framework can be significantly enriched by adopting a conceptual spaces-based analysis of the role of expectations in reasoning. In particular, it will be show that this approach can solve various epistemological issues surrounding nonmonotonic and default logics. [1]

6.1 Introduction

As it was previously explained, classical logic builds on two unwarranted assumptions about reasoning. First, that inference is propositionally-based; and second, that deductive validity is formal. Deductive logic is then *informationally conservative*, in the sense that the information in the conclusion is already implicit in the premises. Following this idea, deductive reasoning has been traditionally conceived as a process that does not require to exploit conceptual knowledge about the premises in order to draw a conclusion. The problem is,

[1] This chapter is based on the paper "Nonmonotonic reasoning, expectation orderings, and conceptual spaces", written in collaboration with Peter Gärdenfors.

as previously explained, that everyday reasoning builds on more than the logical form of explicit premises. Partial information and uncertainty are pervasive; consequently, our inferential mechanisms can hardly afford to be informationally conservative. Instead, we are always using our background knowledge in risky —yet productive— ways to make sense of our environment.(Oaksford & Chater, 2009) In other words, everyday reasoning is strongly nonmonotonic, and the formalist approaches based on classical logic cannot account for this.

One particular way in which this use of background knowledge expresses itself is through our expectations about the world. For instance, if we know that someone is from France, we expect the person to speak French and to have a French passport; or if we are driving a car and we spot a person waiting on one the side of the road, we expect her to intend to cross it. In general, our expectations about the world are crucial for guiding our reasoning and action in everyday life, and they build directly on the structure of our background knowledge.

Gärdenfors and Makinson (1992; 1994) have shown that much of nonmonotonic logic is reducible to classical logic, with the aid of an analysis of the expectations working as hidden premises in arguments. The guiding idea is that when people try to find out whether a conclusion C follows from a set of premises P, the background information used does not only contain the premises in P, but also information about what they expect in the given situation, so that they end up with a larger set of assumptions. Such expectations can be expressed as "default" assumptions, i.e., statements about what the reasoners represent as normal or typical. They include our core conceptual knowledge but also other information that can be regarded as plausible enough to be used as a basis for inference as long as they do not give rise to inconsistencies.

Expectations work as hidden assumptions in reasoning. The main difference they have with explicit premises is that they are "more defeasible." That is, if any of the expectations conflict with some of the explicit premises in P, we do not use them for determining whether C follows from P. However, when evaluating their role in reasoning, it is important to note that they do not all have the same strength. For example, in certain cases, we consider the relation

among two propositions to be strong enough to work as an almost universally valid rule so that an exception to it would be extremely surprising. In other situations, this relation could be better described as a rule of thumb used for drawing more precise conclusions. For instance, while walking on the sidewalk, we expect the ground to be solid enough to support our body weight; but when we are hiking in the snow, this expectation will be weaker, and therefore we will walk carefully to avoid sinking. An exception to the latter type of rule is not unexpected to the same degree as in the former case. In brief, our expectations are all defeasible, but they exhibit varying degrees of defeasibility.

This last point indicates that expectations can be ordered. Gärdenfors and Makinson's (1992; 1994) have shown that expectation orderings contains enough information to express the default assumptions used by everyday reasoning. The main idea is that a default statement of the type "F's are normally G's" can be expressed by saying that "if something is an F then it is less expected that it is non-G than that it is G." This formulation is immediately representable in an expectation ordering $<$ by assuming that the relation $Fx \rightarrow \neg Gx < Fx \rightarrow Gx$ holds for all individuals x.

However, a significant limitation of their work is that it does not provide any explanation about the cognitive origin of expectations, nor does it offer any criteria to determine their relative strength. The purpose of this chapter is to show how an extension of the CS-model developed in Chapter 5 can offer a solutions to these issues. In particular, the main objective here is to study *expectation orderings* as ways of summarizing degrees of defeasibility of our expectations regarding some piece of information about a given object.

In the previous chapter, —core— material inferences were discussed. They involve a kind of reasoning which can also be called "deductive," since it is not perceived as *uncertain* by the reasoner. Now, the model will be extended to account for nonmonotonic —uncertain— inferences *via* the notion of expectation. In particular, the focus will be on how the information a person has about *category structure* influences the person's expectations. We will first see how expectations orderings can be constructed by looking into the prototypical structure of concepts using the CS's built-in distance function. Afterward, the

connections to nonmonotonic logic will be explained. Finally, criteria for updating expectation orderings will be offered, and the explanatory advantages of the CS-model for tackling foundations issues of default logic will be discussed.

6.2 Reasoning with expectations

In brief, the position defended here is that a proper explication of the role of expectations in reasoning must rely on a model of the structure of background knowledge. As discussed in Chapter 3, even though this last notion has played a central role in several areas of philosophy and logic over the last decades, few efforts have been made to define it properly. Classic non-monotonic and default logic has worked within the formalist setting: assuming that both implicit and explicit knowledge are represented in a proposition-based format in some sort of *belief-box* of the cognitive agent. The problem is that the origin of default rules and their use in everyday reasoning remains unexplained. In what follows, we will see how a proper articulation of CS as a model of inference can point towards a solution.

As mentioned earlier, our expectations about the world mirror aspects of the organization of our background knowledge. Consider, for instance, the following hypothetical situation: our friend Maria tells us that she bought a new pet. Since dogs are the most typical pets and we know that Maria does not like cats, we can nonmonotonically infer that she got one, and start wondering about what kind of dog it is. If we get to know later that her new pet is not a mammal, then we might expect it to be a bird, since birds are less typical, but still common kinds of pets. With this new expectation, we can infer that Maria's new pet flies, since birds typically fly. [2] This would be another nonmonotonic inference, since we are still reasoning under uncertainty, and some might even say that this form of reasoning is merely guessing. However, in situations in

[2] Our expectations about the world are, to a big extent, culturally shaped (cf. Lin, Schwanenflugel, & Wisenbaker, 1990; Schwanenflugel & Rey, 1986). It might be the case that in other cultural contexts lizards are more typical pets than dogs, then we would reason accordingly.

which we have environmental pressure to decide or act, we will be willing to draw risky inferences disregarding their uncertain character.

This example highlights two important features of expectations: (1) it is a graded and subjective phenomenon, and (2) the strength of an expectation depends, to a significant extent, on the prototypical structure of concepts and the available information. To be more precise about the second point, when we receive some information about an object falling under some category, we tend to decrease our remaining uncertainty by implicitly assuming that the object also has the typical properties that are associated with the category in question. The effects of this will depend on how we represent the category and its semantic associates. As it will be explained in what follows, CS can give us some useful tool for articulating such relations.

The analysis of expectations proposed here focuses on the agent's representation of categories as associated with clusters of properties. As said before, the main idea is that when we categorize an object x as C, we implicitly form expectations about properties that this object is supposed to have because of falling under C. These possible properties will have different *strengths* according to how typical they are for objects falling under C. Within the framework of CS, we can use this principle to generate an ordering of expected properties by exploiting the prototypical structure of the space for C.

The underlying rationale for this method is the Gricean principle of *maximal informativeness* (Grice, 1975). If we are told that the object x is a bird, and that is all the information we have about it, we will expect that x has all the prototypical properties of birds —that it flies, sings, build nests, has feathers, and so on— because, according to this principle, your informant should have communicated something more specific if these expectations about x are not fulfilled.

Furthermore, when new information is added, expectations are restructured. If, after learning that x is a bird, we learn that it is an ostrich, you will no longer expect that it flies, nor that it sings. Instead, some new expectations will be added, such that x is big, it runs fast and kicks hard. Understanding how expectations are generated and organized when some information is received,

and how they are restructured when new information is added, are the two main aspects that a model of this phenomenon must account for.

6.2.1 CS-based expectation orderings

Let us now give some formal structure to these ideas. A concept M represented in a n-dimensional space $\mathscr{C}(M)$ is a convex sets of points representing possible objects falling under M. If M has a prototype, it is assumed that it corresponds to one of these points: a n-dimensional point $p^M =< p_1^M, p_2^M, \ldots, p_n^M >\in \mathscr{C}(M)$. The central idea is that our expectations are structured around this prototype. In other words, if the only thing we know about x is that it falls under concept M, we will expect it to be —close to— p^M, i.e., to have all the properties of the prototype.[3]

Now, our expectations towards the sentence "$M(x)$" go beyond the specific properties determined by the prototype.[4] They extend to all the possible properties that an object falling under M may have. In the conceptual space framework, this means that representing an object under a concept M implies that the object may occupy any possible position in $\mathscr{C}(M)$. Different positions imply different properties for the object. The properties that do not apply to p^M can be considered as secondary expectations, since they are weaker —more defeasible— than the ones that apply to p^M. In general, for any possible non-prototypical property in $\mathscr{C}(M)$, its *degree of defeasibility* will be a positive function of its distance from the prototype.

Now, the question is whether it is possible to construct an ordering of properties that reflects their "expectedness level" —and thus, their degree of defeasibility— according to their relative distance to the prototype. One way of doing this is by measuring the distance to the closest point where the property is not satisfied. We can use the distance function to obtain this kind of information from the conceptual space with the following criterion:

[3]Formally, if $p_i^M \in R_j \subseteq D_i$, then we expect that $x_i \in R_j \subseteq D_i$

[4]We use the same notation for talking about concepts and their respective lexical counterpart —predicates— in natural language.

Typicality Criterion (TC) 1. *Given domains D_i and D_k in a conceptual space $\mathscr{C}(M)$, for any two properties R_i, R_k, such that $R_i \subseteq D_i$ and $R_k \subseteq D_k$; R_i is more typical than R_k if and only if there is a point $x =< x_1, x_2, \ldots, x_i, \ldots, x_n >\in \mathscr{C}(M)$ with $x_i \in R_i$, and for all points $x' =< x'_1, x'_2, \ldots, x'_k, \ldots, x'_n >\in \mathscr{C}(M)$, $x'_k \in R_k$, it holds $d(x, p^M) < d(x', p^M)$*

The criterion measures the distance of a region to a prototype via its closest point.[5] It is important to note that in TC, we do not count numbers of instances, but the criterion is based on similarity to the prototype. In other words, this model is not probabilistic. Probabilistic models will not give the right results for expectation orderings since some property that is probable may be very atypical. [6]

Notice that TC not only allows us to compare non-prototypical properties; it also applies to properties inside the prototype. For instance, having feathers, building nests, and flying are all prototypical properties of *bird*. However, *flying* is the more defeasible of all three, since instances of birds that don't fly are more common than instances that don't build nests or have feathers. This means that the former two properties will have priority over the latter in an expectation order. In general, for all the properties in a conceptual space, TC will produce an expectation ordering $Exp(M) = \{R_1 > R_2 > \cdots > R_m\}$ when the following criterion is applied: given two properties R_i, R_k in $\mathscr{C}(M)$, R_i is more expected than R_k (i.e. $R_i > R_k$) iff R_i is more typical than R_k. This ordering of properties can be turned into an ordering of atomic sentences by saying that, for all individuals x, $R_i(x) > R_k(x)$ iff $Ri > R_k$. [7]

To see an example, consider the fruit space introduced in Chapter 4, with color, taste, shape, and texture as dimensions. If we are told that a is an apple, our maximal expectations will be that a has the properties of a prototypical

[5]Lewis and Lawry (2016) also use this distance measure for sets. There are, however, other possibilities to define the expectation ordering between properties, for example by using average distances between a prototype and a region. It is a matter of empirical research to determine which method gives the results that best fits with how humans reason.

[6]An example of such a model is Lieto and Pozzato (2019). We will return to a comparison with their model later.

[7]There are other possibilities to define the expectation ordering between properties, for example by using average distances between a prototype and a region. It is a matter of empirical research to determine which method gives the results that best fits with how humans reason.

apple like being red, sweet, round, and smooth. But, as we said above, these properties have different degrees of typicality even if they are all in the prototype. For instance, being sweet is more typical than being red for apples, since it is more surprising to find a non-sweet apple than a non-red one. This means that points representing non-red apples are going to be closer to the prototype than points representing sour or bitter apples in the conceptual space. Similarly, bitterness is a quite odd property for apples, certainly less expected than being yellow . Thus, instances of yellow apples are going to be closer than the prototype than instances of bitter apples. An expectations ordering of properties for Apple can thus look like this: $Exp(Apple) = \{round > red > sweet > soft > green > \cdots > yellow > \cdots > bitter > \ldots\}$.

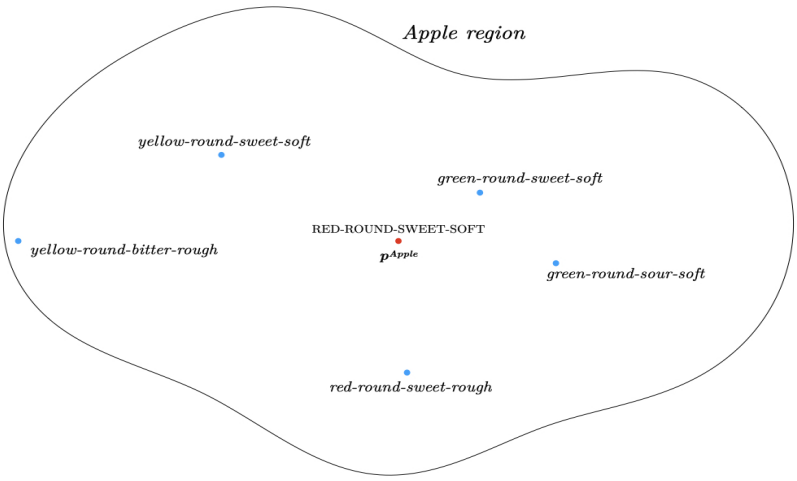


FIGURE 6.1: Illustration of an apple region of the fruit space, with points representing more and less typical instances of apples.

The typicality criterion produces a fine-grained order of expectations that makes it possible to compare individual properties. Thus, TC solves the problem of the origin of the expectation ordering that was mentioned above. We will make use of this advantage later in this chapter.

It should be noted that, since it is based on a distance function, the TC generates a total ordering of individual properties. In many cases, however, this assumption is cognitively unrealistic. For example, which is the most typical property of an ostrich: that it runs fast or that it kicks hard? In many cases, it

would be more natural to avoid judging which property is more typical. This would mean that the expectation ordering would be just a partial ordering of properties. Which ordering to use when analyzing expectation is essentially an empirical question. We will come back to this issue later.

6.3 Relations to Nonmonotonic Logic

Note that a partially ordered set of properties generated by the CS-model can be translated into an ordering of a subset of atomic sentences within a logical language. This allows to connect this treatment of expectations with the results on nonmonotonic logic in (Gärdenfors & Makinson, 1994). The central idea is that p *nonmonotonically entails* p (written $p \mid\!\sim q$) means that q follows from p, together with all the propositions that are "sufficiently well expected" in the light of p. According to the CS-model, these expectations would be those that are associated with the prototype.

As said before, there are many other relevant propositions that are not determined by the properties of the prototype, and can be part of the expectation set. A possible way idea to technically specify what "sufficiently well" means is to demand for any added sentence p to be more expected than $\neg p$ in the expectation ordering.

It can be proved that it follows from the postulates (E1)-(E3) below that this will be a consistent set and that it is a maximal set with this property (Gärdenfors & Makinson, 1994). Thanks to this, we make sure that the set of extra premises added as expectations will not contradict p and that as many as possible of the expectations are included as extra premises. The following definition builds on these ideas:

Definition 1. *$C \mid\!\sim$ is an expectation inference relation iff there is an ordering $\leq$ such that the following condition holds: $p \mid\!\sim q$ iff $q \in Cn(\{p\} \cup \{r : \neg p < r\})$*

Here, Cn denotes the set of logical consequences of the premises in the set.

Gärdenfors and Makinson assume that the expectation ordering $\leq$ —which they call an ordering of *epistemic entrenchment*— satisfies the following properties:

(E1) If $p \leq q$ and $q \leq r$, then $p \leq r$ (Transitivity)

(E2) If $p \vdash q$, then $p \leq q$ (Dominance)

(E3) For any p and q, $p \leq p \wedge q$ *or* $q \leq p \wedge q$ (Conjunctiveness)

(E2) means that if q is a logical consequence of p, then q is always more expected than p. (E1)–(E3) entails that $\leq$ is a total ordering, that is, either $p \leq q$ or $q \leq p$.

Gärdenfors and Makinson used these postulates and the above definition to demonstrate that the nonmonotonic entailment relation $\mid\!\sim$ satisfies the following postulates:

Superclassicality: If $p \vdash q$, then $p \mid\!\sim q$.

Right Weakening: If $\vdash q \rightarrow r$ and $p \mid\!\sim q$, then $p \mid\!\sim r$.

And: If $p \mid\!\sim q$ and $p \mid\!\sim r$, then $p \mid\!\sim q \wedge r$.

Weak Conditionalization: If $p \mid\!\sim q$, then $\mid\!\sim p \rightarrow q$.

Weak Rational Monotony: If p is logically consistent and $\mid\!\sim p \rightarrow q$, then $p \mid\!\sim q$.

Consistency Preservation: If $p \mid\!\sim \bot$, then $p \vdash \bot$.

Cumulativity: If $p \mid\!\sim q$ and $q \vdash p$, then $p \mid\!\sim r$ iff $q \mid\!\sim r$.

Or: If $p \mid\!\sim r$ and $q \mid\!\sim r$, then $p \vee q \mid\!\sim r$.

Rational Monotony: If p does not contradict q and $p \mid\!\sim r$, then $p \wedge q \mid\!\sim r$.

Supraclassicality means that the nonmonotonic inference relation is an extension of the classical inference relation where only explicit premises are assumed.

Nevertheless, the expectations expressed by the underlying ordering generates an inference relation that extends the classical one.

An important thing to mention is that if we assume that the expectation ordering is partial and not total, then the Rational Monotony is no longer satisfied. However, the inference relation $\mid\sim$ has many of the properties that are desired for rational inference.

From a logical point of view, an agent would draw more conclusions based on expectations if the underlying ordering is total. However, this assumption is psychologically unrealistic. The agent may lack sufficient knowledge to compare two partial vectors of information about an object. For example, how can I compare the expectations about an apple's size to expectations about its color? If two domains in a conceptual space are largely independent, it may not be easy to compare values in one domain to those of the other. Consequently, the resulting expectation ordering will only be partial.

6.4 Criteria for updating expectations

Expectation orderings are context-specific. As such, they are dynamic structures that change according to the information available. A central problem for the study of expectation-based inference is understanding the principles according to which these orderings are updated when new information is added. In what follows, it will be shown how the CS approach may bring some light on this issue. In general, different kinds of information, for example, perceptual information, may generate different types of updates. Here, however, the analysis will be restricted to updates produced by information already present in the ordering generated by TC.

First of all, notice that only information that adds specificity to a previous informational state will change the expectation orderings. Trivially, suppose we are told that b is a dog and later that b is a mammal. In that case, the expectation ordering remains the same since the new information is already

implicit in the initial informational state.[8] According to the framework defended in this chapter, reasoning under uncertainty about objects' properties amounts to specifying a point's position in an n-dimensional space. Each new piece of information with a more specific value in one of the dimensions reduces the space of uncertainty.

Prototype shifts

We start by analyzing a kind of updating named "prototype shift". The idea is that given an expectation ordering $Exp(C)$, with p^C as maximal point, if we are told that x is also G, and we know that G is a subordinate concept of C ($G \subset C$), then $Exp(C)$ will be updated into $Exp(C\&G)$, changing its maximal point to the prototype of G, and being re-structured accordingly (following the two criteria explained in section 6.2.1).

Notice that $Exp(M)$ and $Exp(G)$ include properties from the same number of dimensions, since subordinate concepts inherit the dimensionality of their superordinate. However, the regions in each of the domains may be reduced substantially, because adding information that specifies the properties of an objects "shrinks" the initial conceptual space. In other words, $Exp(M)$ will include elements which are not going to be in $Exp(G)$.

To see an example, suppose that we are told that a is a bird. An expectation set will be generated with p^{Bird} as a maximal point. $Exp(BIRD)$ will contain a large number of color properties and shapes (the bird category has a large color and shape variability). If we are then told that a is a penguin, the updated expectation ordering $Exp(BIRD\&PENGUIN)$ will change the maximal to the prototype for penguins; and it will lose all the elements referring to colors which are not black and white; as well as the elements referring to not penguin-like shapes. Also, it will lose the property of flying, in the dimension encoding information about locomotion.

Updating via properties

[8] Note that this correspond to a core material inference, and as such it is supposed to be certain.

A less dramatic kind of update happens when information about specific properties is added. Given an expectation ordering $Exp(M)$, if we get to know that an entity a has property F, and F is a property in $Exp(M)$, a specific region of one domain D_i in $\mathscr{C}(M)$, then the ordering $Exp(M\&F)$ will be equal to $Exp(M)$ minus all the properties that are incompatible with F in D_i. If D_i is partitioned into disjoint regions, then $Exp(M\&F)$ will not include as properties all the other sub-regions of D_i which are not F. If this is not the case, then $Exp(M\&F)$ might lose or may not lose properties from $Exp(M)$, depending on the structure of D_i.

Consider the simple example in which we start with the information $Man(a)$ and we get to know later also $Bachelor(a)$. Then $Exp(MAN)$ will shrink by losing the property *married* (previously in $Exp(MAN)$. However, if we later get to know that a plays golf as a hobby, then we cannot delete from $Exp(MAN\&BACHELOR)$ all the other properties corresponding to the hobby dimension: since this domain is not disjoint, it might be the case that a man has several hobbies. Thus, for properties corresponding to not-disjoint dimensions, the updating will depend in the particular structure of the domain in question.

Updating via properties and correlations

A more complex case of *updating via properties* happens when the new property is correlated to another specific property in other domain. For instance, as said before, in the fruit space, dimensions like ripeness and texture, or color and taste, are strongly correlated. Apples, in general, are expected to be sweet, but green apples are expected to be sour, since these two properties are correlated.

For these kinds of cases, the updating procedure will be the following. Given $Exp(M)$ and two correlated properties G and F from domains D_i and D_k in $\mathscr{C}(M)$. When new information G is added, we will have that:

(i) $Exp(M)$ will be updated via property G in the previously explained way;

(ii) $Exp(M\&G)$ will have F as maximal or close to the maximal;

(iii) For any other $H \subseteq D_k$, F will be more expected than H in $Exp(M\&G)$.

To continue with the apple example, suppose that our initial $Exp(APPLE) = \{round > red > sweet > soft > green > \cdots > yellow > \cdots > bitter > \ldots\}$. If I am then told that apple a is green, properties like red and yellow, will be deleted, while sour will approximate the maximal and be more expected than any other taste-related property. The updated set would look like this: $Exp(APPLE\&GREEN) = \{green > round > sour > soft > sweet > \cdots > bitter > \ldots\}$.

6.5 Defaults

6.5.1 Generating default rules

As explained in previous chapters, everyday reasoning rarely follows a set of formal deductive rules. Instead, it works by exploiting semantic information from our conceptual system. A formalization of this idea was developed by default logics, through the notion of *default rule* (see Brewka, Roelofsen, & Serafini, 2007; Delgrande, 2011; Horty, 2012). Default rules are predicate-specific inference rules. They are meant to capture facts that are normally or typically the case when something is claimed to fall under some predicate X. For instance, the fact that most mammals have fur can be captured by the default $Mammal(x) \rightarrow HaveFur(x)$. Defaults like this can be used as rules of inference in nonmonotonic logical systems to extend derivations.

A foundational problem of default logics concerns the interpretation of the notion of a *default rule* (Delgrande, 2011). Two epistemological issues affect this notion: (1) it is not clear where defaults come from; and (2) before multiple defaults that can be applied to the same object, a decision method has to be applied to determine which defaults have priority over the others (Horty, 2012, p.19). A well-known example is the "Nixon diamond": By default, quakers are pacifists and, by default, republicans are not pacifist, but Richard Nixon is both a quaker and a republican. Thus, when determining to conclude whether Nixon is a pacifist or not, one must violate one of the default rules. Gärdenfors and Makinson (1994, Section 3.3) argued that expectation orderings offer a natural

way of representing defaults. In what follows, this idea will be developed and adapted for the CS framework.

A natural answer to (1) is that defaults express strong regularities about phenomena that we can use as "inference tickets" in reasoning. However, notice that if we focus on cognitive modeling, we should prioritize an internalist interpretation: the idea that defaults express close conceptual relationships codified in our background knowledge. In this sense, the question about the origins of defaults becomes a question about how they emerge from the structure of this background knowledge. As suggested by Gärdenfors (2000, p. 117), the conceptual spaces framework offers an answer to this question. Again, when we categorize an object under a concept, we automatically represent it as having a cluster of properties associated with the concept's prototype. For instance, categorizing something as a fruit, implies also categorizing it as sweet. This feature about categorization can also explain the kind of conceptual transitions that are expressed by default rules. An answer to the problem about the origins of defaults can be found, then, in the explanation that conceptual spaces offer about the prototypical structure of concepts and its exploitation in inference and categorization.

Let's turn now to (2). Notice that a default rule of the form "$F's$ are normally $G's$" can be expressed in terms of expectations as "if something is an F then it is less expected that it is non-G than that it is a G". This can be represented in an expectation ordering as $F(b) \rightarrow \neg G(b) < F(b) \rightarrow G(b)$ for all b. Since the consequence relation $\alpha \mid\sim \beta$ is equivalent to $\alpha \vdash \beta$ *or* $\alpha \rightarrow \neg\beta < \alpha \rightarrow \beta$ (Gärdenfors & Makinson, 1994, p. 124), the fact that $F(b) \mid\sim G(b)$ can be directly proved. This means that an expectation-based inference relation is a nonmonotonic relation with *embedded* defaults. Thus, we don't need any explicit representation of them.

Furthermore, TC gives us information about the comparative strength of each of these defaults. For instance, for the concept *bird*, it is to be expected that (according to TC) the property "build nests" has priority over "fly". This means that the default rule $Bird(x) \rightarrow Fly(x)$ will be weaker than $Bird(x) \rightarrow BuildNest(x)$. In general, the more entrenched a property is in the conceptual

space-generated ordering, the stronger its correspondent default rule will be.

The discussion above shows that the framework introduced in this chapter provide new insights into the two fundamental epistemological problems of default logic, and it suggests constructive solutions to them.

6.5.2 Typicality and the conjunction fallacy

As explained above, one of the central aims of default logics is to capture our use of typical predicate-relations in non-monotonic reasoning. In recent years, various formal models have been developed in order to tackle in a more systematic way the role of typicality in reasoning. For instance, Propositional Typicality Logic (Booth, Meyer, & Varzinczak, 2013) uses a classical propositional language plus a typicality operator for specifying the set of typical situations in which some formula holds. Description logics have been also used for modeling this phenomenon (see Giordano, Gliozzi, Olivetti, & Pozzato, 2008). In particular, Lieto and Pozzato (2019; 2018) developed a rich formal framework that uses description logic to account for typicality in non-monotonic reasoning and conceptual combination.

The model presented here is a natural framework for accounting for the role of typicality in reasoning. No systematic explanation of this will be offered now; but an idea of how this may work will be advanced by an expectation-based explanation of the conjunction fallacy.

As is well known, Tversky and Kahneman (1974; 1983) have shown that intuitive judgments of probability do not mirror the principles of standard probability calculus. They argued that in many cases, people violate these principles because they prioritize intuitive heuristics that exploit typicality relations for estimating the probability that and object has a specific property. Their claims are supported by a famous experimental case called the "Linda problem." In brief, subjects were given the following information: Linda is 31 years old, single, outspoken, and very bright. She has a major degree in philosophy, and while studying, she participated in anti-nuclear demonstrations and was involved with

discrimination and social justice issues. Then subjects were asked which of the following statement is more probable:

(i) Linda is a bank teller.

(ii) Linda is a feminist bank teller.

A large majority of the subjects answered that sentence (ii) is more probable, although by the laws of probability (i) is at least as probable as (ii).

The issue is that ordinary people do not interpret the word "probable" in the same way as people educated in statistics. The paradox disappears when formulated in frequentist terms (see Cosmides & Tooby, 1996). Our Gricean analysis coheres with Tversky and Kahneman's idea that elements of typicality come into play when making the judgment. According to our model, when people are given the initial information, they will represent Linda in a multidimensional "person space". In that space the prototype for *feminist bank teller* is located much closer to the prototype of *feminist*, than to the prototype of *bank teller*. This is mainly because the properties described initially are positively correlated to the feminist property and negatively correlated with the bank teller property. Consequently, feminist will appear much higher in the expectation ordering than a bank teller, making people highly disposed to use it in inference.

Lewis and Lawry (2016) and Lieto and Pozzato (2019) also analyze the conjunction fallacy with the ais of their respective models of concept combination, so it is interesting to compare with their solutions. The goal of the model presented by Lewis and Lawry (2016) is to represent hierarchies of concepts. Their model is similar to ours in that they also use a geometrical approach based on conceptual spaces. Instead of using Voronoi tessellations to determine category membership, they use random set theory. Nevertheless, since their explanation of the conjunction fallacy is also based on distances to prototypes, it is similar in spirit to the one advanced here.

The description logic presented by Lieto and Pozzato (2019) is less similar since their representations are probabilistic rather than geometric. Therefore, it is difficult to compare their account of the conjunction fallacy to the solution

we present. There are, nevertheless, interesting similarities between the models: The logic for their typicality operator **T** can be shown to be equivalent to the nonmonotonic logic presented in Section 4 of this paper. [9]

6.5.3 *Inferential strength*

When an inference is based on defeasible premises, a criterion for judging the strength of the inference is helpful. Deductive inferences, for instance, are often considered as maximally strong since they preserve truth. However, in everyday reasoning, they can be analyzed as having different strengths according to the "quality" of the premises' information. In other words, the structural properties of a deductive inference do not prevent it from being weak from the perspective of everyday reasoning.

The strength of an inductive inference, on the other hand, may depend on the size of the sample —or on the probability— on the premises. Other forms of inductive inference, like category-based induction, depend on a more complex combination of factors related to the exploitation of different conceptual relations —as we will se in the following chapter.

Remember that an expectation ordering is a kind of epistemic entrenchment ordering. The notion of epistemic entrenchment comes from belief revision, and it was meant to capture the idea that within a belief system, certain beliefs are more susceptible to be revised or deleted than others. In an expectation ordering, the position of a belief in the ordering gives information about its comparative degree of defeasibility (Gärdenfors & Makinson, 1994, p. 209). In this chapter, TC was suggested as a criterion for measuring such a degree of defeasibility for properties: The more typical a property is, the less defeasible it is.

Let us go back to the pet example. Suppose I am told that John bought a new pet. I could infer that it flies if I use as an implicit premise the belief from my set of expectations that the pet in question is a bird. However, this

[9] Lieto and Pozzato (2019) remark that their logic is a reformation of Lehmann and Magidor's (1992) rational logic. Gärdenfors and Makinson (1994) show that this locic is equivalent to the logic presented in Section 6.3 above.

inference would be quite weak, since *bird* would typically be less entrenched in $Exp(PET)$ than concepts like *dog* or *cat*. If instead of *bird*, I use as an implicit premise that John's new pet is a dog or a cat, then I would nonmonotonically infer that it does not fly, and this inference would be clearly stronger than the other one.

The above example shows that when we reason nonmonotonically, the strength of our inferences depends on the choice of the information from the set of expectations that we use as implicit premises. In more general terms, we can say that the strength of an inference is a negative function of the degree of defeasibility of the propositions that are used as implicit premises. As we mentioned before, our framework allows for an explication of this criterion by stating that the information to be used is always the one present in the maximal point of the set of expectations, i.e. the prototypical information of the concept in question or the one that results from an updating process previously defined.

As explained in Section 6.2, even among the sufficiently well-expected information in the set of expectations, some propositions are less defeasible than others. As a result, they will produce stronger inferences. For a default logic system, a comparative notion of inferential strength can be defined as a positive function of the strength of the default rules used in specific inferences. As we showed earlier, this strength measure comes from the rule's position in the conceptual space expectation ordering, which depends ultimately on the distance measure defined in TC. To give a simple example, an inference that uses the default $Bird(x) \rightarrow HaveFeathers(x)$ will be stronger than one that uses the default $Bird(x) \rightarrow Fly(x)$ since following TC, $HaveFeathers(x)$ will be more entrenched in the expectation order than $Fly(x)$ even if both pertain to the prototype of *bird*.

Finally, another important issue in default logic concerns the possibility of having defaults expressing conflicting information (Reiter & Criscuolo, 1981; Touretzky, 1984). The most famous example of this situation is the "penguin principle" (Lascarides & Asher, 1993): if we are told that x is a bird and that it is also a penguin, then we will have at our disposal the defaults $Bird(x) \rightarrow Fly(x)$ and $Penguin(x) \rightarrow \neg Fly(x)$. If we don't have a clear criterion to choose which

default has priority, we will end up addressing a contradiction. Above, we saw the "Nixon diamond" where two conflicting defaults seeming lead to the conclusion that Nixon is both a pacifist and not a pacifist. The TC criterion gives a way of resolving the conflict since once it is decided whether Nixon is more atypical as a quaker than as a republican, the resulting expectation ordering determines which of the default rules should yield.

6.6 Conclusions

Despite the significant progress made in the field of nonmonotonic reasoning during the past decades, the available logical models still suffer from various epistemological issues. In particular, several formal systems make strong assumptions about the role of background knowledge and entrenchment relations of beliefs without providing any cognitive or epistemological argument for them. In this chapter it was argued that some of these problems are due to the intrinsic limitations of propositional-based models for capturing the internal structure of conceptual knowledge.

It was shown that combining the CS framework with an expectation-based analysis of nonmonotonic inference is a fruitful way of extending the modelling tools of these logical systems while enriching their theoretical foundations. Furthermore, this analysis implies a "construal" of nonmonotonic inference as a kind of concept-based inference falling under the definition advanced in Chapter 3: a transition between two mental/informational states exploiting properties of a representational system which codifies conceptual information. The model presented here assumes that this representational system is CS, with its typical geometrical properties.

Above, the models suggested by Lieto and Pozzato (2019) have been discussed, and there are other attempts to computationally implement reasoning with default rules (e.g. Brewka et al., 2007; Delgrande & Schaub, 2000). However, in these systems the default rules must be provided by the user and they do not generalize well. Here it was shown that a CS approach to modelling reasoning with default, and nonmonotonic reasoning in general, is a better method

when attempting to build artificial systems with theses capacities. Once the domains with their concept's regions and distance functions have been implemented, then the TC principle provides a direct way to generate expectation orderings and default rules, and thereby a method to calculate nonmonotonic inferences. Implementations, however, still remain to be constructed.

The ideas defended in this chapter are directly related to a kind of reasoning known in psychology as category-based induction. This mechanism consists in judging the strength of arguments that project some property from one or more categories to another category, exploiting conceptual similarities (Feeney, 2017). In the next chapter, a model of this phenomenon based on distances on conceptual spaces will be presented. Explicating both nonmonotonic reasoning and category-based induction with the same modeling framework could be of great value for the various disciplines studying these phenomena, and help to integrate theories of reasoning with theories of concepts.

Finally, one limitation of the research presented here is that the expectations we can model concern only object predication. However, expectations are pervasive in cognition, and a full model of their role in reasoning should also account for more complex lexical items, like verbs or relational and functional predicates. In Gärdenfors (Gärdenfors, 2014, 2020; Gärdenfors & Warglien, 2012), the conceptual space framework has been extended for modeling expectations based on the structure of events, in particular causal inferences. Future work on expectation-based nonmonotonic reasoning will hopefully result in a more general model.

Chapter 7

Category-based induction in conceptual spaces

Category-based induction is an inferential mechanism that uses knowledge of conceptual relations to estimate how likely is for a property to be projected from one category to another. During the last decades, psychologists have identified several features of this mechanism, and they have proposed different formal models of it. In this chapter, a new mathematical model based on distances in conceptual spaces is proposed. It will be argued that this CS model can predict most of the properties of category-based induction, make some new predictions, and provide a solid theoretical foundation for this psychological phenomenon. At the end of the chapter, the relations with other models will be discussed, as well as some methodological considerations. [1]

7.1 Induction and conceptual relationships

Throughout this thesis, the formalist approach has been largely criticized, both for its philosophical commitments and for its inability to explain different forms of concept-based inferences. Formalists generally take deductive reasoning as a paradigm of rational inference, while they see induction as an important but elusive phenomenon that seems to resist formalization within logical frameworks

[1] This chapter is based on Osta-Vélez, M. and Gärdenfors, P. (2020) "Category-based induction in conceptual spaces", *Journal of Mathematical Psychology*, 96, 102357.

(e.g., see Norton, 2003, 2010). The problem is that induction and deduction deal with semantic information in very different ways (see Johnson-Laird, 1983). While deductions are *safe* inferences because they are *informationally conservative*, in inductive reasoning the conclusion of an inference adds semantic information that is not present in the premises in order to tackle uncertainty. In this sense, inductive reasoning is clearly cannot be "formal" in the traditional sense (see also Thagard, 1984, pp. 27-29).

Induction —as a type of personal-level inference—, is a semantic-based mechanism. Understanding it, therefore, requires explaining how background knowledge is exploited in reasoning. Again, this can only be done if we conceive reasoning as an activity which is not strictly propositionally-based, but that combines information codified at the symbolic-propositional level with information encoded at the conceptual level (see Chapter 3).

During the last decades, psychologists have been studying a type of inferential mechanism directly related to this last point. In his pioneering article "Inductive judgments about natural categories," Lance Rips (1975) analyzed a peculiar kind of inductive inference that exploits information about individual categories (and about relations among categories) for estimating the probability of property projection among them. For instance, the inference "*Dogs* have sesamoid bones; thus, *wolves* have sesamoid bones" relies on the conceptual similarities among the categories *dog* and *wolf*, and not on the logical form of the argument or some other propositionally-codified property. Such processes, called category-based inferences (CBI), are intuitive forms of reasoning fundamental to our cognitive lives. On the one hand, they are crucial for dealing with uncertainty: they allow us to reason about some unknown input X by exploiting information stored in our conceptual system about things that resemble X. On the other hand, as Feeney observes (2017, p. 167), they are a clear example of how concepts make our cognition efficient.

Understanding how CBI works, and which properties of our conceptual systems it exploits, can shed light on the general problem of this dissertation. In this chapter, the general features of CBI are discussed, and the CS-model, developed in the previous chapters, is extended to model them. The chapter is

organized as follows. Section 7.2 presents a basic taxonomy of category-based inductions and reviews the central phenomena associated with them. Section 7.3 introduces some of the theoretical and technical aspects of conceptual spaces. Section 7.4 presents the model and explains how the theory of conceptual spaces provides a natural way of modeling CBI's central properties based on the capacity of the theory for representing similarity and typicality relations among categories. Section 7.5 compares the CS model to some of the previous explanations, and Section 7.6 briefly considers some methodological aspects of the approach defended here.

7.2 Category-based induction

7.2.1 The general structure of category-based inferences

Rips' seminal paper (1975) intended to understand the strategies that agents use for reasoning under uncertainty about property projection among biological kinds —such as *hawk*, *bird*, *eagle*, etc. He showed that subjects exploit structural properties of categories for estimating the plausibility of property projection. In particular, Rips saw that similarity among categories, as well as the degree of typicality of premise-categories, were guiding principles of this kind of reasoning.

Most studies on CBI follow Rips' analysis (see for example Carey, 1985; Heit, 1997; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993). They all assume that inductive reasoning is a process that exploits information at the conceptual level, and not at the propositional level. From a methodological perspective, these studies analyze cognitive agents' inferential dispositions by inspecting how they judge the strength of different types of inductive arguments. Various empirical properties of CBI have been identified during the last decades (see Feeney, 2017; B. Hayes, Heit, & Swendsen, 2010, for reviews). Before explaining them, let's see a basic taxonomy of CBI judgments that will help to organize the following analysis.

Category-based inferences are structured as arguments with one or more premises of the form 'X are S' —where X is a category and S a property—,

and a conclusion of the same type with a different category. A typical example would be "Dogs have sesamoid bones; thus wolves have sesamoid bones". In what follows, these arguments will often be abbreviated as $X \rightarrow Y$. This is because, in most studies, subjects have little or no knowledge about the property S, and therefore, it does not influence the strength of the argument.

CBI can be classified in two major ways: according to their number of premises; and according to whether the conclusion is at the same conceptual level as the premises or in some superordinate category. When the premise(s) and conclusion categories are at the same conceptual level, the argument is called "specific; " when the argument involves a generalization —a "jump" to a superordinate conceptual level—, then it's called "general." For instance, arguments with the form *robin* $\rightarrow$ *crow* or *table* $\rightarrow$ *chair* are specific, while arguments like *robin* $\rightarrow$ *bird*, *robin* $\rightarrow$ *animal* or *table* $\rightarrow$ *furniture*, are general. Both specific and general arguments can be composed of one or multiple premises (see Figure 7.1).

	Single Premises	**Multiple Premises**
Specific	(1) Foxes have property S Wolves have property S	(2) Penguins have property S Pigeons have property S Ostriches have property S Sparrows have property S
General	(3) Robins have property S Birds have property S	(4) Polar bears have property S Grizzly bears have property S Bears have property S

FIGURE 7.1: Basic taxonomy of category-based inferences.

In what follows, the main properties of CBI will be reviewed as described by the empirical literature. The idea is that these phenomena say something about what kind of categorical or conceptual relations people exploit when judging category-based inductive arguments.

7.2.2 Premise-conclusion similarity

The primary categorical relation guiding category-based inferences is similarity. Similarity has been considered as a crucial criterion for induction since —at least— Hume (1894). Quine famously argued (1969b; 1974) that similarity might be a fundamental psychological principle in a wide range of cognitive phenomena, like learning, concept formation, and reasoning. In psychology, the notion of similarity has proved to be fruitful since the 1970s. Since the pioneering work of Shepard (1987) and Tversky (1977), formal models of similarity have been developed for explaining concept formation, categorization, and even induction. And since Rosch's work on prototyes (1973; 1983), similarity has been taken as the central criterion for explaining category structure.

Not surprisingly, the empirical literature has shown that the most robust criterion used in CBI is similarity among categories (Carey, 1985; López, Gelman, Gutheil, & Smith, 1992; Osherson et al., 1990; Rips, 1975). This can be formulated by saying that our expectations regarding property projection among two categories X and Y is a positive function of their similarity. For instance, arguments like "Ostriches are S, then emus are S" are generally seen as stronger than arguments like "Ostriches are S, then blue jays are S", since $sim(ostrich, emu) > sim(ostrich, bluejay)$, where $sim(X, Y)$ denotes a measure of the similarity between the categories X and Y.

7.2.3 Typicality

Similarity, as a criterion for category-based inferences, can only be used for categories at the same conceptual level. Still, it is not useful in arguments that generalize a property from the category premise to the conclusion category. In those cases, typicality is what guides categorical inferences. [2]

As explained earlier in this work, the typicality effect is the finding that individuals respond more quickly to typical examples of a category than they do to cases that are considered atypical. For instance, when asked to name a bird, an individual is much more likely to respond with "robin" than with

[2]Typicality is deeply related to similarity (see Hampton, 2001; Rips, 1989).

"penguin". The idea was proposed and tested by Rosch (1973), and it suggests that conceptual structures —especially natural kinds— are articulated around prototypes. Most categories seem to have a graded structure (see Barsalou, 1987; Decock & Douven, 2014), which means that different members of the category are perceived with varying degrees of typicality. For instance, cows are generally seen as very typical representatives of the category *mammal*, while mice are moderately typical, and whales are highly atypical members.

Typicality plays a crucial role in CBI. [3] The most robust effect found in the empirical literature is that expectations of property projection are a positive function of premise-typicality. For instance, the inference "Robins have enzyme E; thus ostriches have enzyme E" is often judged as stronger than "Penguins have enzyme E; thus, ostriches have enzyme E." This is because robins are prototypical birds, and as such, they better represent the category than penguins —which are atypical. To a lesser extent, conclusion-typicality also seems to be a factor in category-based inferences. Hampton and Cannon (2004) have shown that arguments with prototypical conclusion-categories —like *chicken* → *robin*— are judged as stronger than arguments with non-typical conclusion categories —like *chicken* → *vulture*.

Furthermore, the typicality effect produces *asymmetry*, that is, the fact that switching the categories from the premises and conclusion often changes the expectations of property projection, according to the degree of typicality of the category in the premise(s). For instance, arguments like "Cows have enzyme E; thus, otters have enzyme E" is considered stronger than arguments like "Otters have enzyme E; thus, cows have enzyme E" since cows are more typical mammals than otters.

7.2.4 Conclusion homogeneity and premise diversity

Another important aspect is that agents assume a common superordinate category of the premises when making inferences or judging this kind of argument's

[3] For a discussion on the role of prototypicality in reasoning in general —besides the discussion in the previous chapter— see Cherniak (1984).

strength. Sometimes this superordinate category appears explicitly in the conclusion —as in general arguments—; some other times, it is just considered implicitly. For instance, consider the arguments in Figure 7.1. In (1), the implicit superordinate category is *mammal*, while in (2) it is *bird*. In (3) and (4), the superordinate category appears explicitly in the conclusion. Four important phenomena related to such an evoked superordinate category have been studied in the empirical literature: homogeneity, monotonicity, nonmonotonicity, and premise diversity.

(i) *Homogeneity* refers to the idea that the more abstract and less homogenous the category in the conclusion is, the weaker the argument. For instance, arguments like "Robins are S and blue jays are S; thus, all birds are S" are judged as stronger than "Robins are S, and blue jays are S; thus all animals are S". This is not surprising at all. As we said before, when evaluating arguments or making inferences that involve generalizations, we deal with different degrees of uncertainty. The more abstract the category in the conclusion is, the more information we need from the premises to cover it. Hence CBI with abstract categories (like *animal* or *living being*) involve higher degrees of uncertainty and are more difficult to cover by the information from the premises.

Studies of categorization —especially in the prototype tradition— provide some insight into this phenomenon. Categories may have different degrees of generality —e.g. *dog*, *mammal*, *animal*, *living*, *thing*—, and these degrees are related to the computational cost of using them in categorization. Categories with an intermediate level of specificity are preferred in terms of cognitive efficiency. These are called "basic-level categories" —*dog* instead of *mammal*; *chair* instead of *furniture*—, and studies have shown that they are central for carrying out several cognitive tasks (Mervis & Rosch, 1981). Inductive inference seems to follow the same principle. We have a preferred level of induction (Sloman & Lagnado, 2005, p. 106) that coincides with basic-level categories.

A possible way of explaining this is by referring to similarity and typicality as the two main criteria for using categories. Basic level categories are more homogeneous. As such, it is easier for us to apply criterion of similarity among their members. Abstract categories are more diverse and less homogeneous, so

comparing their members in terms of similarity is more complex —for instance, the category *animal* includes highly dissimilar subcategories, such as *elephant* and *starfish*. Along the same line, basic categories have clear prototypes, while it is complicated for us to construct prototypes for abstract categories (see Ungerer & Schmid, 2006, Ch. 2 for an introductory explanation). In this sense, typicality, considered as a criterion for using categories, is stronger in basic-level categories than in abstract ones.

(ii) *Monotonicity* refers to the fact that the addition of premises, as long as their categories are included in the evoked superordinate category, strengthen the argument (Osherson et al., 1990). For instance, an argument of the form $(robin\&hawk) \rightarrow bird$ is weaker that an argument of the form$(robin\&hawk\&pigeon) \rightarrow bird$. However, if we add to the premises a category that is not from the evoked superordinate category, then the argument becomes weaker. This is called nonmonotonicity (iii). For instance, an argument with the categories $(peacock\&crow) \rightarrow bird$ —or $(peacock\&crow) \rightarrow pigeon$— is stronger than an argument that goes from $(peacock\&crow\&rabbit) \rightarrow animal$ —or $(peacock\&crow\&rabbit) \rightarrow pigeon$—.

(iv) Finally, there is the *diversity* phenomenon (Feeney & Heit, 2011; Osherson et al., 1990). Empirical studies have shown that having diverse categories tends to promote expectations regarding property projections. For instance, arguments like "Horses have an ulnar artery and seals have an ulnar artery; thus, all mammals have an ulnar artery" are considered as stronger than the argument "Horses have an ulnar artery, and cows have an ulnar artery; thus, all mammals have an ulnar artery." The less similar the categories in the premises are, the stronger the argument tends to be.

An interesting way of understanding this phenomenon builds on the idea of "category coverage" (Osherson et al., 1990). As we mentioned before, when performing or evaluating categorical inductions, we take as a reference —implicitly or explicitly, according to whether we deal with a specific or general argument— some superordinate category that includes all the categories in the premises. The strength of the argument will depend, to some extent, upon how the categories in the premises cover this superordinate category. For instance, similar

categories like *horse* and *cow* have less coverage of the superordinate category than dissimilar categories like *horse* and *seal*. In this sense, coverage can be described in terms of similarity. We will discuss this idea further in Section 5.

Sloman (1993, pp. 253-254) pointed out that diversity has a limit: if highly dissimilar categories are used in the premises, this can weaken the argument instead of making it stronger. For instance, the argument "German shepherds have sesamoid bones and elephants have sesamoid bones; thus, moles have sesamoid bones" seems stronger than "German shepherds have sesamoid bones, and blue whales have sesamoid bones; thus, moles have sesamoid bones". This indicates that in arguments that include highly atypical categories in some premise —*blue whale* is a highly atypical mammal—, then diversity becomes negative regarding argument strength.

CBI has been analyzed from different perspectives, depending on what kind of categorical relation it is assumed to explicate. Class-inclusion (Inhelder & Piaget, 1964), shared features (Sloman, 1998b), and similarity (Osherson et al., 1990), have been the most explored ones in the literature. However, reasoning about categories seems to be a complex mechanism that involves combining all these relations and probably other sophisticated heuristics. It is a challenge to present a model that can account for all of them. In the following section, a general model of CBI based on conceptual spaces is introduced. Among other things, this model will offer a natural —and relatively simple— way of explaining similarity and typicality, which are the two main categorical relations in CBI.

This model also uses the notion of "expectation." In particular, we talks about "expectations" of property projection among categories instead of *argument strength*. As we saw in the previous chapter, expectations play a crucial role in everyday reasoning. The sentence "John got a new pet" comes associated with a large set of expectations related to the lexical concepts in the sentence. In relation to CBI, the idea is that the agent's inferential dispositions to project a property from one category to another are also determined —to a large extent— by her expectations about regularities in the world, which are codified in the agent's background knowledge (cf., Section 3.4, and Section

5.2.1, this work).

In this sense, the expression $ExpS(X \rightarrow Y)_Z$ will be used as standing for the agent's expectations that the property S is projected from category X to category Y, with Z as the lowest-level superordinate category which contains both X and Y. Let us start the analysis by focusing on the simplest case of category-based inference: single premises/specific arguments. For this kind of inductive inference, we need $ExpS(X \rightarrow Y)_Z$ to satisfy the following criteria:

1. It is positively correlated with $sim(X, Y)$.
2. It is positively correlated with $sim(X, p^Z)$, where p^Z is the prototype of Z.
3. It is positively correlated with $sim(Y, p^Z)$.

The rationale for the first condition is that the more similar the categories X and Y are, the more expected will it be that Y has property S if X has it. Regarding condition (2), the intuition is that the more prototypical X is, the more expected it is that another category Y has property S, given that X has it. Condition (3) is motivated by Hampton and Cannon' (2004) conclusion-typicality: the more prototypical Y is, the more expected it is that Y has property S if X has it.

7.3 A conceptual space-model

7.3.1 A simple model

To illustrate the basic idea of the approach defended here, let's start with a simple model. For the time being, let us assume that X and Y are small regions so that we can identify them with points in a conceptual space. Then, given a conceptual space representing the categories X, Y, and Z and a distance function d of the space, we can account for the three conditions above by the following equation:

$$ExpS(X \rightarrow Y)_Z = (d(X, Y).d(X, p^Z)^a.d(Y, p^Z)^b)^{-1} \tag{7.1}$$

Where a and b are positive constants such that $a > b$. This assumption expresses that premise typicality contributes more to expectations than conclusion-typicality since, according to the literature, the former is a more prevalent phenomenon than the latter. The values of both a and b must be empirically determined from data about CBI judgments.

Now, following Shepard's universal law of generalization (Shepard, 1987), which claims that similarity is an exponentially decreasing function of distance, we can take the logarithm of (1) and obtain:

$$logExpS(X \rightarrow Y)_Z = sim(X,Y) + a.sim(X,p^Z) + b.sim(Y,p^Z) \qquad (7.2)$$

By convention, for any two categories X and Y, $0 \leq sim(X,Y) \leq 1$ and $sim(X,Y) = 1$ if and only if $X = Y$.

Now, equation 7.2 captures the idea that for single-premise specific arguments the expectations of property projection among categories are determined by a weighted sum of three factors: premise-conclusion similarity, premise-typicality, and conclusion-typicality. Equation 7.1, applied to a set of prototypes for categories, captures similarity, premise and conclusion typicality and asymmetry effects in CBI. For instance, when considering the sentence "emus have property S" people expect more that ostriches also have property S than that penguins have it. This is due to the similarity effect since $sim(emu, ostrich) > sim(emu, penguin)$. If we construct a "bird space" through some set of prototypes, this inequality would be immediately represented by the relative positions in the space of the two pairs $< emu, ostrich >$, and $< emu, penguin >$ (see Figure 7.2). And it can be measured via the distance function of the space. Since $sim(emu, ostrich) > sim(emu, penguin)$, if follows from (7.1) that $ExpS(emu \rightarrow ostrich)_{bird} > ExpS(emu \rightarrow penguin)_{Bird}$.

As mentioned before, this model also predicts asymmetry and premise and conclusion typicality. For instance, $ExpS(robin \rightarrow emu)_{Bird} > ExpS(emu \rightarrow robin)_{Bird}$ since $sim(robin, p^{bird}) > sim(emu, p^{bird})$ and $a > b$. Regarding conclusion-typicality assume, following the bird space in Figure 7.2, that $sim(ostrich, vulture) \approx sim(ostrich, robin)$ and that

$sim(ostrich, p^{bird}) \approx sim(vulture, p^{bird})$. Then $ExpS(ostrich \rightarrow robin)_{Bird} > ExpS(ostrich \rightarrow vulture)_{Bird}$ since $sim(robin, p^{bird})$ is significantly larger than $sim(vulture, p^{bird})$.

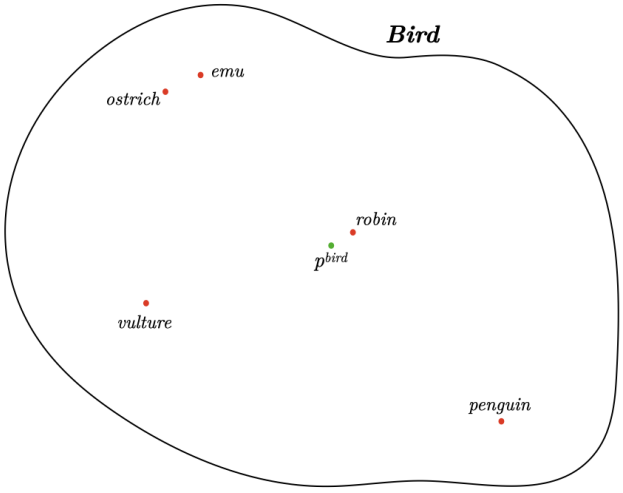


FIGURE 7.2: "Bird space" representing the positions of the different bird categories relative to a prototype.

7.3.2 A more general model

Concepts are represented as regions of conceptual spaces, not as points. We then need a model that accounts for this. One strategy is to consider distances between the prototypes of the categories, [4] and to take the volumes of the regions representing concepts in some CS —areas in the case of a 2-dimensional space— as predictors of expectations, i.e., argument strength in CBI. The volume of a concept in a conceptual space depends on the metric assigned to the space, and it is defined in a standard way. Note that the volume of a concept depends on the variability of properties that can be attributed to an object falling under that concept in each domain. For instance, it is expected that the concept *dog* has a larger volume than the concept *tiger*, since dogs can be of many different colors, shapes, and sizes; while tigers have less variability in these domains. The

[4] This is an idealization that involves to assumes that most concepts have single prototypes that can be represented by some point in the space. An alternative idea is to explicitly introduce distances between regions as a function of distances between their points (e.g., see Niiniluoto, 1987). It would be a matter of empirical testing to determine which model would give the best results.

immediate consequence of this is that the more heterogeneous the concept is, the larger its volume will be in a conceptual space.

Coming back to expectations, we assume that $ExpS(X \to Y)_Z$ is positively correlated with the volume $V(X)$ of X and negatively correlated to the volume $V(Y)$ of Y. The positive correlation is due to the fact that the larger $V(X)$, the more it "covers" —or is more representative of— the superordinate category Z. For example, $ExpS(bear \to wolf)_{Mammal}$ should be larger than $ExpS(polarbear \to wolf)_{Mammal}$ (see Figure 7.3).

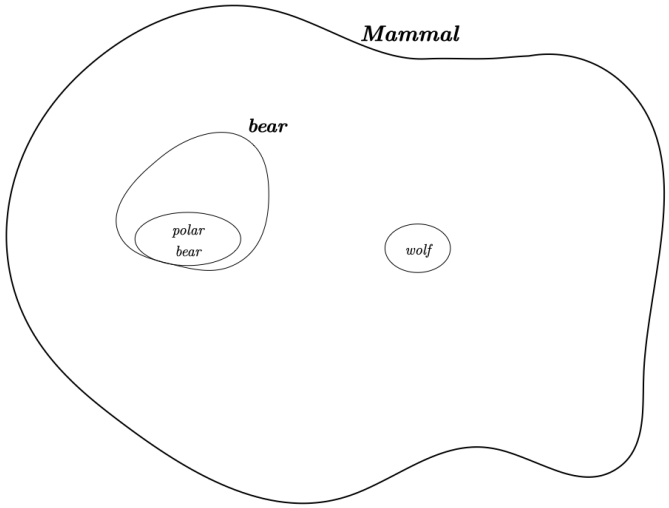


FIGURE 7.3: "Mammal space" representing the difference in volumes of *bear*, *polar bear* and *wolf*.

The negative correlation holds because the smaller the region Y is, the more likely it is for Y to have property S in the inductive argument. If X and Y cover overlapping regions of the space, then the relative sizes of their non-overlapping regions $X - Y$ and $Y - X$, that is , $V(X - Y)/V(Y - X)$ should be considered.

Building on (7.1), and considering the above ideas, the following equation is proposed:

$$ExpS(X \to Y)_Z = \left(d(p^X, p^Y)^{\frac{V(X-Y)}{(Y-X)}} . d(p^X, p^Z)^a . d(p^Y, p^Z)^b \right)^{-1} \tag{7.3}$$

Again, taking the logarithm and considering the relation between distance and similarity, we obtain:

$$logExpS(X \rightarrow Y)_Z = \frac{V(X-Y)}{(Y-X)}.sim(p^X, p^Y) + a.sim(p^X, p^Z) + b.sim(p^Y, p^Z) \tag{7.4}$$

In cases when X and Y are disjoint regions —which are the most typical ones— the quotient reduces to $V(X)/V(Y)$. And in cases when X and Y are represented by small non-overlapping regions, we can take $V(X)/V(Y) = 1$ and then (7.3) and (7.4), respectively, will have (7.1) and (7.2) as limiting cases. Just as (7.3), this new equation predicts premise-similarity, premise and conclusion-typicality, and asymmetry.

To see an example of how it works, consider the conclusion-typicality effect. As mentioned before, some experiments show a robust effect of conclusion typicality in CBI (Hampton & Cannon, 2004). For instance, an argument with categories $koala \rightarrow guineapig$ should be seen as weaker than an argument like $koala \rightarrow tiger$, since tigers are more typical mammals than guinea pigs. Assume, for the sake of argument that $sim(koala, guineapig) \approx sim(koala, tiger)$, and that $V(guineapig) \approx V(tiger)$. Then, using (7.4) we will have that,

$$\frac{V(koala)}{V(guineapig)}.sim(p^{koala}, p^{guineapig}) + a.sim(p^{koala}, p^{mammal}) + b.sim(p^{guineapig}, p^{mammal}) < \frac{V(koala)}{(tiger)}.sim(p^{koala}, p^{tiger}) + a.sim(p^{koala}, p^{mammal}) + b.sim(p^{tiger}, p^{mammal})$$

since, $b.sim(p^{guineapig}, p^{mammal}) < b.sim(p^{tiger}, p^{mammal})$. Then, it follows that $ExpS(koala \rightarrow guineapig)_{Mammal} < ExpS(koala \rightarrow tiger)_{Mammal}$. [5]

Note that (7.4) is not defined when $Y \subset X$, since in that case $Y - X = \emptyset$.

[5] It is possible that a concept with greater volume is less typical than a concept with a smaller volume. For example, *fish* may have a greater volume than *cat*, but being less typical as a *pet*. However, in equation 7.3, the expectation value is not only determined by the volume of a concept but also its typicality. So even though fish may have a larger volume than cat, the greater typicality of cat will counterweight this.

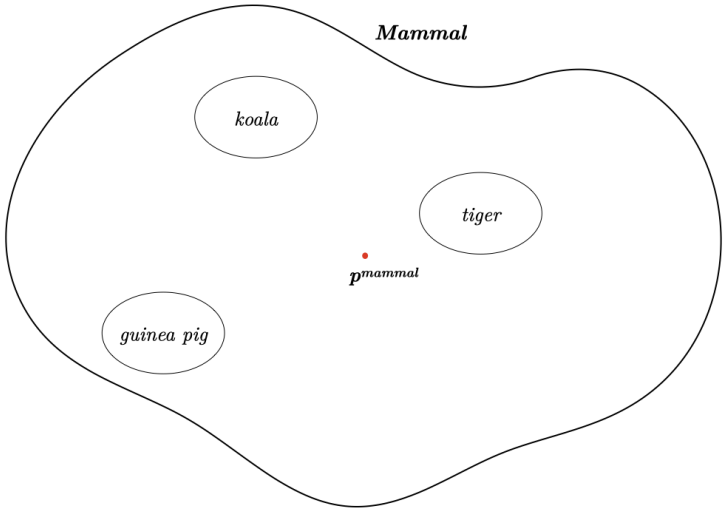


FIGURE 7.4: Mammal space for categories *koala*, *tiger* and *guinea pig*.

However, in this case the expectation of property projection is maximal, and we can define it to be some arbitrary high value.[6] For general judgments, for example $ExpS(vulture \rightarrow bird)_{Bird}$, we have $Y = Z$ and $X \subset Y$, and since $sim(bird, bird) = 1$, (7.4) will consequently take the following form:

$$logExpS(X \rightarrow Z)_Z = a.sim(p^X, p^Z) + b \tag{7.5}$$

This coheres with the idea that single premise/general arguments depend, mostly, on premise-typicality relations, that is, on the idea that agents represent a category with a certain degree of typicality in the context of a more abstract superordinate category.

This is not a minor point. In this model, it is assumed that agents cannot compare categories from different conceptual levels directly in terms of similarity —like comparing *collie* with *mammal*. Instead, the categorical relation that works in these cases is typicality, which comes by default for all categories in a conceptual space, given that conceptual spaces are constructed and articulated

[6] A reason for this assignment is that if Y is a region that partially overlaps X and then shrinks to become a subset of X, then $V(Y - X)$ will be smaller and smaller, which means that $ExpS(X \rightarrow Y)_Z$ will approach infinity. From a psychological perspective, we hypothesize that these cases require agents using an inferential mechanism based on class-inclusion, like property-inheritance. If $Y \subset X$, members of Y inherit all properties of X, thus, $ExpS(X \rightarrow Y)_Z$ is maximal.

around prototypes. As we will see later, this is an advantage over the two classical models of CBI, which have more difficulties representing typicality relations among categories.

Now, Sloman (1993) observes that subjects exhibit an "inclusion fallacy" since the argument "Robins have property S; thus, birds have property S" is judged to be stronger than "Robins have property S; thus, ostriches have property S" despite the fact that ostriches form a subset of birds. The CS model can explain this phenomenon. To see how, note that —since $V(robin - bird) = \emptyset$— we have that $logExpS(robin \rightarrow bird)_{Bird} = a.sim(p^{robin}, p^{bird}) + b$, and that

$$logExpS(robin \rightarrow ostrich)_{Bird} = \frac{V(robin)}{V(ostrich)}.sim(p^{robin}, p^{ostrich}) + a.sim(p^{robin}, p^{bird}) + b.sim(p^{ostrich}, p^{bird}).$$

Then, $logExpS(robin \rightarrow bird)_{Bird} > logExpS(robin \rightarrow ostrich)_{Bird}$, as long as

$$(a.sim(p^{robin}, p^{bird}) - a.sim(p^{robin}, p^{bird})) + (b - sim(p^{ostrich}, p^{bird})) > \frac{V(robin)}{V(ostrich)}.sim(p^{robin}, p^{ostrich}),$$

which would typically be the case since $sim(p^{ostrich}, p^{bird})$ is small.

This shows that this model, unlike the similarity-coverage model (Osherson et al., 1990), also predicts results that are not valid under all conditions, but only under certain specific circumstances.

The CS-model can also predict the conclusion-specificity phenomenon (Osherson et al., 1990, p.187), which says that people tend to judge arguments with more specific categories as stronger than argument with more abstract categories. For instance, an argument from *crow* to *bird* will be judged as stronger than an argument from *crow* to *animal*. This is easily explained by the CS-model because the more abstract the conclusion category is, the bigger its volume in the conceptual space, and the further the prototype of this category will be from the prototype of the premise-category (see Figure 7.5).

Considering the above example, we have that, $logExpS(crow \rightarrow bird)_{Bird} = a.sim(p^{crow}, p^{bird}) + b$, and that
$logExpS(crow \rightarrow animal)_{animal} = a.sim(p^{crow}, p^{animal}) + b$. Since,
$d(p^{crow}, p^{animal}) > d(p^{crow}, p^{bird})$, then
$sim(p^{crow}, p^{bird}) > sim(p^{crow}, p^{animal})$.
Making $a.sim(p^{crow}, p^{bird}) > a.sim(p^{crow}, p^{animal})$. It follows then$ExpS(crow \rightarrow bird)_{Bird} > ExpS(crow \rightarrow animal)_{Animal}$.

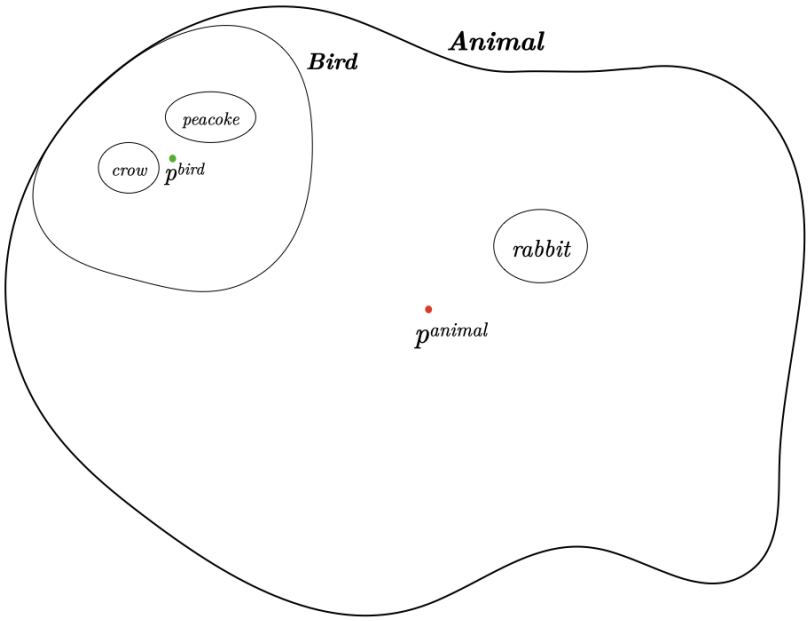


FIGURE 7.5: *Animal space* including the subspace *bird*

7.3.3 Arguments with multiple premises

In single-premise arguments the focus is on the relation between the premise and the conclusion categories; but when dealing with multiple premises, we must also account for premise-premise categorical relations. The main phenomenon to model in these cases is *diversity*, i.e., the more different are the categories in the premises, the stronger the argument. For instance, the argument (i) "Jaguars and leopards have property P; thus, otters have property P," is weaker than the argument (ii) "Jaguars and elephants have property P; thus, otters have

property P." That is because the categories *jaguar* and *leopard* are similar and they provide less "coverage" of the superordinate category *mammal* than *jaguar* and *elephant*.

The diversity phenomenon suggests that argument strength is a negative function of premise-premise similarity. One possible way of modeling this would be to compute the pairwise similarity of category premises, but this would represent a significant increase in the computational complexity of the process, particularly when we have arguments with more than two premises. The proposal presented here tries to avoid this by considering all the premises' categories as part of one large inclusive set. In a sense, the premise categories can be seen as "exemplars" of a more general category. More precisely, we can model n-premises arguments by considering the convex hull of the categories $X_1, X_2, \ldots, X_n$ in the premises. A convex hull of a set S —denoted by $C(S)$— is the smallest convex region containing all elements in S (see Devadoss S., 2011, for a detailed explanation).

Convex hulls are also convex regions of n-dimensional spaces with the same geometrical properties as regions in conceptual spaces. The size of their volumes is positively correlated to the number of convex regions they include, as well as to the distances among these regions. For instance, in a conceptual space in which all the categories have similar volumes, the volume of the hull of two contiguous regions is going to be smaller than the volume of two non-contiguous regions of the space [7]. This is precisely the kind of property of interest to represent the diversity phenomena. For example, if we consider the argument described above, in an animal space the categories *jaguar* and *leopard* would be represented by contiguous (or very close) regions in the space, while *jaguar* and *elephant* would be far from each other. As a consequence, the volume of $C(jaguar \cup leopard)$ would be smaller than the volume of $C(jaguar \cup elephant)$, and then it would provide less coverage of the *mammal* category (see Figure 7.6).

[7] For cases in which this condition does not hold, it is possible that the volume of the hull of two large contiguous regions is larger than the hull of two distant small regions. An empirical study of this fact could be a way of testing the fruitfulness of the notion of volume of a category for the analysis of CBI.

However, one problem of this approach is that we don't have a "natural" prototype —like p^X in equation (7.3)— for the premise anymore. The solution proposed is to consider an "artificial" prototype p^C, at the centroid of the convex hull $C(X_1 \cup X_2 \cup \cdots \cup X_n)$ [8]. For convex hulls, we can then reformulate (7.4) for multiple premises in the following way:

$$logExpS(X_1, X_2, \ldots, X_n \to Y)_Z = \frac{V(C(X_1 \cup X_2 \cup \cdots \cup X_n) - Y)}{V(Y - C(X_1 \cup X_2 \cup, \ldots, \cup X_n))}.sim(p^C, p^Y) + a.sim(p^C, p^Z) + b.sim(p^Y, p^Z) \tag{7.6}$$

To see how this formula predicts diversity, consider the example at the beginning of this section. According to (7.6), we have that

$$ExpS(jaguar, leopard \to otter)_{Mammal} = \frac{V(C(jaguar \cup leopard)}{V(otter)}.sim(p^C, p^{otter}) + a.sim(p^C, p^{mammal}) + b.sim(p^{otter}, p^{mammal});$$

This is smaller than

$$ExpS(jaguar, elephant \to otter)_{Mammal} = \frac{V(C(jaguar \cup elephant)}{V(otter)}.sim(p^{C*}, p^{otter}) + a.sim(p^{C*}, p^{mammal}) + b.sim(p^{otter}, p^{mammal}).$$

since $V(C(jaguar \cup elephant)) > V(C(jaguar \cup leopard))$. Which makes

$$\frac{C(jaguar \cup elephant)}{V(otter)}.sim(p^{C*}, p^{otter}) > \frac{V(C(jaguar \cup leopard)}{V(otter)}.sim(p^C, p^{otter})$$

when $sim(p^{C*}, p^{otter}) \geq sim(p^C, p^{otter})$.

[8] This assumption is not meant to be psychologically realistic. Prototypes are hardly centroids of the convex regions that represent them, even for natural categories (see Douven, 2019). However, according to the empirical literature, the typicality effect holds for multiple-premise arguments as a compound measure of the degree of typicality of some of the categories in the premises. Considering the lack of robust evidence about how these degrees of typicality interact, we introduced the centrality of the artificial prototype as a formal idealization that seems to respond well to the classical examples.

The result also follows in the case in which $sim(p^{C*}, p^{otter}) < sim(p^{C}, p^{otter})$, when the difference between $\frac{V(C(jaguar \cup elephant))}{V(otter)}$ and $\frac{V(C(jaguar \cup leopard))}{V(otter)}$ is enough to make $\frac{V(C(jaguar \cup elephant))}{V(otter)}.sim(p^{C*}, p^{otter}) > \frac{V(C(jaguar \cup leopard))}{V(otter)}.sim(p^{C}, p^{otter})$. Again, this is a conclusion that is not always valid, but depends on the relations between the categories involved.

If Y is a sub-region of $C(X_1 \cup X_2 \cup ... \cup X_n)$, then $Y - C(X_1 \cup X_2 \cup ... \cup X_n) = \emptyset$, so (7.6) is undefined. For the same reasons as before, we can set this to some maximal value. For example, if *buzzard* belongs to the convex hull of *eagle*, *kite*, and *harrier*, it would follow that $ExpS(eagle, kite, harrier \rightarrow buzzard)_{Bird}$ would be maximal. This is the first prediction that emerges from the CS model. It is an interesting empirical problem, whether this would correspond to the judgment of real subjects. As far as we are aware, this phenomenon has not been tested.

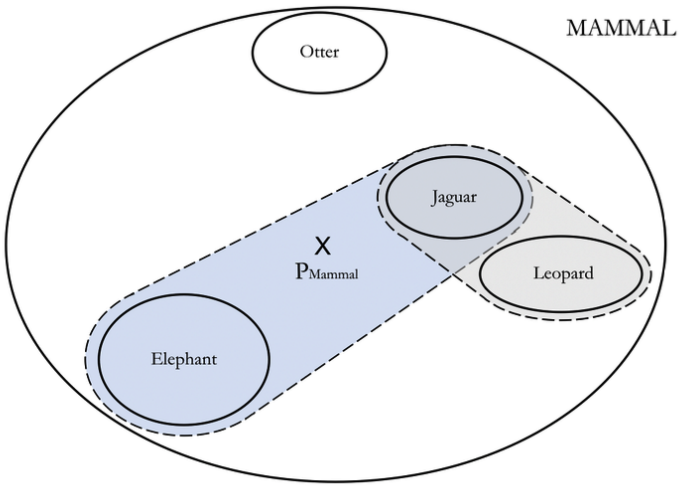


FIGURE 7.6: Mammal space illustrating that the volume of $(elephant \cup jaguar)$ is larger than the volume of $(leopard \cup jaguar)$

If equation (7.6) is applied to multiple premise general arguments then $C(X_1 \cup X_2 \cup ... \cup X_n) - Y = \emptyset$,since $C(X_1 \cup X_2 \cup ... \cup X_n) \subset Y$. Then, (7.6) reduces to

$$logExpS(X_1, \ldots, X_n \rightarrow Y)_Z = a.sim(p^C, p^Z) + b.sim(p^Y, p^Z) \qquad (7.7)$$

Note that when $Y = Z$, $sim(p^Y, p^Z) = 1$ and (7.7) reduces to

$$logExpS(X_1, \ldots, X_n \rightarrow Y)_Z = a.sim(p^C, p^Z) + b \tag{7.8}$$

A problem with this expression is that is does not account for the diversity of $X_1, X_2, \ldots, X_n$, but only the prototype p^C. One way of solving this problem is to let the constant a depend on the proportion of Z that is covered by $X_1, X_2, \ldots, X_n$, that is, $V(C(X_1, X_2, \ldots, X_n))/V(Z)$. However, since empirical evidence for this case seems to be lacking, this there will not be pursued here.

Let us now see how this model can deal with monotonicity. As it was mentioned before, the monotonicity effect states that adding premises to a CBI argument increases expectations of property projection when the new premise-categories are also included in the original evoked superordinate category of the argument. For instance, adding a premise with the category *pig* to the argument $(fox\&wolf) \rightarrow mammal$ is going to strengthen it. Note that adding a premise-category to an argument will increase the volume of the convex hull of the premises. And in most cases, the volume of that set is negatively correlated to the distance between p^C and p^Y , that is, the more $V(C(X_1, X_2, \ldots, X_n))$ approximates $V(Y)$, the closest p^C is to p^Y . Then, for the above arguments we have that if p^C is the centroid of $C(fox \cup wolf)$ and p^{C*} is the centroid of $C(fox \cup wolf \cup pig)$, since $V(C(fox \cup wolf \cup pig)) > V(C(fox \cup wolf))$ then $sim(p^{C*}, p^{mammal}) > sim(p^C, p^{mammal})$, and as a consequence $ExpS(fox, wolf, pig \rightarrow mammal)_{Mammal} > ExpS(fox, wolf \rightarrow mammal)_{Mammal}$.

This model can also predict Sloman's (1993) observation that diversity has a limit. To analyze his example (mentioned in Section 7.2.4), note that $sim(p^C, p^{mole})$, where p^C is the prototype of $C(germanshepherd \cup elephant)$, is considerably larger than $sim(p^{C*}, p^{mole})$, where p^{C*} is the prototype of $C(germanshepherd \cup bluewhale)$. Similarly $\text{sim}(p^C, p^{mammal}) >$

$sim(p^{C*}, p^{mammal})$. Then it typically follows that:

$$\left(\frac{V(C(germanshepherd \cup elephant))}{V(mole)}.sim(p^C, mole) + a.sim(p^C, mammal) + b.sim(p^{mole}, p^{mammal})\right) > \left(\frac{V(C(germanShepherd \cup blueWhale))}{V(mole)}.sim(p^{C*}, mole) + a.sim(p^{C*}, mammal) + b.sim(p^{mole}, p^{mammal})\right)$$

Next, suppose that we add to some premise-set a new premise with a category that is not included in Z. What will happen is that the new modified argument will have a different (and more abstract) evoked superordinate category $Z*$ such that $Y \subset Z*$ with $Z \subset Z*$. According to the empirical literature, subjects should perceive the new argument as weaker than the original one, making CBI nonmonotonic. Remember the example of nonmonotonicity that we gave in Section 2.4: the argument $(peacock\&crow) \rightarrow bird$ is stronger than the argument $(peacock\&crow\&rabbit) \rightarrow bird$ (see Figure 7.5). According to the analysis presented here, adding the premise *rabbit* change the evoked superordinate category (Z) from *bird* to *animal*. Then, the CS-model correctly predicts that $logExpS(crow, peacock \rightarrow bird)_{Bird} > logExpS(crow, peacock, rabbit \rightarrow bird)_{Animal}$; because $a.sim(p^C, p^{bird}) + b > a.sim(p^{C*}, p^{animal}) + b.sim(p^{bird}, p^{animal})$. Since $a.sim(p^C, p^{bird})$ should larger than $a.sim(p^{C*}, p^{animal})$, and $b.sim(p^{bird}, p^{animal}) < b$.

7.3.4 Knowledge effects and nonblank properties

The model presented so far only focuses on two types of semantic relations among premise and conclusion categories, namely similarity and typicality. However, newer experimental results on CBI have shown that there are other cognitive mechanisms that also influence inductive judgments. Beyond similarity and typicality relations, different kinds of knowledge about premise and conclusion categories (Coley, Shafto, Stepanova, & Baraff, 2005; Shafto, Coley,

& Vitkin, 2007) or different reasoning heuristics (Rehder, 2006) might shape inductive inferences. For instance, there is evidence that knowledge about thematic relations of the categories involved in the arguments (Coley et al., 2005), as well as expertise in some domain related to the topic of the arguments (Proffitt, Coley, & Medin, 2000), can play an important role in the agent's expectations of property projection. Furthermore, in most CBI cases, people also use knowledge (or make hypotheses) about the property projected for estimating the strength of the argument. In any case, a full model of CBI should account for the effects of background knowledge and consider non-blank properties. The conceptual spaces approach seems rich enough to deal with CBI's most studied knowledge effects concerning non-blank properties. In what follows, a strategy on showing how this can be done will be briefly explained.

An influential criticism of the similarity-based models of CBI was presented by Heit and Rubinstein (1994). They showed that it is not possible to account for some knowledge effects that influence inductive inferences using only a single similarity measure. In particular, they showed that in category-based arguments with nonblank properties, the agents' knowledge about the property S that was projected modulated the similarity measure that was used for comparing the categories in the premise and conclusion. For instance, they showed that arguments of the form *chicken* $\rightarrow$ *hawk* are judged as stronger than arguments of the form *tiger* $\rightarrow$ *hawk* when the property projected is anatomical —such as "has a liver with two chambers." But the opposite holds when the property projected is behavioral —such as "prefer to feed at night."

For explaining this phenomenon, we will no see an extension of the CS-model that includes a similarity measure that puts larger weights on the categories involved in the projected properties. This focus would be determined by the dimensions of the non-blank property in the arguments. When the agent has little knowledge of the property S that is projected —which is by definition the case for a blank property—, she will compare categories using a general similarity measure. However, if the agent has more precise knowledge about S —at least about what kind of property S is—, it is expected for her to use a similarity measure that gives more weight to the dimensions related to S.

Formally this can be done by using a weighted distance function like the one introduced in Chapter 4.3. Equation 7.1 would be reformulated in the following way:

$$ExpS(X \rightarrow Y)_Z = \left(d^{(S)}(X,Y).d^{(S)}(X,p^Z)^a.d^{(S)}(Y,p^Z)^b\right)^{-1}$$

Where the function $d^{(S)}(x,y)$ is defined as the distance between x and y when the domains relative to S are salient. In the example from Heit and Rubinstein (1994), when the projected category refers to anatomical properties —"has a liver with two chambers"—, the model will predict that the argument *chicken* → *hawk* will be judged as stronger than *tiger* → *hawk*. On the other hand, if the projected category refers to behavioral properties —"prefer to feed at night"—, then the model will predict the converse relation — since now the weight to the behavioral domain will make *tiger* more similar to *hawk*, just as it was found in the experiments by Heit and Rubinstein.

Medin, Coley, Stroms, and Hayes (2003) showed that property effects also show up in arguments with blank properties. In particular, they discovered a non-diversity effect by property reinforcement that occurs when some salient feature shared between the premise-categories leads the agent to produce hypotheses about the nature of the property S that is projected (cf. Shafto et al., 2007). As a result, the agent will use a weighted similarity measure that can override normal diversity effects. For instance, according to what we saw so far, the argument "Polar bears and antelopes have property S; thus, all animals have property S" should be considered weaker than the argument "Polar bears and penguins have property S; thus, all animals have property S," since the first premise set is less diverse than the second. However, in this case, both polar bears and penguins inhabit cold areas, leading agents to hypothesize that property S is related to this shared feature. That will weaken the argument since properties of this kind are atypical regarding animals in general.

The interpretation of this example defended here is that the conjunction of *polar* and *penguin* evokes a new —non-taxonomic— minimal superordinate, namely *animal in cold areas* and thereby that the property S somehow is related to this superordinate. The superordinate animal in cold areas generates a new

way of classifying the similarity animals, that is a new distance function $d*$. As a result after applying Equation 7.7, we expect that

$$logExpS(polarbear, antelope \rightarrow animal)_{Animal} > logExpS(polarbear, penguin \rightarrow animal)_{Animal}$$

since $a.sim(p^C, p^{Animal}) + b$ will be bigger than $a.sim(p^{C*}, p^{Animal}) + b$, because the distance between p^C and p^{Animal}will be smaller than the distance between p^{C*} and p^{Animal} since *animal in cold areas* is a rather small and atypical region of *animal*.

Finally, similar ideas can be applied for explaining some of the effects of expertise in CBI (Proffitt et al., 2000). For a non- expert, the two inferences "Dutch elms have disease A; thus, ginkgo trees have disease A" and "River birches have disease A; thus ginkgo trees have disease A" would, for lack of knowledge, be judged to be equally strong. However, for a tree expert, the knowledge that ginkgo trees are more similar to Dutch elms when it comes to which diseases affect them would make the first inference stronger than the second. In brief, for experts, the distance measure $d^{(S)}$ in the model would be dependent on that S relates to diseases, while this would not affect the non-experts' judgments.

These examples of how knowledge effects can be handled by the CS-model show that it is able to cover a wide variety of experimental findings from the literature. [9]

[9]One issue that will not be considered here is the influence of causal relations between the concepts involved. Various experimental studies have shown that causal knowledge is important in CBI, sometimes overriding standard similarity and typicality relations (e.g., see Bright & Feeney, 2014; Medin et al., 2003; Rehder & Hastie, 2001; Shafto et al., 2007). For example, "Grass has enzyme E; thus, cows have enzyme E" is judged to be stronger than "Cows have enzyme E; thus, grass has enzyme E" since there may exist a causal link from the enzyme of the grass to the enzyme of the cows. One possible way to use the CS model for these phenomena is that causal connections may introduce a different kind of typicality relations between the concepts so that the presence of the enzyme is more typical for grass than for cows.

7.4 Previous models of CBI

7.4.1 The Similarity-coverage model

The first formal model of CBI was the similarity-coverage model (SCM), proposed by Osherson et al. (1990). In this model, argument strength in CBI is judged on the basis of two factors: (i) premise–conclusion similarity, and (ii) the degree of coverage that the premise's category has regarding the lowest superordinate category that includes both the category of the premise and the category of the conclusion.

For specific arguments, argument strength depends only on (i). If the argument has multiple-premises, the model uses a maximum rule that estimates premise–conclusion similarly by focusing on the premise with the most similar category to the conclusion's category. For instance, for an argument like "Horses and bats have property S; thus, cows have property S", argument strength will be determined by $Maxsim\big[(horse, cow), (bat, cow)\big]$, which will return $sim(horse, cow)$.

Coverage is a more complex notion. The model assumes —as we did here— that CBI with natural categories involves "evoking" an implicit superordinate category that includes all the categories in the argument. Coverage is then a relation between the premises' categories with that superordinate category, and it is also explained in terms of similarity. More specifically, coverage is an average measure of several pairwise similarity judgments that compare the premise's category with members – "examples" – of the superordinate category in question; and it is a negative function of similarity among premises. For instance, consider the following arguments:

(a) Horses have sesamoid bones.
Cows have sesamoid bones.
∴ Mammals have sesamoid bones.

(b) Horses have sesamoid bones.
Rats have sesamoid bones.
∴ Mammals have sesamoid bones.

(a) is weaker than (b) because the pair (*horse*, *cow*) provides less coverage of *mammals* than the pair (*horse*, *rat*). In particular, the degree of coverage

can be associated with the extension of the set which includes all the categories similar to those of the premises. In (a), that set is relatively small because most categories that are similar to *horse* are also similar to *cow*. In (b), however, that set is bigger, since most categories similar to *rat* are not in the set of categories that are similar to *horse*.

Coverage is also related to typicality. The SMC assumes that typical categories are associated with larger sets of similar categories (of the same conceptual level) than atypical ones. For instance, the argument "Horses have property S"; thus, mammals have property S" is stronger than "Bats have property S; thus, mammals have property S" because the set of mammals similar to horses is larger than the set of mammals similar to bats.

Despite being a very successful model thanks to its predictive power, the SCM has various limitations. One of them is that it does not build on a psychologically grounded notion of similarity. For instance, the model does not include any precise notion of similarity relations among categories. It uses similarity as an empty notion that can be filled out with different specific measures. As it was mentioned before, it is desirable for a theory of CBI to build on some fundamental theory of conceptual knowledge; one that provides a basic notion of conceptual similarity and that can be used to give a unified explanation of the diversity of concept-based cognitive phenomena (categorization, concept formation, language-learning, etc.). Furthermore, as observed by Tenenbaum, Kemp, and Shafto (2007), the SCM lacks a systematic mathematical foundation. This is also related to the previous point. The formal structure of this model is not based on any formal model of inference or categorical relation, but it was directly designed to model the properties of CBI as described by the empirical studies.

The approached presented here, while it is close to the SMC model in various respects, does not suffer from the aforementioned problems since both the formal and the psychological foundations of the model come from the general theory of conceptual spaces. Furthermore, the CS-model can account for the same range of CBI phenomena than the SMC model, while also predicting some results that are valid in special cases, something that the SMC model cannot do.

7.4.2 The feature-based model

The other well-known model of CBI was proposed by Sloman (1993) as an alternative to the SCM. Sloman started by criticizing the assumption that reasoning with categories involves the necessary representation of their hierarchical structure. He argued that inclusion fallacies in reasoning form strong evidence against that idea. As an alternative, he proposed to understand categorical relations as based on the overlap of features. "Features", Sloman claims, "*represent a large number of interdependent perceptual and abstract attributes. In general, these values may depend on the context in which categories are presented*" (Sloman, 1993, p. 237).

Sloman develops his feature-based model within a connectionist framework in which categories are represented as sets of features described by vectors of real numbers from the $[0, 1]$ interval. With it, he is able to explain ten of the patterns explained by SCM and three new ones, not treated by Osherson et al. (1990). He also presents empirical support for the new patterns. The central idea of this model is that argument strength is positively correlated with the proportion of features in the conclusion category that are also included in the premise categories. For instance, in the simple case "All Xs are S; thus, all Ys are S", the premise category X, and the conclusion category Y can be represented by two vectors $F(X)$ and $F(Y)$ of feature values. The strength of the inductive argument is determined by the following expression: $\frac{F(X)\cdot F(Y)}{|F(Y)|^2}$, where $F(X) \cdot F(Y)$ can be seen as a measure of the overlap of the features of X and Y, and $|F(Y)|^2$ a measure of the magnitude of the conclusion category vector. [10]

Unlike the SCM, the feature-based approach does not have foundational issues, since it is developed within a connectionist framework. [11] One could think that this connectionist background leaves no space to the CS approach. However, as it has been argued before (Gärdenfors, 1997; Lieto et al., 2017), CS is compatible with connectionist approaches.

[10] $F(X) \cdot F(Y)$ is the inner product of the two vectors, defined as $\sum_i F(X)_i.F(Y)_i$ and $|F(Y)|^2$ is the inner product of $F(Y)$ with itself, defined as $\sum_i F(Y)_i^2$.

[11] See Rogers and McClelland (2004) for a connectionist approach to semantic cognition.

In general, the main ideas of Sloman's model are not in contradiction with the CS approach. In fact, they could be translated into this framework. The theory of conceptual spaces also assumes that concepts are represented as collections of properties from different domains. The feature-overlap measure that Sloman's use to determine argument strength could be replaced by a similarity measure in a conceptual space covering the dimensions of the feature vector.

One important advantage of the CS-model over the feature-based approach concerns the representation of typicality relations. In Sloman's model, there is no specific mechanism for accounting for typicality. Both typicality and similarity relations are reduced to feature-overlap. The model can account for typicality effects in general arguments because it assumes that typical categories —such as *apple*) share more features with their immediate superordinate category (*fruit* in this case— than non-typical categories. However, this model cannot account for independent premise-typicality effects in specific arguments. For instance, if we have three categories A, B and C, and A is more typical than B but both categories A and B have the same feature overlap with C, then the model would predict the arguments $A \rightarrow C$ and $B \rightarrow C$ to be equally strong (Heit, 2000, p. 586). The CS approach does not have this limitation since it is able to explicitly represent independent typicality relations both in general and specific arguments. To give an example, consider two arguments of the form *quince* $\rightarrow$ *pineapple* and *apple* $\rightarrow$ *pineapple*. The categories *apple* and *quince* have the same feature overlap with *pineapple*, but since *apple* is a more typical fruit than *quince*, $sim(p^{apple}, p^{fruit})$ is going to be significantly larger than $sim(p^{quince}, p^{fruit})$, consequently $ExpS(apple \rightarrow pineapple)_{Fruit} > ExpS(quince \rightarrow pineapple)_{Fruit}$. This is a second example of a new prediction that follows from the CS model.

In general, the two models presented here provide different insights into CBI. One interesting thing about the CS-model, is that it combines the two main features of the SMC and Sloman's model: it is a similarity-based model that includes a feature-based view of categories. Furthermore, the CS approach has an important theoretical advantage regarding these other two models; it inherits from the theory of conceptual spaces an explanation of how knowledge

domains are formed and structured, and how they are grounded on perception an action. In that way, the approach presented here is grounded on a systematic psychological theory about the nature and structure of conceptual systems. At the same time, this psychological theory comes with a formal model of some of the main cognitive mechanisms behind conceptual processes. As mentioned before, the CS-model leverages this formal model and, in that sense, builds on a solid mathematical foundation. Another difference is, as Feeney (2017, pp. 172-173) notes, that neither SMC, nor the feature-based model can explain the conclusion effect reported by Hampton and Cannon (2004).

7.4.3 Bayesian models

Besides these two classical models of CBI, Bayesian accounts have recently become influential in the literature. The first proposal in this area was advanced by Heit (1998) and consisted of a computational-level analysis that puts the agent's knowledge about properties at the center stage of the process of CBI. His idea is that while evaluating a CBI argument, the agents estimate the probability of property projections among categories based on her estimation of the range of the property projected —i.e., the set of categories for which the property is true and the set of categories for which the property is false—. For doing so, the agent exploits prior knowledge about other familiar properties, under the assumption that the property projected is distributed in a similar manner.

For instance, in an argument of the form "X has property S; thus, Y has property S", the agent will reason from a set of four basic hypotheses about the possible range of S: (1) S is true of X and Y, (2) S is true of X and false of Y, (3) S is false of X and true of Y, and (4) S is false of X and Y. The prior probability distribution for these hypotheses may vary according to the similarity between X and Y or other categorical relations. Then, using the premise of the argument as evidence, the agent will update their beliefs about the set of hypotheses and estimate the probability of the conclusion using Bayes' theorem.

Heit showed that his approach predicts as many properties of CBI as Osherson's and Sloman's models. However, it has also an important drawback: it does not include any mechanism for estimating the prior distribution over the hypotheses about the range of the property. This is mainly because, unlike the other models (including ours), the Bayesian approach is centered on property-relations instead of categorical-relations.

Tenenbaum et al. (2007) followed Heit's approach and made some important improvements regarding the above problem. Their strategy consists of defining a set of structures with information about the agent's knowledge of categorical relations in different domains and knowledge about the compatibility of different properties with these relations. These structures will determine the prior probability that some property P may be projected from one category X to a category Y from the same domain. Then, these probabilities may be updated according to standard Bayesian rules when considering specific category-based arguments.

This approach can work with different types of knowledge structures. Taxonomic systems of categories, causal structures, or spatial knowledge are some of the knowledge structures that have been studied for CBI in the Bayesian tradition. This represents an important advantage over the SCM and Sloman's model, which have serious troubles for dealing with forms of inductive reasoning that do not involve natural categories.

There are, however, considerable drawbacks of Bayesian models of CBI. One is that there is no natural way to represent similarity and typicality in these models. Another is that probabilistic reasoning is very resource-demanding when implemented computationally. These drawbacks make the Bayesian models psychologically unrealistic. [12]

[12] Yang and Long (2020, p. 9) recently found some empirical evidence for this claim. See also Jones and Love (2011) for a general criticism of the use of Bayesian models in cognitive science.

7.5 Methodological considerations

Empirical studies are crucial for research on category-based induction. A major challenge for is then to develop quantitative tests for testing the models. The framework presented in this chapter opens up for a new methodology of investigating category-based induction. The distance measure and the similarity and betweenness it generates will allow new and more precise quantitative predictions. We have already seen that, according to the CS-model, when Y is a subregion of $C(X_1 \cup X_2 \cup \cdots \cup X_n)$, the prediction is that $ExpS(X_1, X_2 \rightarrow Y)_Z$ should be maximal. For regions Y, X_1 and X_2 ; Y lies *between* X_1 and X_2 if, for every y in Y, there are points x_1 and x_2 in X_1 and X_2 respectively, such that y is between x_1 and x_2. Given this definition, a special case of the prediction above is that when Y lies between X_1 and X_2, then $ExpS(X_1, X_2 \rightarrow Y)_Z$ should be maximal. A second new prediction concerns explicit representations of independent typicality relations as was discussed before.

Some other predictions are related to the introduction of the notion of *volume* of a category. First, the CS-model predicts that premise-specificity is negatively correlated to argument strength. More formally, it is to be expected that for categories Y, X_1 and X_2, if $X_1 \subset X_2$ then $ExpS(X_2 \rightarrow Y) > ExpS(X_1 \rightarrow Y)$ because $V(X_2) > V(X_1)$. For instance, arguments of the form *germanshepherd* $\rightarrow$ *cow* (or *mammal*) should be considered as weaker than arguments of the form *dog* $\rightarrow$ *cow* (or *mammal*). Second, the model predicts that for categories X_1, X_2 and Y, if it is the case that X_1 , X_2 are equally typical, but that $V(X_1) > V(X_2)$ then $ExpS(X_1 \rightarrow Y) > ExpS(X_2 \rightarrow Y)$. These two predictions hold *ceteris paribus*.

These predictions are interesting ways of testing this model. However, doing that depends on having operational procedures for determining betweenness, similarity and distances. There are general methods for estimating psychological distances, such as Multi-Dimensional Scaling (MDS) (see Hout, Papesh, & Goldinger, 2013, for a review) and Principal Component Analysis (PCA) (see Abdi & Williams, 2010, for a review). For example, by asking subjects to judge the similarities of a number of different categories, the data can be analyzed by

MDS or PCA in order to generate a low-dimensional conceptual space with a distance measure. Once the distance is established, similarity and betweenness can be determined, and the predictions presented above can be tested.

As an example of the relevant type of data collection, Hampton and Cannon (2004) asked subjects to rate the premise typicality, conclusion-typicality and premise–conclusion similarity on a seven-graded scale. This data could also have been used to estimate an underlying distance measure that would have allowed Eqs. (1) or (2) to be tested. [13]

7.6 Conclusions

Category-based induction is a fundamental cognitive mechanism for everyday reasoning that has become a focus of research only during the last decades. In this chapter, a new mathematical framework of such inferences, that can explain almost all of the available empirical data, was presented. The model subsumes the earlier SCM by Osherson et al. (1990) and Sloman's (1993) feature-based model and it generates new predictions. Furthermore, it builds on solid formal foundations and it relies on a systematic theory of conceptual knowledge that has been proven successful in the explanation and modeling of others concept-based cognitive phenomena.

From a philosophical point of view, CBI is an interesting phenomenon because it clearly shows the degree to which concepts and reasoning are intertwined, challenging the formalist thesis. In Chapter 5, we saw how material inferences exploit core semantic knowledge about the structure of lexical concepts, and Chapter 6 showed how nonmonotonic reasoning uses the prototypical structure of categories to tackle uncertainty. Here, we saw a more complex inferential mechanism that exploits a combination of properties of category representation to make inductive inferences with sparse information. All things considered, I believe that the ideas so far defended make a case for developing a richer view of reasoning that goes beyond the rule-based and propositional approach and takes conceptual representation as constitutive of the very process of inferring.

[13] Also Rips (1975) uses data from MDS for analyzing CBI arguments.

Furthermore, the models presented so far complement well other attempts of explaining forms of semantic-based inference with conceptual spaces, like reasoning with analogies and metaphors in (Gärdenfors, 2000, 2008), or interpolative inference in (Schockaert & Prade, 2013). If such a program can be worked through, it would form a unified basis for human inference that considerably extends the classic logicist and probabilistic approaches.

Chapter 8

Beyond language: Model-based inference in science

In this chapter, I analyze the role of concepts and representation in scientific reasoning. I reconstruct Stephen Toulmin's procedural theory of concepts and explanations to develop two overlooked ideas from his philosophy of science: *methods of representations* and *inferential techniques*. I argue that these notions, when adequately articulated, could be useful for shedding light on how scientific reasoning is related to representational structures, concepts, and explanations within scientific practices. I explore and illustrate these ideas by studying the development of the notion of instantaneous speed during the passage from Galileo's geometrical physics to analytical mechanics. In the end, I argue that methods of representations could be considered as constitutive of scientific inference; and I show how these notions relate to other similar ideas from contemporary philosophy of science, like those of models and model-based reasoning.[1]

[1]This chapter is based on the article "Methods of Representation as Inferential Devices", published in 2019 in the *Journal for General Philosophy of Science* (Osta-Vélez, 2019).

8.1 Introduction

Throughout this dissertation, I have argued in favor of understanding reasoning as a mechanism exploiting much more than propositional structure. In Chapter 3, a *pluralistic* view of inference was advanced: depending on the cognitive task, we represent conceptual information in different ways —images, formulas, diagrams, natural language, and so on. Since inferential mechanisms depend on the representational structures that support them, the diversity of representation formats implies a variety of ways of inferring. [2] Nevertheless, most of the discussion here developed concerns forms of concept-based inferences that are quite similar, since they are all supposed to use conceptual spaces as representational format. Furthermore, these inferences are also language-based, even if they do not build on syntax, but on our semantic intuitions.[3]

In this chapter, I discuss a type of reasoning that is not exclusively language-based, but builds on complex representational structures, external to the cognitive agent and whose use requires extensive prior instruction. When we think about reasoning in this sense, scientific thinking comes to mind as the paradigmatic example. Scientists train for years to make inferences involving highly abstract concepts mediated by complex informational structures like models, formulas, graphs, and/or computer simulations. Considering that scientific reasoning involves different strategies and mechanisms than everyday reasoning, it is important to see how it fits within a general theory of inference.

The classic literature on scientific reasoning —historically monopolized by philosophers— has mainly understood this process from the formalist perspective, with logic and probability as its central engines, and deductive inference as its paradigmatic case. In general terms, reasoning was pictured as an individual ability, depending exclusively on some "hardwired" cognitive mechanisms,

[2] In Section 3.3, I explained how this idea comes from a discussion in AI regarding the relation between the structure of information and the computational efficiency of the mechanisms that exploit them.

[3] The inferential mechanisms studied so far fall under what Boghossian calls —following the terminology of the dual-system view (Frankish, 2010)— "1.5 inferences" (Boghossian, 2014, 2018). That is, inferences which are fast and intuitive but still at the personal-level —i.e., the agent is conscious of why and how she draw them. In this chapter, we will discuss inferences that are typical of system 2, i.e., fully conscious, effortful, and resource consuming.

and specified by some set of domain-general rules operating over a sentence-like representational system. In recent decades, however, this view has been progressively abandoned by many cognitive scientists (Mercier & Sperber, 2017) and, to a minor extent, by philosophers (Clark, 2006; Hacking, 2009; Harman, 1986; Wartofsky, 1987).The main reason behind abandonment is that this highly idealized view of thinking neglects some relevant factors involved in reasoning, notably, its socio-cultural dimensions and its dependence on non-linguistic and external devices.

A good theory of reasoning should explain not only the internal processes taking place in the mind/brain of cognitive agents, but also how these processes are influenced by the socio-cultural context in which the agents are embedded and the role of tools and external devices that agents use while performing different (high-level) cognitive tasks. Following this direction, cognitive scientists have proposed other alternative theories to account for this "situated" character of reasoning. To name some: Mercier and Sperber's "argumentative theory" (Mercier & Sperber, 2017) which proposes that the key element for understanding the origins and function of reasoning is that of social interaction; Hutchins's "distributed approach", which sees cognition as a process distributed across people and artifacts, and dependent both in internal and external representations (Hutchins, 2010); or "the extended mind" theory, which proposes that high-level cognitive processes are driven, to a great extent, by elements that are external to the cognitive agent herself (Clark, 2006; Clark & Chalmers, 1998).[4]

This tendency was echoed in the philosophy of science —where the classical view was deeply rooted in— and has encouraged some philosophers to propose more realistic accounts of scientific thinking. Some notable examples are Hacking's "styles of reasoning" (Hacking, 1994), describing scientific reasoning as a multi-dimensional notion depending on domain-specific forms of reasoning that are socially and culturally developed; Nersessian and Magnani's work on model-based reasoning, which shows how scientific models are a central part of scientific reasoning abilities and practices (Magnani, 2004; Nersessian, 2010); or

[4]From a more fundamental perspective, all these theories are a reaction to the formalist view of high-level cognition promoted by the computationalist approach discussed in previous chapters.

Wartofsky's "constructive mentalism", which claims that understanding scientific cognition implies inquiring into the historical processes that give rise to the specific forms of culturally-situated cognitive practices (Wartofsky, 1987).

Before all these proposals, within the framework of a systematic critique of classical logic as an adequate theory of reasoning (see Toulmin, 1971, 2003), Stephen Toulmin developed an "ecological" approach that understands scientific thought through two central notions: "method of representation" (MR) and "inferential technique" (IT). For Toulmin, scientific reasoning cannot be explained by reducing it to a set of hardwired cognitive capacities working over amodal inputs, and independently of any socio-cultural context. On the contrary, it can only be understood as a socially-embedded activity, which depends on mastering different methods for representing information that make possible specific forms of inferences. Toulmin uses these two notions in his two most important works on the philosophy of science (Toulmin, 1953, 1972a), and they play a central role in his explanatory-based approach to science. However, he does not provide analytical definitions for them, nor does he use them systematically throughout his works.

In what follows, I intend to explore the notions of MR and IT and articulate them with Toulmin's procedural view of concepts and explanation. I will show that they constitute interesting analytical categories for studying scientific practice. In particular, they can shed some light on the complex relationship between scientific reasoning, models, and conceptual change. I will later explain how Toulmin's analysis fits the general definition of inference proposed in this work and gives some interesting insights into the complexity of scientific concepts and their relation with reasoning.

The rest of the chapter is organized as follows. In Section 8.2, MRs and ITs are characterized, and their relations to other central notions in Toulmin's philosophy of science, like explanation, conceptual use, and conceptual change, are explained. Section 8.3 applies the previous ideas to a case study from the history of mathematical physics. More precisely, it will be shown how the development of an explicit notion if instantaneous speed in physics was possible thanks to introducing a new MR that allowed for a new way of reasoning about

motion. Section 8.4 finally shows how Toulmin's ideas relate —and anticipated in some cases— some important notions from the contemporary philosophy of science like "models" or "model-based reasoning". At the end, the relations between model-based reasoning and inferential pluralism will be discussed.

8.2 Representational methods and inferential techniques

First, it is important to keep in mind that Toulmin's philosophical project was, to a big extent, a reaction to some foundational views in epistemology —notably Frege's program (Toulmin, 1972a, pp. 52-55)— that had a deep impact in philosophy of science at the beginning of the twentieth century. [5] These views —that Toulmin called "absolutists"— assumed that the job of philosophy was to look for the ahistorical and immutable principles of rationality that would underlay scientific knowledge. In particular, Toulmin was a fierce opponent of logical positivists, and as such, he was against formalist views of reasoning. According to him, logical positivism made the mistake of identifying scientific rationality with "logicality", that is, to suppose that "the rationality of a science could be explained in terms of the logical attributes of the propositional systems intended to express its intellectual content at one time or another" (Toulmin, 1974, p. 404). Toulmin thought, on the contrary, that scientific rationality was historically and culturally situated, and as such, dynamic in nature. In this sense, he proposed that elucidating the principles underlying a rational enterprise like science, required to dig into the complexities of the "intellectual ecology" that characterize it in different periods of times (Toulmin, 1972a, Ch. 4.3).

Secondly, Toulmin defended an explanatory-centered view of science in which its main role was to provide explanations of phenomena as a way for obtaining understanding—in contrast to the traditional way of thinking of science as allowing for prediction. In order to understand what scientists consider

[5] Toulmin discontent is related to some of the problems of Frege's views on inference and meaning discussed in Chapter 2 and 3.

"rational" in some specific period of time, philosophers have to analyze the "explanatory practices" used in scientific disciplines in that period. Toulmin thought that these practices were articulated around *ideals of natural order* —also called "explanatory ideals"—, that is, background explanatory structures that determine some "natural way" in which some class of phenomena is supposed to behave (Toulmin, 1961, p. 79). They establish what is the "normal"—and so the "expected"—behavior of the phenomena studied. When some phenomenon deviates from the ideal, it needs to be explained, and for doing that, scientists propose laws that encode specific explanatory procedures to account for them. For instance, one of the explanatory ideals which articulate classic geometrical optics is the principle of rectilinear propagation of light. Refraction is a phenomenon that deviates from the ideal, and as such it begs for an explanation. In this sense, Snell's law of refraction is an ad-hoc law which accommodates the deviant phenomenon to the background ideal of natural order.

Toulmin used to call his approach to explanation "procedural" because the focus was not in abstract patterns of arguments —like in the deductive-nomological model— but on concepts-in-use. He conceived conceptual use in science as depending on the mastering of standardized procedures for representing and modeling natural phenomena. For instance, possessing the concept of *refraction* from geometrical optics does not only involve knowing its abstract definition, but requires to master those symbolic techniques —geometrical and algebraic— used for representing and reasoning about cases of refraction. Most scientific concepts require, in order for someone to grasp them, mastering some of these techniques for representing information (Toulmin, 1972b, p. 161). In this sense, and because Toulmin thinks that concepts and conceptual systems are the fundamental units of science, understanding scientific practices —notably reasoning and explanation— requires to understand the MRs used and developed by scientific communities.

But, what exactly are MRs? While Toulmin used this notion extensively in his two most ambitious works on philosophy of science, he did not provide an analytical definition of it. Roughly, MRs are "intellectual techniques" that allow

scientists to construct and use models of target phenomena. In general, any standardized symbolic system that scientists use for representing phenomena —diagrams, pictures, mathematical formulas, computer programs, etc.— counts as a MR. Toulmin suggests that the crucial role that MRs play in scientific theorizing is related to how they make up for the representational limitations of natural language. He writes:

> ..."representation techniques" include all those varied procedures by which scien- tists demonstrate—i.e. exhibit, rather than prove deductively—the general relations discoverable among natural objects, events and phenomena: so, comprising not only the use of mathematical formalisms, but also the drawing of graphs and diagrams, the establishment of taxonomic "trees" and classifications, the devising of computer programmes, etc. (Toulmin, 1972a, pp. 162–163)

Some of the most important features of MRs are: (1) they are associated with the explanatory ideals of scientific disciplines; (2) they are generative, since they establish the rules and the symbolic resources that will constitute particular models that scientists will use to represent, understand, and reason about phenomena; (3) they are part of the "collective methods of thought" (Toulmin, 1972a, viii), and as such they are "communal" in nature —their use and "validity" depends on the agreement of the scientific community; (4) they play a central role in discovery; and (5) they are essential to scientific reasoning because they bring with them new ITs. In the remainder of the chapter, I will focus on (4) and (5) and their mutual relation.

Regarding the notion of "inferential technique," it can be roughly defined as the set of procedures that allow scientists to draw model-based inferences within the context of a particular MR. More specifically, when scientists try to explain some phenomenon, they represent it by building a model using some specific symbolic resources. Many of the inferences that scientists are going to make using this model, depend on procedures for manipulating these symbolic structures. As we will see, ITs are important because they challenge the classical

view of reasoning that builds on the formality thesis (see Toulmin,1972a, pp. 487–488; or Toulmin, 1953, p. 25).

Regarding the relation between (4) and (5), Toulmin thinks that MRs play a central role in discovery because they allow for the introduction of new ways of thinking about phenomena:

> The heart of all major discoveries in the physical sciences is the discovery of novel methods of representation, and so fresh techniques by which inferences can be drawn. (Toulmin, 1953, p. 34)

The above quotation only makes sense when we consider Toulmin's ideas regarding the "ecological" relation between conceptual use, representation, and inference. As suggested before, Toulmin's views about concept possession are also procedural. Scientific concepts are neither linguistic entities nor abstract ideas; instead, they obtain their content by being part of "communal" practices that involve mastering representational techniques for explanatory purposes. Scientific concepts are not abstract entities that agents somehow "grasp" —as it would be from a Fregean perspective. They are acquired when agents learn how to produce explanations by following the symbolic —and inferential— procedures established by the MR of the scientific discipline to which the concepts belong. Furthermore, Toulmin believes that this procedural character of concepts is not lost when scientists do their thinking "in their heads". But internalized thinking that tokens some scientific concept reflect the external symbolic procedures that characterize the MR (Toulmin, 1972a, p. 163).

Toulmin illustrates these ideas by analyzing the concept *light* in geometric optics. As with many other concepts, light has a scientific version and an everyday version of it. Propositions including *light* in the everyday sense, are not necessarily supported by a MR, and so their inferential role is different from the scientific version. In its everyday use, *light*'s inferential role is associated with concepts like *vision*, *shadow*, *darkness*, *color*, and so on. For instance, it can take part of inferences like "If there is not light in the room, I will not be able to see." But within the context of geometrical optics *light* shows a different inferential role due to its association with a MR. In particular, the introduction of the

principle of rectilinear propagation allows to model optical phenomena with geometrical methods and brings a "fresh way of drawing inferences" (Toulmin, 1953, p.25) based on the reading/manipulation of geometrical diagrams and other symbolic procedures. For instance, our reasoning about the length of a shadow cast by a wall depends on the construction of geometrical diagrams and the application of arithmetic and trigonometrical techniques to it. In other words, this kind of reasoning build on a set of model-based inferences that exploit formal properties of the model in order to draw some conclusion about the target phenomenon (Figure 8.1).

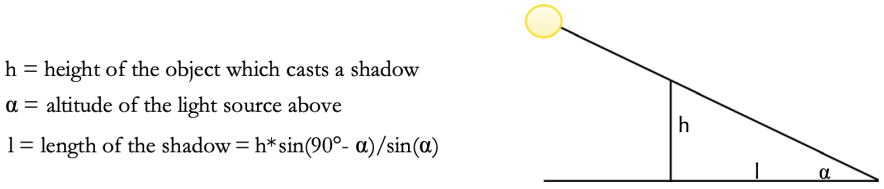


FIGURE 8.1: Geometrical diagram used for calculating the length of the shadow in basic geometrical optics.

These symbolic procedures are not heuristic tools that facilitate reasoning. They constitute the inferential technique of geometrical optics and, according to Toulmin, they cannot be fully translated to any language-based logical scheme:

> If the novel techniques of inference-drawing here used have not been recognized by logicians for what they are, that is probably because in geometrical optics one learns to draw inferences, not in verbal terms, but by drawing lines. (Toulmin, 1953, p. 26).

In this sense, the content of the scientific concept *light* becomes associated with the kind of procedures aforementioned:

> The view of optical phenomena as consequences of something travelling and the diagram-drawing techniques of geometrical optics are introduced hand-in-hand: to say that we must regard light as travelling is to say that only if we do so can we use these techniques to account for the phenomena being as they are. (Toulmin, 1953, p. 26)

These ideas bring Toulmin closer to an *inferentialist* view of (scientific) meaning (see Section 2.2). But with the peculiarity that for him, the inferences contributing to the meaning of scientific concepts are not propositional, but model-based. Furthermore, Toulmin's ideas regarding inference and MRs have as corollary that there is no unique —cross-disciplinary— form of scientific inference, but different ITs are constituted by situated practices involving different systems of representation and procedures of explanation. The central epistemic value for these practices is not only to explain the data but to understand phenomena in new and productive ways. Since different disciplines use different representational methods and explanatory ideals, ITs are diverse. In this way, Toulmin's model-based approach involves a form of inferential pluralism similar to the one defended in Chapter 3.

Now, coming back to (4), Toulmin seems to think about scientific discovery mainly as conceptual innovation—at the level of individual concepts or at the level of "conceptual populations" (Toulmin, 1971). In this sense, and because of the previously explained relation between concepts, inference and representation, conceptual change could involve changes at the level of the MR and its associated IT used within a scientific discipline. Toulmin endorses this idea in different ways throughout various of his works. However, he does not give any detailed example of it in the history of science. In what follows, I will analyze the development of the concept of instantaneous speed during the passage from geometrical physics to analytical mechanics using the analytical tools explained above. With this, I will try to illustrate how representational techniques are involved in conceptual development in science, and how scientific concepts and reasoning are interwoven with representational structures.

8.3 From Geometrical Physics to Mathematical Physics

Conceptual change has classically been understood as a problem circumscribed to the linguistic dimension of science. Philosophers of science have tended

to focus on language-related issues, like referential stability or the possibility of translations between successive theories, disregarding the relations between scientific language and other elements of scientific theorizing like models and reasoning.[6] Toulmin's approach to conceptual change is centered in the latter relation. For him, the "symbolic aspect" of scientific practice —the one related to concepts— comprises both natural language and MRs (Toulmin, 1972a, p. 163). In this sense, conceptual changes should be studied by analyzing the dynamic relation between MRs, concepts and procedures of explanation. I will apply these ideas to the study of the development of the notion of instantaneous speed during the passage from Galilean geometrical physics to analytical mechanics. For doing that, I follow two similar analyses of this episode by Michel Blay (1992; 1998) and Marco Panza (2002). This case study will allow me to make two points: one about the strong dependency relationship between the MR and reasoning, which exemplifies the relation between inference and representation discussed in Chapter 3; and a second one concerning the role of the MR in the development of concepts, a point which, I believe, is often overlooked in the literature on conceptual change in science.

8.3.1 Galileo's geometrical method

The MR that characterized Galileo's physics was based on two main mathematical tools: the "method of the configuration of qualities" (an application of two-dimensional geometry to kinematics) and the eudoxian theory of proportions. The first one was developed in the Middle-Ages following the work of Nicolas Oresme. Oresme's intention was to represent the variations of qualities (phenomena enduring in time like velocity, temperature or luminosity). According to this method, qualities are measured through its *intensio* and *extensio*. The intensio (rate of change of the quality) is measured in degrees represented as straight lines associated with different points in a horizontal line representing the extensio of the quality.

[6] Except for the cases of Nersessian (1999; 2010) and Thagard (1992).

For example, when we consider motion, *speed* is seen as an intensive quantity in relation to the extensive quantity time. At different points at the extensio line, speed has a correspondent degree, which can be represented by a perpendicular line. Oresme saw that the figures formed by delimiting the lines of extensio and intensio could represent different types of motion, and what is more, that the area of that figures was equivalent to the total space traversed by the body in motion (Clagett & Oresme, 1968, p.15). As Schemmel explains (2008, p. 65), this equivalence is due to the medieval idea that speed is space traversed in a specific period of time. In this sense, a uniform quantity may be represented by a rectangle, a *uniformly difform* quality by a triangle or a trapezium (see Figure 8.2) and a *difformly difform* quality by various kinds of irregular figures.

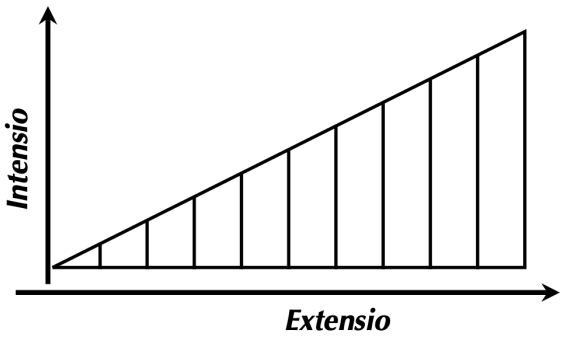


FIGURE 8.2: Representation of an uniformly difform quality, like, for example, uniformly accelerated motion.

Galieo's reasoning about motion was based on this very MR, [7] that implied a great deal of diagrammatic manipulation and visual thinking. Along with it, eudoxian theory of proportions was used as the tool for studying the mutual dependencies of magnitudes by comparing ratios. This last method was the precursor of the functional analysis of motion, which defines velocity as $v = \frac{\Delta s}{\Delta t}$. But Galileo could not arrive to this last definition because the theory of proportions included an "homogeneity constraint" establishing that only magnitudes of the "same kind" could be compared in a ratio (see Def. 3 and 4, Book V, *The Elements*). Since time and speed are represented by lines, and space by areas,

[7] There are, however, some differences concerning the interpretation of the extension of the quality in Oresmes and Galileo (see Palmerino, 2010; Schemmel, 2008, for a detailed explanation).

it was impossible to form ratios between them. Hence, Galileo could not define speed explicitly, but he had to use a more complex formula for expressing proportional relations between different quantities (Guicciardini, 2013) : $\frac{v_1}{v_2} = \frac{s_1}{s_2}.\frac{t_1}{t_2}$ (using contemporary algebraic notation).

This initial constraint for finding a simple definition to the notion of velocity clearly illustrates Toulmin's claims that the explicit definition of some concepts depends on the availability of a MR that supports their inferential role. In general, as various scholars have observed (cf. Giusti, 1994; Palmerino, 2010; Palmieri, 2003; Sellés, 2006), the MR that Galileo used allowed for the development of some concepts but also imposed serious constraints to the development of others:

> [W]hen a mathematical theory is chosen to describe the phenomena (but very often there is no freedom in this choice, and Galileo had only the theory of proportions at his disposal), the mathematical language will condition not only the manner of exposing, but also sometimes the way of conceiving the very nature of things, to the point that it is not always easy to separate what belongs to the author's thought from what is instead determined by the underlying mathematical theory, which organizes the phenomena of nature according to its own structures. (Giusti, 1994, p. 493, *my translation*)

Giusti's observation is perfectly aligned with Toulmin's ideas about how MRs set the limits of conceptual development in scientific practices.[8] As we just saw, Galileo's analysis of motion is seriously limited by the mathematical —representational— structure that guides his reasoning. In particular, because it prevents Galileo from reasoning with the notion of *continuity* and with a proper notion of instantaneous speed. As Blay explains (1992, pp. 133-151), within the framework of geometrical physics —the tradition of Galileo, Descartes and Newton— there was an operational, yet non-explicit, notion of instantaneous speed. And it was necessary to wait until the development of

[8]Following Roux (Roux, 2010, p. 3), we can say that Galileo's case shows how mathematical language is not "conceptually neutral".

analytical mechanics to have the representational tools that will make possible one.

Galileo used the notion of *degree of speed* as an informal tool for capturing the general idea of instantaneous speed. We can see this notion operating in Galileo's proof of the Mean Speed Theorem, which establishes a relation between uniform motion and uniformly accelerated motion. The Theorem I, Proposition I of the *Dialogue* expresses the following:

> [T]he time in which any space is traversed by a body starting from rest and uniformly accelerated is equal to the time in which that same space would be traversed by the same body moving at a uniform speed whose value is the mean of the high- est speed and the speed just before acceleration began. (Galilei, 1954 [1632], p. 173)

Galileo's reasoning in the proof is built around the diagrammatic representation of a particular case (see Figure 8.3):

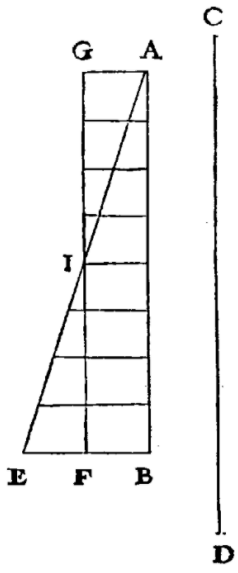


FIGURE 8.3: Diagrammatic representation used for reasoning about the mean speed theorem (Galilei, 1954 [1632], p. 173).

The line AB represents the time in which a body travers the space CD falling from rest from point C. All the lines parallel to EB represent the degrees of speed at different instants starting form A, while EB itself represents the highest value of speed gained during AB. F bisects EB and FG is drawn parallel to AB until

reaching GA, which is parallel to FB. The formed figures AGFB and AEB represent the two different forms of motions mentioned in the above quotation. And their relation—which is the central point of the theorem—is determined by studying the geometrical properties of the fig- ures that represent them. In particular, the areas of each figure is assumed to be equal to the infinite "aggregate" (*totidem velocitatis momentum*) of degrees of speed. Since both areas are equal, then the equality of both overall speeds is inferred. And as a consequence, Galileo concludes that the space traversed by each motion is also the same.

This proof clearly illustrates Toulmin's ideas regarding how scientific reasoning depends upon representational techniques.Galileo's reasoning is clearly model-based, in the sense that it exploits formal properties of the model —according to the rules established in the MR— in order to make inferences about the phenomenon represented. One of the central points of Galileo's reasoning is —as it was just said— the idea that the area of a geometrical figure is made up of an aggregate of an infinity of lines (see Clavelin, 1968, p. 316), and in that way the geometrical model represents the increase of momentum in time. This is so due to the representational properties of the model, since geometrical entities are considered continua like momentum and time (see Ducheyne, 2008, for a detailed explanation). A similar thing happens with the notion of degree of speed, that tries to capture the idea of instantaneous speed. Galileo did not have an explicit definition of instantaneous speed, because within the framework of the MR he used for analyze motion, speed was considered as an intensive magnitude increasing by "successive additions of degrees" (Blay, 1998, p. 72), so his reasoning does not use an explicit definition of instantaneous speed but it depends entirely on the formal properties of the MR. This was also the case in Oresme's proof of the same theorem. As Panza writes:

> Though it explicitly deals with speed as an instantaneous (or punctual) quality, the proof of this theorem is not founded on any explicit definition, either of speed in general, or of instantaneous (or punctual) speed. It simply works because of a diagrammatic formalism

> associated to the metaphysical idea of speed (or quality, in general). (Panza, 2002, p. 260).

8.3.2 Towards an analytical method of representation

In the *Principia*, Newton takes a big step towards the development of the formal notion of instantaneous speed, and with that, towards the mathematization of motion.[9] That implied the development of a new MR that will play a crucial role in the conceptual development of analytical mechanics. *Grosso modo*, Newton developed the concept of *motion* within the framework of *dynamics*, this is, explaining motion in relation to the notion of *forces as cause*. According to Newton, motion is a relation between two magnitudes represented by a curve, which can be described within the framework of Cartesian geometry. The use of analytical methods in geometry for analyzing motion implied a substantial advance in relation to the Galilean MR. Still, Newton was tied to geometrical ideas when reasoning about motion. As Panza emphasizes:

> As long as motion is, for Newton, a mathematical object, it is essentially a geometrical one. [...] One of the aims of analytical mechanics in 18th century is that of trans- forming the Newtonian science of motion in an analytical science, i.e. to pass from a geometrisation of the science of motion to a new theory of motion where the latter is just an analytical object. (Panza, 2002, p. 263)

Regarding the notion of instantaneous speed, as Blay showed, Newton worked with an"operative" notion of it, but without any explicit definition. For example, in the proof of the Proposition V, Theorem III, in the *Principia I*, he uses the idea of instantaneous speed with a variable segment which represents the speed in an instant as the areas of the figures under the motion curve (see Figure 8.4). The reasoning is still geometrical, but it's also based on an infinitesimal idea that he cannot really represent formally: supposing that

[9] It is common in the literature to talk about mathematization in Galileo, but I believe, following Blay and Panza, that there is an important difference between *geometrization* of motion in the tradition of Galileo–Descartes–Newton and the kind of *mathematization* in the development of analytical mechanics.

the area can be divided into innumerable equals intervals (see Blay, 1992, pp. 135–137).

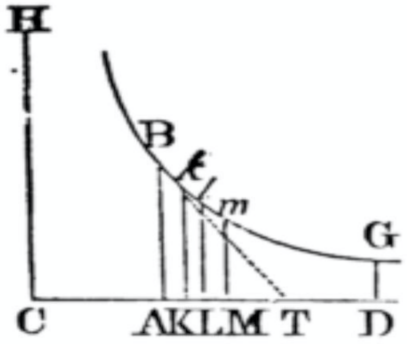


FIGURE 8.4: Proposition V, Theorem III, Principia I (Newton, 1999 [1687], p. 642).

However, since an explicit definition of the notion of instantaneous speed is not at disposal, it is impossible to explicitly reason with that concept. And this could be easily seen as a conceptual limitation due to the MR in use. And as Panza said (2002, p. 246), this limitation is not an expository choice, but a consequence of the very method that Newton was using to reason about motion.

Another central change made by Newton was to eliminate the metaphysical interpretation of motion as a quality (ibid., 263). That implied a big advance in the explanatory structure behind the Galilean interpretation of motion, and it will have a deep impact on the MR, notably because the new —analytical— notion of motion will not be affected by the homogeneity constraint of the theory of proportions.

The crucial step for leaving behind the geometric MR for understanding motion, as well as the conceptual limitations that came with it was taken by Pierre Varignon. His interpretation of the concept of speed at each instant, which is valid for both rectilinear and curvilinear motion, was a crucial step towards the consolidation of a new MR in physics: analytical mechanics. Varignon built on Leibniz's *differential* —using the method developed by l'Hôpital in *Analyse des infiniment petits* (1969)— in order to define the velocity at instant t as valid for every infinitesimal interval of time dt. His argument was straightforward: since $t + dt \approx t$, speed does not vary and so $v(t) \approx v(t + dt)$. This allowed Varignon

to give an explicit algebraic definition of instantaneous speed as: $v = \frac{ds}{dt}$, as well as of other crucial functional relations like $dv = \frac{d^2s}{dt}$ and $y = \frac{dv}{dt}$ ("force of instantaneous acceleration").

Using Toulmin's analytical tools again, we can see in this example how the explicit definition of the notion of instantaneous speed depended on the MR in use; clearly illustrating his idea regarding the role of MRs in conceptual development. The mathematical language used in physical theories —and more generally, the symbolic structure used in a MR— is not "conceptually neutral" (cf. Roux, 2010), but quite the opposite. In fact, it is not a "language" in which we can express any content, or any "abstract idea" but, in many cases, it is the condition of possibility of the very emergence of the content of these ideas and of their proper systematic use within a particular IT. Furthermore, with the new MR of analytical mechanics comes a whole new IT for reasoning about motion: algebraic-based reasoning instead of geometrical case-based reasoning. As Blay explains:

> [T]he figure, essential to the development and the organization of the geometric- infinitesimal thinking, gradually becomes simple diagram with Varignon. This is to say, that the figure loses its traditional value of intellection in order to take a secondary value of mere illustration. ((Blay, 1992, p. 16), *my translation*).

Furthermore, one of the main advantages of the new algebraic-based IT is that it allows reasoning with generality, and is not tied to any particular figure like in the Galilean-style of physics. As Panza explains:

> This representation enables Varignon to eliminate any constraint of homogeneity, since it allows him to compare spaces, times, speeds and accelerative forces by means of a comparison of segments [...]. These identities make the solution of a number of cinematic and dynamic problems independent of a geometric analysis of the data are expressed by means of suitable equations, the solutions can also be expressed by other equations derived by using the algorithm of the calculus. (Panza, 2002, p. 265)

In summary, this case study illustrates the two main features of MRs. First, they are involved in conceptual change and conceptual use.Galileo's way of understanding motion imposed serious limits to the development of an explicit notion of instantaneous speed. However, he managed to reason with an operational version of this concept by exploiting the geometrical properties of his MR. Furthermore, as it was explained above, proposing an explicit notion of instantaneous speed required the development of a new MR for thinking about motion. Secondly, this case study shows how scientific reasoning relies heavily on representational techniques. As it was explained before, the two MRs studied —the geometrical and the analytical one— provide their own symbolic resources to build models for analyzing motion. These models work as inferential devices that make possible different patterns of inferences, and in this sense, different "styles of reasoning" about motion in physics.

8.4 Models, Model-based Reasoning, and Inferential Techniques

Toulmin's notion of IT was a pioneering attempt to propose a more realistic alternative to the traditional view of scientific reasoning based on the formality thesis. However, looking beyond its historical merit, one could argue that Toulmin's view is not useful today because it is redundant with the new "environmental" perspectives that proliferate in contemporary philosophy of science. Particularly with the notion of model-based reasoning developed mainly by Nersessian (1999; 2010), Giere (1999; 2004; 2010) and Magnani (2002; 2004).[10] I will argue that, on the contrary, Toulmin's approach is not redundant but complementary to the model-based reasoning approach (MBRA). For that, I will briefly compare these two views in order to show their strong coincidences and their relevant differences.

[10]Toulmin's ideas on representation, specially those developed in his 1972 book, are also a clear antecedent of Suarez's inferential view of scientific representation (Suárez, 2004). However, discussing this particular relation exceeds the scope of this article.

The first common point concerns the role of the notion of "model" in these two views. Both Toulmin and the MBRA consider models as the key element for understanding scientific practice, emphasizing their role in scientific reasoning, concept formation, and discovery in general (Giere, 2004; Morrison & Morgan, 1999; Nersessian, 1999).

One important account of models within the MBRA was developed by Giere (1999). He defends a "representational" view of models [11] which sees them as the structures which enable us to 'access' phenomena, playing a central role in scientific theorizing. They are not just tools for interpreting how mathematical formulae and theoretical principles connect with actual phenomena; but they are the very target of these principles and formulae. Scientists can speak of mathematical structures as describing phenomena thanks to the supposed representational properties of models.

In Giere's view, models are constructed according to theoretical principles (like Newton's laws, Darwin's principles, etc.), which at the same time gain meaning thanks to models, since we cannot understand them literally as laws of nature. These principles work as "general templates" (Giere, 2004, p. 745) for the construction of models:

> [S]cientists generate models using principles and specific conditions. The attempt to apply models to the world generates hypotheses about the fit of specific models to particular things in the world. Judgments of fit are mediated by models of data gener- ated by applying techniques of data analysis to actual observations. Specific hypoth- eses may then be generalized across previously designated classes of objects. (Giere, 2004, pp. 60-61)

Toulmin also understands models as representational. However, his central unit of analysis of science is not strictly models but MRs, this is —again— the generative techniques and procedures that underlies the construction of particular models. In Toulmin's view, the epistemological dimension is center

[11] Against what he calls an "instantial view", which understands models as instantiations of the axioms and mathematical structures of theories and focuses on the problem of truth and reference in the relationship between models and target phenomena.

stage. Explanation, conceptual use and reasoning are all interrelated aspects of a model-based practice whose central goal is to provide understanding. MRs, hand in hand with ITs, constitute new and productive ways of thinking about already known phenomena, and in that way, they provide scientific understanding:

> By making the journeys (inferences) so licensed, the physicist finds his way around phenomena: by thinking of the systems he studies in terms of appropriate models, he sees his way around them and comes to understand them. (Toulmin, 1953, p. 104)

Toulmin, as Giere, does not see theoretical principles as laws of nature that could be literally interpreted as talking directly about phenomena. They gain their meaning only in association with MRs and explanatory procedures that influence the construction of models that allow us to reason (indirectly) about phenomena:

> If the layman is told only that matter consists of discrete particles, or that heat is a form of motion, or that the Universe is expanding, he is told nothing or rather, less than nothing. If he were given a clear idea of the sorts of inferring techniques the atomic model of matter, or the kinetic model for thermal phenomena, or the spherical model of the Universe is used to interpret, he might be on the road to understanding; but without this he is inevitably led into a cul-de-sac. (Toulmin, 1953, p. 39)

It seems fair to see Toulmin as one of the pioneers of the representational view of models, and in particular, his approach is very close to Giere's, but —I believe— with some additional virtues. Notably, it does not understand models as abstract and independent objects but as part of a situated, cognitive practice connecting understanding, inference, and explanation. In Giere's view, these elements are (supposedly) connected, but he does not specify how, while Toulmin's procedural view of concepts and explanation does.

Coming back to reasoning, within the MBRA, models also play a central role in it, and it's at this point where the deep similarities between the MBRA

and Toulmin's approach become clear. Nersessian (1999) defines "model-based reasoning" as an inferential process that involves the construction and manipulation of various kinds of representations, not mainly sentential and/or formal, but associated with multiple formats of representation of information. As she explains:

> In model-based reasoning, inferences are made by means of creating models and manipulating, adapting, and evaluating them. A model, for my present purposes, can be characterized loosely as a representation of a system with interactive parts and with representations of those interactions. Model-based reasoning can be performed through the use of conceptual, physical, mathematical, and computational models, or combinations of these. (Nersessian, 2010, p. 12)

ITs are the model-specific patterns of reasoning that characterize the cognitive practice of scientists working within a specific discipline which use a specific MR. These patterns are rule-based procedures that make use of multi-modal elements of the model and follow the explanatory schemes involved in the MR. It is easy to see that ITs and model-based reasoning are deeply similar notions, but still, they are not exactly the same.

The main difference between these two notions is that the study of model-based reasoning is generally oriented to the specification of some domain-general cognitive mechanisms that are involved in the cognitive manipulation of models —for example: analogical reasoning, mental modeling, manipulative abduction (Magnani, 2011, Ch. 3), or visual reasoning. While ITs, on the other hand, are concerned with how the different model-specific symbolic systems, schemes of explanations, and concepts interact in order to conform to a particular historically and socially-situated "procedure of reasoning".[12] In this sense, Toulmin's analytical tools could be useful for studying the procedures of reasoning that are not the direct product of our "hardwired" cognitive capacities, but those which

[12]In this sense, inferential techniques are close to Hacking's notion of "style of reasoning" (Hacking, 1994).

depends upon the collective use of normative symbolic systems, like scientific language and scientific systems of representation.

8.5 Conclusion

In the previous chapters, we have focused on the organization of concepts in agents' background knowledge. That is, we have remained "inside" the mind/brain of reasoners. However, as argued in chapter 3, conceptual information can also be distributed in symbolic structures external to the cognitive agents. Since these structures' primary function is to support reasoning, an explanation of how they are used in inference is required. In this chapter, a possible explanation in this sense was proposed focusing on scientific models, i.e., hybrid representation systems whose use require a lot of previous training.

I revisited two notions of Toulmin's philosophy of science that are a central part of his procedural theory of concepts and explanation. I showed that Toulmin's ideas parallel some important new currents in philosophy of science regarding models, conceptual change, and symbolic reasoning. And that they point in the same direction as some new trends in cognitive science that understand reasoning as socially and culturally situated. I explained in what specific sense MRs could be considered as constitutive of scientific inference and how these elements are of central importance to conceptual development in science. I further showed how MRs (and models) in science play more than a mere representational role, since they are central to inference. It is in this sense that I speak of MRs as inferential devices, this is: symbolic systems that, when correctly manipulated, allow users to arrive to conclusions that could not be inferred otherwise. Furthermore, since inferential practices in science are specific of MRs, the diversity of MRs across scientific disciplines implies a diversity of inferential practices. This point could be developed as an argument in favor of some forms of scientific pluralism, especially those forms associated with the notion of style of reasoning (Ruphy, 2011). But specially, this last point is clearly an argument in favor of inferential pluralism, as it was described in Chapter 3.

The discussion about models in science has advanced tremendously since Toulmin (see Frigg & Hartmann, 2020). However, I devoted this chapter to him because I believe that he pointed out to some connections between concepts, models, and inference that are generally overlooked in philosophy. In particular, Toulmin's ideas opened up for a way of seeing concepts that was rather unusual in the field. He took scientific concepts "out of the mind" of scientists, and situated them within the context of socially regulated procedures of representation, manipulation, and action. He emphasized the normative character of scientific concepts, claiming that they are "*intellectual micro-institutions*" (Toulmin, 1972a, p. 166); but without overlooking their psychological complexity.

Chapter 9

Summary and concluding remarks

In this last chapter, I summarize what I consider the main results of this dissertation and point out some future directions in which the approach outlined here can be developed.

The previous eight chapters tried to challenge, on different fronts, the idea that a psychologically plausible theory of reasoning can build on the formality thesis. In Chapters 1 and 2, I have analyzed the philosophical origins of this idea and its influence in psychology and semantics. Chapter 3 consisted of a theoretical discussion on the relationship between inference and representation, which motivated a pluralistic approach for theorizing about reasoning. Chapter 4 introduced Conceptual Spaces, the theoretical and formal framework that I consider most suitable to explicate the role of concepts in reasoning. Chapter 5 showed that material inferences —which are not accountable from a logicist perspective— can be modeled with CS. Chapters 6 and 7 extended these ideas to nonmonotonic reasoning and category-based induction. Two CS-models of these inferential mechanisms —developed in collaboration with Peter Gärdenfors— were presented. Finally, Chapter 8 studied how model-based inference builds on conceptual information encoded in external and hybrid symbolic structures.

What have we learned? One of the main points made by this dissertation is that logical formality should not be given a psychological interpretation within an explanation of reasoning. Those theories endorsing the formality thesis are ill-equipped to explain an overwhelming amount of data showing that reasoning

is sensitive to content. In addition, they have various theoretical limitations, such as not being able to account for several inference-types that are intuitively correct —or that count as intuitively rational inferential moves— but formally invalid.

While classical logic fails as a model of everyday reasoning, some might argue that it can be "enriched" in different ways to accomplish this task.[1] As a matter of fact, the model presented in chapter 6 can be understood in this sense. However, notice that these attempts generally involve the negation of the formality thesis and, in particular, of the idea that conceptual content does not play a role in inferential validity. As said before, a central aim of the analysis in Chapter 6 was to show how integrating a model of conceptual knowledge to the classical treatment of nonmonotonic inference can solve various foundational and epistemological issues of this approach.

Does the denial of the formality thesis imply the assumption that reasoning is not formal in any sense? Not really. The approach defended here shows that several inferential mechanisms can be explicated with CS. Consequently, it assumes that reasoning can be specified through some mathematical —"formal"— structures. Still, this kind of *formality* is not equivalent to *logical formality*, since it does not require us to accept that inferential validity responds only to the truth-functional structure of natural language sentences. In contrast, instead of accepting that inference pivots in the truth-functional structure of propositions, the CS-model assumes that inference pivots on formal properties of conceptual structure.

The other crucial point emerging from this thesis concerns the explanatory advantages of giving to semantics a central place in the theorization about reasoning. In particular, when *meaning* is understood as *conceptualization* — as proposed by cognitive semantics— one can naturally address the elusive problem of the role of knowledge in reasoning, as well as the largely neglected relationship between *understanding* and *inference*.

The model presented in Chapter 7 illustrates this last point. Category-based

[1] Relevance logic, default logics, and description logics are some examples of such attempts.

induction is an inferential mechanisms that directly builds on our understanding of concepts and their interrelations. As such, it shows what Evans (1989) suggested and this dissertation tried to prove: *No concepts, no understanding. No understanding, no inference.*

Finally, as said at the end of Chapter 7, I believe that the fact that CS can naturally model several types of inferential mechanisms shows that it is a promising framework for a unified theory of reasoning.

Inferential pluralism

Another view defended in this dissertation was *inferential pluralism*, i.e., the idea that from the diversity of representational structures encoding conceptual information, it follows a diversity of inferential mechanisms exploiting them. Inference is plural at different levels. Chapters 5, 6, and 7 showed how different inference-types exploit different properties of conceptual representation. For instance, nonmonotonic reasoning and CBI make essential use of concepts' prototypical structure, while material inferences only require *core* conceptual knowledge. Furthermore, from the fact that different word classes have different underlying representational structures (Gärdenfors, 2014), it follows that they should also have different associated inference-patterns —even if they can be all represented in CS.

In a different level of analysis, when we get "out of the head" of cognitive agents, we can see how inference is plural in the diversity of (external) representational devices used both in everyday and scientific cognition. The analysis developed in Chapter 8 shows some possible ways to understand the complex relationship between external models and rational inference.

Future work

Many of the ideas presented in this dissertation open up new research lines and complement others that already exist. First, if the preliminary model of material inference presented in Chapter 5 is correct, it should be possible to extend it to most word classes with inferential properties. Verbs and sentences expressing events are good candidates for continuing this line of work since

there are some available CS-model of them (Gärdenfors, 2014; Gärdenfors, Jost, & Warglien, 2018; Warglien, Gärdenfors, & Westera, 2012). Furthermore, I believe that the CS-model of material inferences could help the study of sense-relations in lexical semantics. Some initial suggestions in this line were advanced in Section 5.4.3 regarding semantic compatibility and co-hyponymy. Finally, this framework could also be used as an explication of the kind of semantic modulation affecting the process of model-construction in the MMT. It seems to me that this could be an interesting research line to develop.

There is a lot of work to be done building on the ideas presented in Chapter 6. In particular, our model of expectation-based inference (Osta-Vélez & Gärdenfors, n.d.) can be adapted for the analysis of *prototypical reasoning*, a kind of inference that has recently attracted attention in the Computer Science and AI (e.g., see M. Lewis & Lawry, 2016; Lieto, Minieri, Piana, & Radicioni, 2015; Lieto & Pozzato, 2018). It remains to be seen how it performs compared to the other alternative models, or if they can be complementary in some way. In addition, the CS-model should be extended beyond expectations related to object properties. For instance, future work may focus on nonmonotonic reasoning about events, actions, or intentions. Furthermore, it could be interesting to use this model to analyze the relation between expectations and vague concepts (Douven et al., 2013); as well the relation between expectations and generic statements (Cimpian et al., 2010).

Concerning CBI, the CS-model needs to be also extended to better account for "property-effects" in inductive reasoning. In particular, to effects associated to properties describing causal relations. Some suggestions in this sense were advance in Chapter 7, but they need to be further developed and formally implemented to see if they are able to explain the available data (e.g., Bright & Feeney, 2014; Rehder & Hastie, 2001). Moreover, this model can be empirically tested. Some methodological ideas on how to do that were also discussed in the chapter mentioned above.

Lastly, I believe that the analysis in Chapter 8 can be developed in various ways. In particular, the claim that the *concrete* procedures of model-manipulation are constitutive of scientific reasoning could find some grounding

in some recent views about situated conceptualization in cognitive science.

A hint on how to bridge them can also be found in Toulmin. He wrote:

> After a time, no doubt, any experienced scientist begins to do much of his thinking "in his head", just as we all of us learn to do elementary arithmetic "in the head"$\cdots$ our "internalized thinking" conforms to the same arithmetical, zoological or physical procedures, and criteria of "correctness", as the thinking we do overtly or out loud. (Toulmin, 1972a, p. 163)

While vague, his suggestion points to the idea that, instead of using abstract rules or inner language, scientific thinking continues to conform to symbolic manipulation procedures used with the concrete models.

These ideas relate to recent views of complex thinking as involving perceptual and sensorimotor simulations of concrete situations, like Barsalou's "situated conceptualization" (Barsalou, 2003; Barsalou, Santos, Simmons, & Wilson, 2008) or the "perceptual account of symbolic reasoning" (Landy et al., 2014). For instance, this latter approach conceives symbolic thinking as a special kind of embodied mechanism in which mathematical formulas are "internalized" by simulating the external procedures by which agents manipulate the symbols (Ibid., p. 4).

There have been some attempts to combine this discussion in philosophy with cognitive science (e.g., see Carruthers, Stich, & Siegal, 2002; Nersessian, 2006). However, as far as I know, none of them have used the above-mentioned theories. I believe that the insights in Chapter 8 can be naturally connected to situated accounts of conceptualization, possibly enriching our understanding of how scientific reasoning works.

Appendix A

Résumé détaillé en Français

Le raisonnement et les concepts sont deux sujets centraux en philosophie et en psychologie cognitive. Curieusement, ils ont été traités comme des thèmes de recherche indépendants dans la littérature. Cela est particulièrement déconcertant si l'on considère le large consensus sur le rôle essentiel de ces deux notions dans l'explication de la cognition. Historiquement, les concepts ont été conçus comme les "éléments constitutifs" de la pensée. En même temps, le raisonnement est un type de transition entre des pensées censé guider l'action et la fixation des croyances dans les agents rationnels. À première vue, les notions semblent être intrinsèquement liées. La question est donc de savoir pourquoi les théories du raisonnement, tant en psychologie qu'en philosophie, évitent la notion de concept dans leur structure explicative.

Une explication possible de cette situation est que les théories du raisonnement ont été dominées par une approche logiciste qui voit l'inférence comme un processus purement formel-syntaxique (i.e., non sémantique) qui s'appuie sur un ensemble de règles générales et indépendants du contenu (*topic-neutral*). De ce point de vue, les concepts lexicaux sont considérés comme non pertinents pour le processus d'inférence rationnelle. Cela n'est pas du tout fortuit; au contraire, cela répond à une interprétation spécifique de la notion de "forme logique" qui a dominée la logique depuis Aristote (see Etchemendy, 1983). Dans cette optique, la validité inférentielle est une question de forme, et non de contenu. En d'autres termes, les inférences déductives sont valides en vertu de ses structures logiques, indépendamment de la relation entre les termes (*concepts*) extra-logiques dans les prémisses et la conclusion.

Cette conception, synthétisée dans l'affirmation d'Inhelder et Piaget selon laquelle "le raisonnement [humain] n'est rien d'autre que le calcul propositionnel lui-même". (1958, p. 305), a encouragé les psychologues et les philosophes cognitifs à considérer le raisonnement déductif comme le paradigme de l'inférence rationnelle, et à concevoir le contenu sémantique comme sans intérêt pour l'explication du raisonnement.

Parallèlement à cela, la sémantique philosophique a été dominée par une vision da la signification qui confirme la déconnexion entre les concepts et le raisonnement. Dans une large mesure, les philosophes croyaient pouvoir expliquer ce qu'était la signification lexicale sans introduire la notion d'inférence — ou toute autre notion faisant référence à un mécanisme cognitif. La sémantique était alors pensée comme quelque chose exclusivement concernée par la relation entre le langage et le monde. Dans ce sens-là, elle est réduite aux notions de *référence* et *conditions de vérité*.

Je suis persuadé que ces idées sont fondamentalement erronées et qu'il n'y a pas moyen d'expliquer le raisonnement sans concepts et vice versa. Cette thèse est une tentative de justifier cette conviction.

Je ne suis pas le premier à affirmer cela. Jonathan Evans, une figure centrale dans le domaine de la psychologie du raisonnement, a écrit que les concepts et l'inférence sont "inextricably entangled" (Evans, 1989, p. 29). Il était convaincu que *connaissance* —C'est-à-dire que le "corps des concepts" est constitutif du processus même du raisonnement; et que les théories psychologiques doivent en tenir compte. En particulier, il a affirmé que le raisonnement ne pouvait pas être "aveugle", mais qu'il nécessitait un certain degré de compréhension du sujet en question. Étant donné que la compréhension suppose la possession de concepts, il ne peut pas avoir du raisonnement sans concepts.

Dans la même optique, Hugo Mercier et Dan Sperber ont récemment développé une théorie générale du raisonnement qui vise à expliquer ses dimensions sociale, individuelle et évolutive (Mercier & Sperber, 2017). L'idée fondamentale est que les mécanismes inférentiels sont censés exploiter les régularités empiriques de l'environnement qui sont codifiées dans des systèmes de représentation. Leur théorie est essentiellement anti-formaliste, et comprend

l'inférence comme *inextricablement enchevêtré* avec la représentation. Cependant, elle n'explique pas comment cet enchevêtrement fonctionnerait, ni comment les informations conceptuelles sont structurées au sein dans systèmes de représentation. En fait, —et encore une fois— leur approche néglige la notion de *concept* (see Osta-Vélez, 2019).

Cette thèse défends —par des stratégies diverses– des idées qui sont similaires à celles mentionnées ci-dessus. Je me focalise sur le point négligé par la théorie de Mercier et Sperber, à savoir, la question relative à l'imbrication des mécanismes d'inférence avec les structures de représentation. En particulier, je soutiens que l'inférence exploite différentes propriétés des systèmes de représentation utilisés par la cognition humaine pour coder les informations conceptuelles.

Notez que cela ne contredit pas les affirmations des logicistes. Ceux derniers pensent que l'inférence déductive exploite la forme logique, et que la forme logique est une propriété (implicite) du langage naturel (i.e., un système de représentation). Cependant, je crois que l'inférence logique joue un rôle plutôt marginal dans la cognition de haut niveau, et que la plupart des inférences basées sur le langage dépendent des propriétés de la représentation sémantique. Maintenant, qu'est-ce que la "représentation sémantique"? Suivant une tradition en sémantique cognitive, et par opposition aux approches conditionnelles de la vérité, je suppose que ce type de représentation fait référence aux structures mentales évoquées par les concepts lexicaux lors du traitement du langage (language processing).

L'un des objectifs principaux de cette thèse est de développer cette dernière idée en détail. J'utilise la théorie des espaces conceptuels de Peter Gärdenfors (2000; 2014) pour expliquer différentes formes d'inférences basées sur la sémantique de façon telle qu'elles correspondent à la définition donnée ci-dessus.

Les espaces conceptuels sont un programme de recherche en science cognitive et en représentation des connaissances qui affirme que le contenu conceptuel est organisé en différentes structures topologiques et géométriques à un niveau sous-symbolique de représentation de l'information. Il fournit de nombreux outils formels et théoriques pour expliquer comment les concepts sont utilisés

dans les processus cognitifs tels que la catégorisation, l'induction, la formation de concepts ou l'apprentissage des langues.

Dans cette thèse, j'ai l'intention de montrer comment cette approche peut être utilisée pour expliquer le rôle des concepts dans le raisonnement. En ce qui concerne l'inférence basée sur la sémantique du langage naturel, un résultat de l'analyse développée ici sera que les classes de mots ont des patterns inférentiels spécifiques qui leur sont associés en raison de leur dépendance à différents espaces conceptuels. Par ailleurs, on montrera comment l'inférence inductive et le raisonnement non monotone reposent sur des propriétés spécifiques des structures conceptuelles comme la similarité et la typicalité.

La notion de inférence défendue ici est *pluraliste*. L'idée sous-jacente est simple: la cognition humaine utilise de nombreux types de systèmes de représentation différents. Le langage naturel est, sans doute, le plus important. Cependant, nous pensons également avec des images mentales et une pléthore de structures symboliques externes comme des diagrammes, des formules mathématiques et des modèles scientifiques. Ces systèmes codifient également les connaissances conceptuelles, et nous les utilisons pour raisonner par le biais de mécanismes inférentiels qui exploitent des propriétés qui leur sont propres. Pour appuyer cette affirmation, le dernier chapitre de cet ouvrage est consacré à l'étude du raisonnement scientifique basé sur des modèles et, en particulier, à la relation entre les concepts, les modèles et le raisonnement scientifiques.

Avant d'expliquer la structure de ce travail, il est important de noter deux choses. Premièrement, le contenu de plusieurs chapitres a déjà été publié sous la forme d'articles. En particulier, les chapitres 6 et 7 s'appuient sur deux collaborations avec Peter Gärdenfors (Osta-Vélez & Gärdenfors, n.d., 2020a); tandis que le chapitre 8 est basé sur (Osta-Vélez, 2019). Deuxièmement, la thèse se déroule en deux étapes. Les trois premiers chapitres offrent une analyse critique du cadre historique et philosophique qui motive ce travail. Ces chapitres sont censés être le "ciment" qui relie le reste du contenu. En revanche, les cinq derniers chapitres sont plutôt constructifs et proposent différentes façons d'explique et modéliser les questions abordées dans cette thèse.

La thèse formaliste

Le chapitre 1 porte sur la *thèse formaliste*, une idée centrale pour les théories classiques du raisonnement qui affirme que l'inférence est un processus formel effectué sur un "langage de la pensée" qui compte avec une structure propositionnelle. Le but de ce chapitre c'est de montrer comment cette idée est basée sur la distinction conceptuelle entre *forme* et *contenu* héritée de la logique classique.

La logique a été traditionnellement conçue comme une théorie abstraite du raisonnement. Elle a joué un rôle essentiel dans le développement de sciences cognitives et, en particulier, dans des théories du raisonnement (Harman, 1984; Henle, 1962). Elle a été utilisé comme modèle de compétence pour le raisonnement déductif; comme cadre normatif pour évaluer nos performances dans les tâches de raisonnement (Osherson, 1975b; Stenning & van Lambalgen, 2011); et comme outil méthodologique pour modéliser la structure formelle des opérations cognitives de haut niveau (Piaget, 1957). En général, les différentes formes sous lesquelles la logique a influencée l'étude du raisonnement au fil des ans partagent une hypothèse sous-jacente: les propriétés logiques sont des propriétés formelles-syntaxiques et, si l'inférence humaine est "logique," alors elle doit être également "formelle."

Cette idée s'inscrit dans une longue tradition qui consiste à concevoir l'esprit comme une machine/ordinateur. Elle est née avec Thomas Hobbes, et a été développée par George Boole, Charles Babbage, Alan Turing, Warren McCulloch et Walter Pitts, et Jerry Fodor —entre autres–, jusqu'à devenir l'un des paradigmes centraux de la science cognitive (Boden, 1988; Gigerenzer & Goldstein, 1996). La thèse générale qui sous-tend ce point de vue est qu'un ensemble d'opérations formelles/mathématiques constitue la base de la cognition humaine. Si nous pouvions les comprendre *via* les algorithmes mathématiques appropriés, alors le raisonnement pourrait être formellement expliqué, et éventuellement reproduit par un dispositif non biologique.

Cependant, l'idée que le raisonnement peut être formel dans un sens logique n'est pas exactement équivalente à l'idée que la cognition peut être décrite par

un formalisme mathématique particulier. La première thèse s'applique exclusivement à l'inférence rationnelle et elle revendique que la logique classique peut spécifier les mécanismes qui la sous-tendent. En revanche, la seconde s'applique à tout processus cognitif quel qu'il soit et elle n'est pas engagée à une structure mathématique particulière comme modèle.

Étant donné que nous nous concentrons sur la formalité logique, la question est de comprendre les origines de cette notion. Selon MacFarlane (2000), la formalité logique remonte à l'utilisation de la distinction aristotélicienne entre forme et contenu pour analyser le raisonnement et l'argumentation. En gros, cette tradition affirme que le raisonnement a des propriétés à la fois formelles et matérielles, mais que la "validité" déductive est une question de forme, et non de contenu.

Cet approche (appelé "hylomorphique") a été renforcé lors de la *mathématisation* de la logique grâce aux travaux de Boole et Frege (see, Van Heijenoort, 1967), conduisant à ce que Warren Goldfarb a appelé la *conception schématique de la logique* (Goldfarb, 2001). Selon ses mots, d'après la conception schématique :

> ...the subject matter of logic consists of logical properties of sentences and logical relations among sentences. Sentences have such properties and bear such relations to each other by dint of their having the logical forms they do. Hence, logical properties and relations are defined by way of the logical forms; logic deals with what is common to and can be abstracted from different sentences. (Goldfarb, 2001, p. 26)

L'idée c'est que il y a des formes logique "cachées" dans les énoncées du langage natural; des relations structurels entre ces formes est ce qui permet les mouvements inferentiels entre premise(s) et conclusion. Dans la tradition schématique ces formes son déterminées par les constantes logiques, i.e., des particules linguistiques sans contenu conceptuel. Dans ce sens-là, les prédicats n'ont aucun "puissance inférentielle".

En résumé, la thèse de la formalité trouve ses sources dans l'idée que la validité déductive dépend exclusivement de la distribution des constantes logiques dans les propositions de langage naturel, et non du contenu des prédicats impliqués dans ces propositions.

La thèse formaliste dans les sciences cognitives

L'un des principaux défis pour la psychologie cognitive est de fournir une explication scientifique d'un phénomène à la fois *intentionnel* et matériel (Horst, 1999a). En particulier, les pensées sont des entités intentionnelles avec du contenu sémantique – ils sont *à propos* ("about") quelque chose– articulées de manière non arbitraire. Le raisonnement est un cas spécifique de transition entre des pensées qui préserve (idéalement) de la cohérence sémantique. Si l'on veut donner une explication empirique de ce processus, la question centrale est alors, comment le raisonnement est-elle mécaniquement (matérielment) possible? (see Rescorla, 2012).

Jerry Fodor a vu qu'une façon de répondre à cette question était de recourir à une interprétation psychologique de la formalité logique (voir Fodor, 1975; 1987; 2008; 2015). En gros, il affirmais que la psychologie doit expliquer l'esprit comme une machine syntactique effectuent des opérations formelles sur des entités semblables à des propositions linguistiques —pensées—, avec des propriétés à la fois syntactiques et sémantiques. Les transitions causales entre les pensées sont possibles grâce à leurs propriétés syntaxiques. Comme ces propriétés *reflètent* le contenu sémantique des pensées, (Fodor, 1985, p. 93) la pensée rationnelle est également possible. Comme l'explique Fodor :

> ...you connect the causal properties of a symbol with its semantic properties via its syntax. The syntax of a symbol is one of its higher-order physical properties . To a metaphorical first approximation , we can think of the syntactic structure of a symbol as an abstract feature of its shape. Because, to all intents and purposes, syntax reduces to shape, and because the shape of a symbol is a potential determinant of its causal role, it is fairly easy to see how there could

> be environments in which the causal role of a symbol correlates with its syntax. It' s easy, that is to say, to imagine symbol tokens interacting causally in virtue of their syntactic structures . The syntax of a symbol might determine the causes and effects of its tokenings in much the way that the geometry of a key determines which locks it will open . (Fodor, 1987, pp. 18-19)

Cette idée a eu un profond impact dans la psychologie du raisonnement au XXème siècle. La deuxième partie du chapitre 1 montre comment la thèse formaliste a façonné trois théories différentes dans cette discipline: La théorie du développement cognitif de Piaget ; la théorie de la logique mentale, et la théorie des modèles mentaux.

La thèse formaliste dans la philosophie du langage

En même temps que la logique excluait la sémantique de l'analyse de l'inférence, les philosophes commençaient à penser à la "signification" comme quelque chose de complètement déconnecté de la cognition. En particulier, des philosophes comme Quine, Carnap et Putnam promouvaient l'idée que la signification linguistique était quelque chose qui appartenait exclusivement à la relation entre le langage et le monde. De cette façon, ils ont promu l'idée que la philosophie du langage n'avait pas besoin de coopérer avec une théorie de l'inférence et vice versa.

Par exemple, Quine pensait qu'aucune théorie systématique du langage ne pouvait émerger de l'association de la notion de signification à des entités mentales ou abstraites comme les idées ou les intentions. Pour lui, les notions d'extension et de référence étaient suffisantes pour analyser la signification des termes extra-logiques dans le langage naturel. Notons que ce dernière idée implique qu'on peut construire une théorie sémantique sans l'aide de théories psychologiques ou cognitives.

En résumé, les extensionalistes pensaient que la sémantique était complètement indépendante des théories de l'inférence. En ce sens, ils ont également suivi la thèse formaliste : une théorie de l'inférence doit être axée sur les constantes

logiques, tandis que la sémantique concerne tous les éléments extra-logiques du langage naturel. Jerrold Katz appelle ça le “dogme extensionaliste”:

> The article of faith is that there exists a justifiable distinction between the logical and nonlogical components of sentences, one that enables us to divide a theory of connectives and quantifiers from a theory of the meaning of nouns, verbs, adjectives, etc., that form the expressions and sentences they connect and quantify. (Katz, 1975, p. 77)

Comme l’a montré Katz (Ibid.), cette dernière idée a de nombreuses limitations. En général, toute théorie sémantique qui déconnecte la signification de la cognition aura plusieurs problèmes pour expliquer des phénomènes que sont directement liés à la sémantique, comme par example la communication, la compréhension linguistique, la catégorisation et le changement conceptuel. Cela a conduit les philosophes à élaborer des théories de la signification plus compréhensives. La tentative la plus importante a était la "sémantique du rôle conceptuel", qui affirme que les significations des mots émergent du rôle des concepts lexicaux —et des attitudes propositionnelles— dans “l’écologie cognitive complexe des agents”(Block, 1986; Brandom, 1998b; Harman, 1982).

Même si la sémantique du rôle conceptuel est une tentative intéressante de faire le lien entre l’inférence et la signification, je soutiens dans le Chapitre 2 que cette théorie n’est pas assez systématique pour expliquer la relation entre la signification et le raisonnement. Au lieu de cela, je propose la Sémantique Cognitive comme une alternative prometteuse pour accomplir ce tâche. La sémantique cognitive est un programme de recherche dont l’objectif principal est de développer une analyse du langage en tant que système d’information qui sert de médiateur dans notre interaction avec le monde grâce à la coopération de plusieurs facultés cognitives telles que la perception, la catégorisation, le raisonnement et la mémoire (voir Geeraerts & Cuyckens, 2007, p. 5).

Contrairement à la sémantique traditionnelle, focalisée sur la signification propositionnelle et la vérité, la tradition cognitive prend la signification lexical (lexical meaning) comme central. Les notions de *référence* et *vérité* jouent un

rôle relativement marginal dans sa structure théorique. L'idée principale de la sémantique cognitive peut être résumée par la devise suivante: "meaning is conceptualisation". En d'autres termes, le traitement sémantique implique la mobilisation constante des structures de connaissance pour décoder la signification lexicale et propositionnelle. Comme Langacker l'explique: "*Semantic structure is conceptualization tailored to the specifics of linguistic convention. Semantic analysis therefore requires the explicit characterization of conceptual structure*" (Langacker, 1987, p. 99). Les éléments constitutifs de ces structures de connaissances - ou comme dirait Quine, les "véhicules da la signification"- ne sont pas des propositions, mais des notions comme les prototypes conceptuels, les cadres (frames) ou les "image schemes" de Lakoff et Johnson.

Représentation et inférence

La notion d'inférence est au centre de plusieurs domaines de recherche. Par exemple, elle fait partie du répertoire conceptuel de la philosophie, les sciences cognitives, l'informatique, l'IA, et les statistiques. Le problème est que cette notion est si fondamentale qu'on l'utilise souvent sans la définir. Au chapitre 3, je propose une définition de la notion d'inférence dans le cadre d'une perspective pluraliste.

Les inférences sont des transitions entre des états mentaux qui participent à ce que l'on pourrait appeler — selon la terminologie de William James– "the stream of thought". Il est évident que ce *stream of thought* ne se limite pas aux transitions inférentielles. Nous sommes tous familiarisés avec différents types d'associations mentales —entre des croyances, perceptions, souvenirs, etc. — que nous ne considérerions pas comme *inférientielles*. Par exemple, je peux avoir une disposition personnelle à penser à des guitares chaque fois que je vois un bateau; ou me souvenir de la maison de mon enfance quand je pense à des chats; mais es transitions mentales ne peuvent pas être évaluées de façon normative car elles ne répondent à aucune logique spécifique, et elles ne semblent suivre

aucun critère informationnel. Les transitions inférentielles, en revanche, sont censées satisfaire ces deux derniers points.

La tradition dominante en philosophie considère les inférences comme des transitions entre des jugements avec une forme linguistique, qui suivent des règles logiques. Parmi les nombreux problèmes que pose cette tradition, je n'en analyse que deux : (1) elle n'inclut pas de contraintes informationnelles sur le type d'inférences qu'il est raisonnable de tirer d'une certaine information ; et (2) elle a une notion très limitée de la représentation. Dans ce qui suit, je développerai brièvement ce deuxième point.

Le "conservatisme représentationnel"

La philosophie analytique a été fidèle au schéma explicatif suivant pour la pensée et le raisonnement: quel que soit le sujet, les unités de pensée sont des croyances, et les croyances sont des entités avec structure propositionnelle et propriétés logiques —sémantiques et syntaxiques. Le raisonnement, *modulo* thèse de la formalité, consiste en des transitions entre les croyances générées par des mécanismes qui exploitent leur structure syntaxique. Comme il est évident, c'est l'idée fondamentale qui sous-tend la théorie computationnelle de l'esprit (CTM) examinée dans le premier chapitre. Nous allons maintenant la revisiter, en nous focalisant sur son usage de la notion de représentation.

La CTM (en particulier Fodor et Pylyshyn (Fodor & Pylyshyn, 2015)) défends un approche "conservateur" du format de représentation: toute représentation psychologique est basée sur un système amodal qui ressemble un langage, la *langue de la pensée* (LOT) ou *mentalase* (Fodor, 1975, 2008). Selon Fodor, LOT est "the only game in town" (Fodor, 1975, p.55), c'est-à-dire la seule hypothèse plausible pour construire une psychologie scientifique, parce que le LOT a les bonnes propriétés pour expliquer la *productivité* et la *systématicité* de la pensée rationnelle. En gros, toutes les différentes modalités de l'information –visuelle, auditive, tactile, etc. – que le cerveau traite pour alimenter les mécanismes cognitifs doivent être traduites dans le LOT afin d'être utilisées dans la cognition de haut niveau.

Le principal problème du conservatisme représentationnel est qu'il n'explique pas comment l'information est organisée dans la mémoire sémantique. En conséquence, cette perspective ne peut pas résoudre le "problème du cadre" (*frame problem*, c'est-à-dire la question de savoir comment extraire les connaissances pertinentes d'une riche data-base dans des contextes de résolution de problèmes. Par exemple, disons que je dois décider comment aller de chez moi au cinéma. La plupart des informations contenues dans ma data-base ne sont absolument pas pertinentes pour répondre à cette question; il doit donc y avoir un mécanisme de recherche d'informations qui sélectionne parmi ce corpus diverse les éléments qui sont utiles pour la tâche – par exemple, dans la catégorie *moyens de transport*, des concepts comme *train*, *bus*, *voiture*, ou *vélo*.

Je défends l'idée que pour résoudre le problème du cadre, il faut postuler l'existence d'une structure sub-linguistique de représentation des informations qui a comme des concepts comme des unités de base, en lieu des croyances, comme l'indique la tradition susmentionnée. Plus précisément, j'affirme que le raisonnement a besoin de l'interaction d'informations qui sont explicitement représentées dans le langage – ou dans un autre format de représentation externe –, avec des informations implicites et codifiées dans une structure de représentation sub-symbolique au sein de la mémoire sémantique. Je propose que cette structure sub-linguistique peut être modélisée en utilisant la théorie des espaces conceptuels de Peter Gärdenfors.

Les espaces conceptuels

La théorie des Espaces Conceptuels (EC) (Gärdenfors, 2000, 2014) est un programme de recherche en sciences cognitives visant à modéliser plusieurs phénomènes cognitifs liés à la conceptualisation — par exemple, le l'apprentissage, le raisonnement, la catégorisation, la formation de concepts, etc. Contrairement à la tradition computationaliste prédominante en sciences cognitives, EC ne part pas du principe que la pensée est basée sur un langage mental —comme le LOT. Au contraire, EC es fondeé sur l'hypothèse qu'il existe

un système représentationnel intermédiaire qui codifie l'information sémantique avec une structure spatiale.

EC s'appuie sur deux notions fondamentales: *dimension qualitative* (DQ) et *domaine*. Les premières sont les éléments constitutifs des concepts. Elles représentent les différentes *qualités* des objets qui servent de base pour juger des similitudes entre différents stimuli (Gärdenfors, 2000, p. 6). Par exemple, "hauteur" est une DQ des stimuli auditifs; en se concentrant sur la hauteur, on peut comparer et classer différents sons. Les DQ sont diverses, elles peuvent être innées, culturellement acquises, phénoménales ou abstraites selon le concept.

Un point fondamental est que las DQ peuvent être représentées par différentes structures géométriques (voir Gärdenfors, 2000, Chapitre 1). Par exemple, le poids et la hauteur peuvent être tous deux représentés par une ligne isomorphe aux nombres réels non négatifs. D'autres DQ ont une structure discrète et correspondent à des qualités qui sont représentées sous forme d'ensembles disjoints.

Les DQ peuvent être *intégrales* ou *séparables*. Elles sont intégrales lorsqu'il est impossible d'attribuer à un objet une valeur dans une dimension sans lui attribuer une autre valeur dans une autre dimension (see Maddox, 1992). Par exemple, nous ne pouvons pas représenter un son avec une hauteur spécifique mais sans valeur pour son intensité sonore. En revanche, certaines DQ peuvent être représentées indépendamment les unes des autres, comme *height* et *wealth* lorsque l'on pense aux personnes. Dans ces cas, on parle de dimensions *séparables*. Les dimensions intégrales sont souvent modélisées avec une métrique euclidienne, tandis que les dimensions séparables avec une métrique "city block".

Un ensemble de dimensions intégrales qui sont séparables de toutes les autres dimensions est appelé un *domaine*. L'exemple classique d'un domaine est le "espace du couleur". Il est composé de trois dimensions intégrales: *ton*, *saturation*, et *luminosité*. La représentation géométrique du ton est le cercle chromatique. La saturation ou l'intensité est représentée comme un intervalle de la ligne réelle, tandis que la luminosité varie du blanc au noir et est donc une dimension linéaire avec des points terminaux. Ensemble, ces trois dimensions

intégrales, une à structure circulaire et deux à structure linéaire, constituent l'espace couleur (voir figure 4.2, Chapitre 4).

Les domaines d'un espace conceptuel sont reliés de plusieurs façons puisque les propriétés des objets modélisés dans les espaces co-varient. Par exemple, dans l'"espace des fruits", les dimensions de *maturité* et *couleur* co-varieront, ainsi que *taille* et *poids*. Ces co-variations sont le support de différentes procédures inférentielles qui exploitent les propriétés conceptuelles, comme nous le verrons au chapitre 6.

Un *espace conceptuel* est défini comme une collection d'un ou plusieurs domaines avec une fonction de distance —a *métrique*— qui représente les propriétés, les concepts, et leurs inter-relations de similarité. La similarité entre des concepts peut être facilement estimée puisqu'il s'agit d'une fonction monotone décroissante de leur distance dans l'espace (Shepard, 1987). Dans ce cadre, les concepts sont compris comme une sous-région d'un certain espace conceptuel.

Pour illustrer ces idées avec un exemple, considérons un l'EC du concept pomme, qui serait un sous-ensemble du produit Cartésien des domaines de la couleur, du goût, de la forme, de la maturité et de la texture. Cet espace s'étendrait à travers certaines régions de chacun de ces domaines —ceux qui représentent les propriétés communes des pommes—, tout en laissant d'autres régions "intactes" – par exemple, nous ne représentons pas les pommes de forme pyramidale, ou étant noires, donc ces propriétés ne sont pas couvertes dans l'espace conceptuel. Le concept *pomme* a plusieurs corrélations entre ses propriétés : le degré de douceur et d'aigreur, ainsi que la texture, sont corrélés au niveau de maturité. Ces corrélations peuvent également être représentées dans l'espace conceptuel par différents outils mathématiques (voir la figure 4.3, Chapitre 4).

Dans les chapitres 5, 6 et 7, je propose différents modèles de processus inférentiels qui s'appuient sur la théorie des espaces conceptuels.

Les inférences matérielles modélisés avec des EC

Dans le chapitre 5, je montre comment la notion d'inférence matérielle, proposée par Wilfrid Sellars et Robert Brandom, peut être développée et modélisée avec la théorie des espaces conceptuels. Dans ce qui suit, je vais expliquer les points centraux de mon argumentation.

Wilfrid Sellars était l'un des plus importants critiques de l'approche formaliste de l'inférence. Dans plusieurs de ses travaux, il a essayé de montrer que la plupart de nos inférences dans le langage naturel ne sont pas formelles au sens logique du terme, mais qu'elles sont "matérielles". En quelques mots, une mouvement inférentiel est *matériel* lorsqu'il est basé sur une relation conceptuelle entre les prédicats de la prémisse et de la conclusion. La validité matérielle n'a rien à voir avec la forme logique. Elle ne dépende pas de l'organisation des constantes logiques, mais elle est liée à la façon dont les concepts sont articulés dans une pratique inférentielle structurée de façon normative.

Deux exemples classiques d'inférence matérielle sont "Fido est un chien, alors Fido est un mammifère" et "Munich est au sud de Berlin, donc Berlin est au nord de Munich". En gros, Sellars pense que notre pratique inférentielle est surtout "matérielle", en raison du rôle que les règles qui sous-tendent ces inférences — "règles matérielles"— jouent dans la construction et l'utilisation des concepts dans le langage et la pensée. Selon Sellars, sans règles matérielles tenant compte de la façon dont les prédicats sont liés, aucune analyse logique ou philosophique du langage ne serait exacte. On pourrait dire que les inférences matérielles peuvent être analysées comme des enthymèmes (inférences avec des prémisses implicites), mais Sellars montre que cette stratégie échouerait.

Le problème chez Sellars (et Brandom) c'est que l'origine de ces règles matérielles d'inférence, et la manière dont les agents les appliquent dans le cadre d'un raisonnement au niveau personnel, reste inexpliqué. En général, une analyse approfondie des fondements psychologiques de l'ingérence matérielle manque dans le cadre inférentialiste. Sellars partent plutôt du principe que cet type d'inférence n'est qu'une question de suivre des règles. Brandom va même

plus loin en affirmant que toute théorie sémantique doit être précédée d'une théorie pragmatique. Dans la section 5.2.3, je critique cette dernière idée en montrant qu'il n'est pas possible d'éliminer les facteurs cognitifs (non pragmatiques) lors de l'explication du contenu conceptuel. En particulier, je montre que l'existence de contraintes cognitives et/ou de mécanismes d'apprentissage innés derrière la formation des concepts doit être une hypothèse inévitable dans la plupart des théories psychologiques des concepts. Cela suggère qu'une explication du contenu sémantique ne peut être développée en termes purement pragmatiques, mais doit tenir compte des processus cognitifs "câblés" qui sous-tendent l'acquisition des concepts.

Classes de mots et types d'inférences matérielles

Le modèle proposé ici commence par analyser les types d'inférences matérielles en fonction des types d'éléments lexicaux. Je m'appuie sur la classification standard des *catégories lexicales* utilisée par les linguistes qui distinguent entre *nouns*, *adjectifs*, *verbes*, et *prépositions* —entre autres– (Baker, 2003). Je suivrai l'analyse de Gärdenfors qui montre que les différentes catégories lexicales ont des structures de représentation différentes du point de vue de la théorie des espaces conceptuels (Gärdenfors, 2014).

De plus, mon analyse se base sur une hypothèse théorique concernant la relation entre l'attention et l'inférence. Je propose qu'un mécanisme cognitif central derrière les inférences matérielles es le "re-profiling" (Langacker, 1987). Selon les termes de Langacker, cela c'est un déplacement attentionnel au sein d'une base conceptuelle qui produit des transformations sémantiques minimales. Je soutiens que les inférences matérielles sont des cas de propositions de re-profilage au sein de leurs structures conceptuelles correspondantes. Par exemple, une inférence comme $chien(x) \rightarrow mammifere(x)$ consiste à "reprofile" l'objet x dans l'espace conceptuel de *chien* vers l'espace conceptuel de *mammifère*, ce qui est implicite dans la représentation du chien parce que le premier est une sous-région du second.

Voyons maintenant comment modéliser certains cas d'inférences matérielles

avec des EC. Commençons par discuter des inférences matérielles avec des noms ("nouns"). Selon le modèle de l'espace conceptuel, les noms correspondent à des concepts et, en tant que tels, ils constituent une région convexe dans un espace conceptuel: i.e., un sous-ensemble de l'espace-produit de l'ensemble des dimensions qui constituent l'espace. À partir d'un concept M, $\mathscr{C}(M)$ correspond à un sous-ensemble du produit cartésien des domaines n. Comme expliqué dans le chapitre 4, un objet x catégorisé comme M correspond à un point à n-dimensions $x =< x_1, x_2, \ldots, x_n >\in \mathscr{C}(M)$ avec x_i les coordonnées du point dans chaque dimension.

Un type fréquent d'inférénce materielle avec des noms c'est "bottom-up", où le concept dans la prémisse est un *subordonné* du concept sur la conclusion, comme dans $Chat(x) \rightarrow Mammifere(x)$. Du point de vue de l'espace conceptuel, la validité matérielle de ce type d'inférence réside dans un simple relation ensembliste. Comme nous l'avons dit plus haut, lorsqu'une entité x est classée comme étant N, elle est représentée comme un point dans $\mathscr{C}(N)$. Ainsi, pour tout concept M tel que $\mathscr{C}(N) \subseteq \mathscr{C}(M)$, si $x \in \mathscr{C}(N)$ alors $x \in \mathscr{C}(M)$. En outre, comme la relation d'inclusion est transitive, $\mathscr{C}(N)$ sera inclus dans tous les concepts supérieurs de M. Ainsi, le fait de classer x comme N en fera (par défaut) un membre de chaque concept supérieur de N.

Avec la même logique, on peut expliquer les inférences avec la négation. Considérons un ensemble des catégories $M_1, \ldots, M_n$, co-hyponymes de N, qui vont être représentés comme des sous-régions disjointes des $C(N)$. Cela signifie que $M_i \cap M_k = \emptyset$ pour tout $i \neq k$. Puisque l'expression "$M_i(x)$" est représentée par un objet $x \in M_i \subseteq \mathscr{C}(N)$, alors $x \notin M_k \subseteq \mathscr{C}(N)$ puisque M_i et M_k sont des sous-régions disjointes de $\mathscr{C}(N)$. Ce simple fait ensembliste justifie toutes les inférences matérielles de la forme "$M_i(x) \rightarrow \neg M_k(x)$" pour chaque catégorie M_k co-hypothèse de M_i. Normalement, les types biologiques sont de bons exemples de co-hyponymes dans ce sens. Par exemple, représenter l'énoncé "$Chien(a)$" implique matériellement "$\neg Chat(a)$", "$\neg Pigeon(a)$", "$\neg Aigle(a)$", et ainsi de suite, pour toute catégorie animale, des co-hyponymes de *chien* et au même niveau conceptuel. Ce type d'analyse est étendu dans le chapitre 5 aux adjectifs, aux relations de parenté et aux prépositions spatiales.

Les déductions matérielles sont sans incertitude. Elles dépendent de notre connaissance des relations sémantiques de base entre les concepts. Cependant, la plupart de nos raisonnements sont soumis à l'incertitude. Dans le chapitre 6, je propose un modèle d'inférences non monotones basé sur des espaces conceptuels.

Inférence non monotone et expectatives

Issu d'un article en collaboration avec P. Gärdenfors, le chapitre 6 est consacré à la relation entre les inférences non-monotones et les représentations conceptuelles. Le problème est, comme a été expliqué précédemment, que le raisonnement quotidien repose sur plus que la forme logique des prémisses explicites. L'information partielle et l'incertitude sont omniprésentes chez la pensé humaine. En conséquence, nos mécanismes inférentiels peuvent difficilement se permettre d'être "conservateurs" par rapport à l'information disponible. Au lieu de cela, nous utilisons systématiquement nos connaissances de base de manière risquée, mais néanmoins productive, pour donner un sens à notre environnement. En d'autres termes, le raisonnement quotidien est fortement non monotone, et les approches formalistes basées sur la logique classique ne peuvent pas expliquer cela.

Cette utilisation des connaissances de base s'exprime notamment à travers nos expectatives sur le monde. Par exemple, si nous savons qu'une personne vient de France, nous attendons d'elle qu'elle parle français et qu'elle ait un passeport français ; ou si nous conduisons une voiture et que nous apercevons une personne qui attend sur le bord de la route, nous attendons d'elle qu'elle ait l'intention de la traverser. En général, nos expectatives vis-à-vis du monde sont cruciales pour guider notre raisonnement et notre action dans la vie quotidienne, et elles s'appuient directement sur la structure de nos connaissances de base.

Gärdenfors et Makinson (1992; 1994) ont montré qu'une grande partie de la logique non monotone est réductible à la logique classique, à l'aide d'une

analyse des expectatives fonctionnant comme des prémisses occultes dans les arguments. L'idée directrice est que lorsque les gens essaient de savoir si une conclusion C découle d'un ensemble de prémisses$P$, les informations de base utilisées ne contiennent pas seulement les prémisses dans P, mais aussi des informations sur ce qu'ils expectent dans la situation donnée, de sorte qu'ils se retrouvent avec un ensemble plus large d'hypothèses. Ces expectatives peuvent être exprimées comme des hypothèses "par défaut", c'est-à-dire des énoncés sur ce que les raisonneurs représentent comme normal ou typique. Elles comprennent nos connaissances conceptuelles de base, mais aussi d'autres informations qui peuvent être considérées comme suffisamment plausibles pour servir de base à une inférence, pour autant qu'elles ne donnent pas lieu à des incohérences.

La principale différence entre les expectatives et les prémisses explicites c'est qu'elles sont "plus défaisables". En d'autres termes, si l'une des expectatives est en conflit avec certaines des prémisses explicites du P, nous ne les utilisons pas pour déterminer si le C découle du P. Toutefois, lors de l'évaluation de leur rôle dans le raisonnement, il est important de noter qu'ils n'ont pas tous la même force. En bref, nos expectatives sont toutes défaisables, mais elles présentent des degrés variables de défaisabilité.

En bref, la position défendue ici est qu'une bonne explication du rôle des expectatives dans le raisonnement doit s'appuyer sur un modèle de la structure des connaissances de base. Comme nous l'avons vu au chapitre 3, même si cette dernière notion a joué un rôle central dans plusieurs domaines de la philosophie et de la logique au cours des dernières décennies, peu d'efforts ont été faits pour la définir correctement. La logique classique non monotone a toujours travaillé dans le cadre formaliste : elle suppose que les connaissances implicites et explicites sont représentées sous forme de propositions dans une sorte de *belief-box* de l'agent cognitif. Le problème est que l'origine des règles par défaut et leur utilisation dans le raisonnement quotidien restent inexpliquées. Dans ce qui suit, nous verrons comment une bonne articulation des EC en tant que modèle de l'inférence non monotone peut indiquer une solution.

Donnons maintenant une structure formelle à ces idées. Un concept M représenté dans un espace n-dimensionnel $\mathscr{C}(M)$ est un ensemble convexe des

points représentant des objets possibles tombant sous M. Si M a un prototype, on suppose qu'il correspond à l'un de ces points : un point à n-dimensionnel $p^M =< p_1^M, p_2^M, \ldots, p_n^M >\in \mathscr{C}(M)$. L'idée centrale est que nos expectatives s'articulent autour de ce prototype. En d'autres termes, si la seule chose que nous savons sur x est qu'il tombe sous M, nous nous attendrons à ce qu'il soit — proche de p^M, c'est-à-dire qu'il ait toutes les propriétés du prototype.

Maintenant, nos expectatives par rapport à l'énonce "$M(x)$" vont au-delà des propriétés spécifiques déterminées par le prototype. Ils s'étendent à toutes les propriétés possibles qu'un objet tombant sous M peut avoir. Dans le cadre de la théorie des EC, cela signifie que la représentation d'un objet sous un concept M implique que l'objet peut occuper toute position possible dans $\mathscr{C}(M)$. Des positions différentes impliquent des propriétés différentes pour l'objet. Les propriétés qui ne s'appliquent pas à p^M peuvent être considérées comme des expectatives secondaires, car elles sont plus faibles —plus défaisables— que celles qui s'appliquent à p^M. En général, pour toute propriété non typique dans $\mathscr{C}(M)$, son *degré de défaisabilité* sera une fonction positive de sa distance par rapport au prototype.

Nous pouvons construire un ordre des propriétés qui reflète leur " degré de expectative " —et donc, leur degré de défaisabilité– en fonction de leur distance relative au prototype. Une façon de le faire est de mesurer la distance par rapport au point le plus proche où la propriété n'est pas satisfaite. Nous pouvons utiliser la fonction de distance pour obtenir ce type d'information à partir de l'espace conceptuel avec le critère suivant :

Typicality criterion (TC) Étant donnés les domaines D_i et D_k dans l'espace conceptuel $\mathscr{C}(M)$, pour deux propriétés quelconques R_i, R_k, tels que $R_i \subseteq D_i$ et $R_k \subseteq D_k$; R_i est plus typique que R_k sii s'il y a un point $x =< x_1, x_2, \ldots, x_i, \ldots, x_n >\in \mathscr{C}(M)$ avec $x_i \in R_i$, et pour tous les points $x' =< x'_1, x'_2, \ldots, x'_k, \ldots, x'_n >\in \mathscr{C}(M)$, $x'_k \in R_k$, il détient $d(x, p^M) < d(x', p^M)$

Pour voir un exemple, considérez l'espace conceptuel du fruit (Chapitre 4)

qui a comme des dimensions la couleur, le goût, la forme et la texture. Si l'on nous dit que a est une pomme, nos expectatives maximales seront que a a les propriétés d'une pomme prototypique: rouge, sucrée, ronde et lisse. Mais ces propriétés ont des degrés de typicité différents, même si elles sont toutes présentes dans le prototype. Par exemple, une pomme sucrée est plus typique qu'une pomme rouge, car il est plus surprenant de trouver une pomme non sucrée qu'une pomme non rouge. Cela signifie que les points représentant des pommes non rouges vont être plus proches du prototype que les points représentant des pommes acides ou amères dans l'espace conceptuel. De même, l'amertume est une propriété atypique pour les pommes, certainement moins attendue que le fait d'être jaune . Ainsi, les pommes jaunes seront plus proches du prototype que les pommes amères. Un classement des propriétés attendues pour la pomme peut donc ressembler à ceci : $Exp(Pomme) = \{ronde > rouge > sucr > lisse > vert > \cdots > jaune > \cdots > amer > \ldots\}$.

Le critère de typicité produit un ordre d'expectatives qui permet de comparer les propriétés individuelles. Cela résout le problème de l'origine de l'ordre d'expectatives qui a été proposé par Gärdenfors et Makinson, et permet de connecter la théorie des espaces conceptuels à la logique non monotone. Par ailleurs, on montre dans ce chapitre que cette stratégie peut également expliquer l'origine et les forces relatives (degrés de défaisabilité) des règles par défaut dans la logique par défaut. Pour finir, il est montré comment cette approche offre une nouvelle solution à le "problème de Linda" (aussi connue comme "l'erreur de conjonction", voir Tversky and Kahneman (1983)).

L'induction et la représentation des catégories

L'induction basée sur des catégories (CBI, par "category-based induction") est un mécanisme inférentiel qui exploite notre connaissance des relations conceptuelles pour estimer la probabilité qu'un propriété soit projeté d'une catégorie à une autre. Au cours des dernières décennies, les psychologues ont identifié plusieurs caractéristiques de ce mécanisme, et ils en ont proposé différents

modèles formels. Dans le Chapitre 7, un nouveau modèle mathématique basé sur les distances dans les espaces conceptuels est proposé. On montrera que ce modèle basé sur les EC peut prédire la plupart des propriétés de CBI, faire quelques nouvelles prédictions et fournir une base théorique solide pour ce phénomène psychologique. À la fin du chapitre, les relations avec d'autres modèles sont examinées, ainsi que certaines considérations méthodologiques.

Dans son article pionnier "Jugements inductifs sur les catégories naturelles", Lance Rips (1975) a analysé un type particulier d'inférence inductive qui exploite les informations sur les catégories individuelles (et sur les relations entre les catégories) pour estimer la probabilité de projection de propriété parmi elles. Par exemple, l'inférence "Les *chiens* ont des os sésamoïdes, donc les *loups* ont des os sésamoïdes" repose sur les similitudes conceptuelles entre les catégories *chien* et *loup*, et non sur la forme logique de l'argument ou sur une autre propriété codifiée de manière propositionnelle. Ces processus sont des formes intuitives de raisonnement fondamentales pour notre vie cognitive. D'une part, ils sont essentiels pour faire face à l'incertitude: ils nous permettent de raisonner sur un objet inconnue X en exploitant les informations stockées dans notre système conceptuel sur des choses qui ressemblent à X. D'autre part, comme l'observe Feeney (2017, p. 167), ils sont un exemple clair de la manière dont les concepts rendent notre cognition efficace.

Les inférences basées sur les catégories sont structurées comme des arguments avec une ou plusieurs prémisses de la forme "Les X sont S" —où X est une catégorie et S une propriété—, et une conclusion du même type avec une catégorie différente. Les CBI peuvent être classées de deux manières: selon la quantité de prémisses; et selon que la conclusion se situe au même niveau conceptuel que les prémisses ou dans une catégorie supérieure. Lorsque les prémisses et les catégories de conclusions sont au même niveau conceptuel, l'argument est dit "spécifique" ; "lorsque l'argument implique une généralisation —un "saut" vers un niveau conceptuel supérieur—, alors il est dit "général". Par exemple, les arguments de la forme $rouge - gorge \rightarrow corbeau$ ou $table \rightarrow chaise$ sont spécifiques, alors que des arguments comme $rouge\text{-}gorge \rightarrow oiseau$, $rouge\text{-}gorge \rightarrow animal$ ou $table \rightarrow Mobilier$, sont généraux. Les arguments

spécifiques et généraux peuvent être composés d'une ou de plusieurs prémisses (voir figure 7.1, Chapitre 7).

Il existe plusieurs propriétés empiriques du CBI, et le modèle présenté dans cette thèse peut les expliquer toutes. Toutefois, pour ce résumé, je n'en considérerai que deux : la similarité et la typicité.

La principale relation catégorielle qui guide les CBI est la similarité. En psychologie, la notion de similarité s'est avérée fructueuse depuis les années 1970. Depuis les travaux pionniers de Shepard (1987) et Tversky (1977), des modèles formels de similarité ont été développés pour expliquer la formation des concepts, la catégorisation et même l'induction. Et depuis les travaux de Rosch sur les prototyes (1973; 1983), la similarité a été prise comme critère central pour expliquer la structure des catégories. Il n'est pas surprenant que la littérature empirique ait montré que le critère le plus solide utilisé dans la CBI est la similarité entre les catégories (Carey, 1985; López et al., 1992; Osherson et al., 1990; Rips, 1975). Cela peut être formulé en disant que nos expectatives concernant la projection de la propriété entre deux catégories X et Y est une fonction positive de leur similarité. Par exemple, des arguments tels que "Les autruches sont S, puis les émeus sont S" sont généralement considérés comme plus forts que des arguments tels que "Les autruches sont S, puis les geais bleus sont S", puisque $sim(autruche, emeu) > sim(autruche, geaibleu)$, où $sim(X, Y)$ indique une mesure de la similarité entre les catégories X et Y.

Une autre relation conceptuelle derrière CBI est la typicité. L'effet le plus robuste constaté dans la littérature empirique est que les expectatives sur la projection de la propriété dans un argument inductif sont une fonction positive de la typicité de la catégorie dans la prémisse. Par exemple, l'inférence "Les rouges-gorges ont l'enzyme E, donc les autruches ont donc l'enzyme E" est souvent jugée plus forte que "Les pingouins ont l'enzyme E; donc les autruches ont l'enzyme E". Cela s'explique par le fait que les rouges-gorges sont des oiseaux prototypiques, et qu'en tant que tels, ils représentent mieux la catégorie que les pingouins – qui sont atypiques. Dans une moindre mesure, la typicité de la conclusion semble également être un facteur dans les inférences basées sur la catégorie. Hampton et Cannon (2004) ont montré que les arguments avec

des catégories de conclusions prototypiques —comme $poule \rightarrow rouge - gorge$— sont jugés plus forts que les arguments avec des catégories de conclusions non typiques —comme $poule \rightarrow vulture$.

Par aillieurs, l'effet de typicité produit une *asymétrie*, c'est-à-dire que le fait de changer les catégories des prémisses et de la conclusion modifie souvent les expectatives de la projection de la propriété, selon le degré de typicité de la catégorie dans les prémisses. Par exemple, des arguments tels que "Les vaches ont l'enzyme E, donc les loutres ont l'enzyme E" sont considérés comme plus forts que des arguments tels que "Les loutres ont l'enzyme E, donc les vaches ont l'enzyme E" puisque les vaches sont des mammifères plus typiques que les loutres.

Un modèle basé sur des espaces conceptuels

Ce modèle utilise également la notion d'expectative. Nous parlons notamment de l'expectative de projection de la propriété parmi les catégories au lieu de *force de l'argument*. Comme nous l'avons vu dans le chapitre précédent, les expectatives jouent un rôle crucial dans le raisonnement quotidien. La phrase "Jean a un nouvel animal de compagnie" est associée à un large ensemble d'expectatives liées aux concepts lexicaux de la phrase. En ce qui concerne le CBI, l'idée est que les dispositions inférentielles de l'agent à projeter un propriété d'une catégorie à une autre sont également déterminées — dans une large mesure– par ses expectatives concernant les régularités dans le monde, qui sont codifiées dans les connaissances de base de l'agent (cf., section 3.4, et section 5.2.1, ce travail).

En ce sens, l'expression $ExpS(X \rightarrow Y)_Z$ sera utilisée pour représenter les expectatives de l'agent selon lesquelles la propriété S est projetée de la catégorie X à la catégorie Y, avec Z comme catégorie de niveau supérieur qui contient à la fois X et Y. En générale, on propose que les expectatives $ExpS(X \rightarrow Y)_Z$ doivent répondre aux critères suivantes: pour satisfaire les critères suivants :

1. Ils sont positivement corrélées avec $sim(X, Y)$.

2. Ils sont positivement corrélées avec $sim(X, p^Z)$, où p^Z est le prototype de Z.

3. Ils sont positivement corrélées avec $sim(Y, p^Z)$.

Le rationale de la première condition est que plus les catégories X et Y sont similaires, plus on s'attend que Y ait les memes propriétés que X. En ce qui concerne la deuxième condition, l'intuition est que plus la catégorie X est prototypique, plus on s'attend à ce qu'une autre catégorie Y ait la propriété S, étant donné que X la possède. La condition (3) est motivée par l'effet de typicité de Hampton et Cannon (2004): plus le Y est prototypique, plus on s'attend à ce que Y ait la propriété S si X l'a.

Le modèle présenté ici est basé sur les distances entre les catégories représentées dans les espaces conceptuels. Comme nous l'avons dit précédemment, les concepts sont des régions d'espaces conceptuels. La stratégie consiste ici à considérer les distances entre les prototypes des catégories (représentés par des points), et à prendre les volumes des régions représentant les concepts dans l'EC comme prédicteurs des expectatives, c'est-à-dire de la force des arguments dans le CBI. Le volume d'un concept dans un espace conceptuel dépend de la métrique attribuée à cet espace, et il est défini de manière standard. Notez que le volume d'un concept dépend de la variabilité des propriétés qui peuvent être attribuées à un objet relevant de ce concept dans chaque domaine. Par exemple, on s'attend à ce que le concept *chien* ait un volume plus important que le concept *tigre*, car les chiens peuvent avoir de nombreuses couleurs, formes et tailles différentes; alors que les tigres ont une variabilité moindre dans ces domaines. La conséquence immédiate est que plus le concept est hétérogène, plus son volume sera important dans un espace conceptuel.

Nous supposons que $ExpS(X \rightarrow Y)_Z$ est positivement corrélé avec le volume $V(X)$ de X et négativement corrélé avec le volume $V(Y)$ de Y. La corrélation positive est due au fait que plus le volume $V(X)$ est grand, plus il "couvre" —ou est plus représentatif de la catégorie supérieure Z. Par exemple, $ExpS(ours \rightarrow$

$loup)_{Mammifre}$ devrait être plus grand que $ExpS(ourspolaire \rightarrow loup)_{Mammifre}$ (voir figure 7.3). La formule proposée pour modéliser ces idées est la suivant:

$$logExpS(X \rightarrow Y)_Z = \left(d(p^X, p^Y)^{\frac{V(X-Y)}{(Y-X)}}.d(p^X, p^Z)^a.d(p^Y, p^Z)^b\right)^{-1} \quad (A.1)$$

Cette équation peut être transformée en la suivante en prenant le logarithme et en considérant la relation entre la distance et la similarité, établie par la loi de Shepard:

$$logExpS(X \rightarrow Y)_Z = \frac{V(X-Y)}{(Y-X)}.sim(p^X, p^Y) + a.sim(p^X, p^Z) + b.sim(p^Y, p^Z) \quad (A.2)$$

Pour voir comment cette formule prédit l'effet de similarité, supposons que nous avons les arguments (i) "Les chiens ont la propriété S, donc les ours ont la propriété S" et (ii) "Les chiens ont la propriété S, donc les loups ont la propriété S". Supposons que $V(ours)$ est identique ou très similaire à $V(loup)$, alors, $\frac{V(chien-ours)}{(ours-chien)} \approx \frac{V(chien-loup)}{(loup-chien)}$; et que $d(p^{ours}, p^{mammifere}) \approx d(p^{loup}, p^{mammifere})$. Alors, comme $d(p^{ours}, p^{chien}) > d(p^{loup}, p^{chien})$, on a que $sim(p^{loup}, p^{chien}) > sim(p^{ours}, p^{chien})$, et donc $ExpS(chien \rightarrow loup)_{mammifere} > ExpS(chien \rightarrow ours)_{mammifere}$.

De la même manière, voyons maintenant comment le modèle prédit l'effet de la typicité. Considérez les arguments suivants: "Les pigeons ont la propriété S, donc les hérons ont la propriété S" et "Les pingouins ont la propriété S, donc les hérons ont la propriété S". Selon les études empiriques, le premier argument est plus fort que le second car les pigeons sont des oiseaux plus typiques que les pingouins. Ce dernier fait implique que $d(p^{pigeons}, p^{oiseaux}) < d(p^{pingouins}, p^{oiseaux})$, et donc $sim(p^{pigeons}, p^{oiseaux}) > sim(p^{pingouins}, p^{oiseaux})$. Si l'on suppose que toutes ces catégories aient un volume similaire, alors l'équation ci-dessus va donc donner que $ExpS(pigeons \rightarrow herons)_{oiseaux} > ExpS(pingouins \rightarrow herons)_{oiseaux}$.

Ce modèle peut être adapté à des arguments inductifs avec multiples

prémisses. Dans ces cas, nous utilisons un *enveloppe convexe* ("convex Hull") pour traiter l'ensemble des catégories dans les prémisses comme une seule catégorie avec un volume spécifique et un prototype artificiel qui sera le centroïde de ce nouvel ensemble. Les enveloppes convexes sont également des régions convexes d'espaces à $n$dimensions avec les mêmes propriétés géométriques que les régions dans les espaces conceptuels. La taille de leurs volumes est positivement corrélée au nombre de régions convexes qu'elles comprennent, ainsi qu'aux distances entre ces régions. Par exemple, dans un espace conceptuel dans lequel toutes les catégories ont des volumes similaires, le volume de le enveloppe convexe de deux régions contiguës va être plus petit que le volume de deux régions non contiguës de l'espace

Dans ces cas, l'équation prendrait la forme suivante:

$$logExpS(X_1, X_2, \ldots, X_n \rightarrow Y)_Z = \frac{V(C(X_1 \cup X_2 \cup \cdots \cup X_n) - Y)}{V(Y - C(X_1 \cup X_2 \cup, \ldots, \cup X_n))}.sim(p^C, p^Y) + a.sim(p^C, p^Z) + b.sim(p^Y, p^Z) \tag{A.3}$$

Au final du chapitre 7, on montre que cette dernière équation prédit un phénomène très important dans le CBI avec de multiples prémisses, appelé "diversité", parmi beaucoup d'autres. Par ailleurs, le modèle présenté ici englobe la plupart des modèles formels disponibles dans la littérature, et fait quelques nouvelles prédictions.

Au-delà du langage : Modèles, inférence, et concepts scientifique

Au chapitre 8, j'analyse le rôle des concepts et de la représentation dans le raisonnement scientifique. Je reprends la théorie procédurale des concepts et des explications de Stephen Toulmin pour développer deux idées négligées de sa

philosophie de la science : *méthodes de représentation* et *techniques interférentielles*. Je soutiens que ces notions, lorsqu'elles sont correctement articulées, pourraient être utiles pour éclairer la façon dont le raisonnement scientifique est lié aux structures de représentation, aux concepts et aux explications au sein des pratiques scientifiques. J'explore et illustre ces idées en étudiant le développement de la notion de vitesse instantanée lors du passage de la physique géométrique de Galilée à la mécanique analytique. En conclusion, je soutiens que les méthodes de représentation pourraient être considérées comme constitutives de l'inférence scientifique ; et je montre comment ces notions sont liées à d'autres idées similaires de la philosophie scientifique contemporaine, comme celles des modèles et du raisonnement basé sur des modèles.

La littérature classique sur le raisonnement scientifique —historiquement monopolisée par la philosophie– a principalement compris ce processus selon la perspective formaliste, avec la logique et la probabilité comme moteurs centraux, et l'inférence déductive comme cas paradigmatique. En termes généraux, le raisonnement était représenté comme une capacité individuelle, dépendant exclusivement de certains mécanismes cognitifs "câblés", et spécifié par un ensemble de règles générales de domaine opérant sur un système de représentation de type phrase.

Stephen Toulmin a été l'un des premiers philosophes de la science à critiquer ces idées (qui étaient très fortes grâce à l'influence du positivisme logique). Selon lui, une bonne théorie du raisonnement doit expliquer non seulement les processus internes qui se déroulent dans l'esprit/le cerveau des agents cognitifs, mais aussi comment ces processus sont influencés par le contexte socioculturel dans lequel les agents sont intégrés et le rôle des outils et des dispositifs externes que les agents utilisent lorsqu'ils effectuent différentes tâches cognitives (de haut niveau).

Toulmin a développé une approche "écologique" qui aborde la pensée scientifique à travers deux notions centrales : "méthode de représentation" (MR) et "technique inférentielle" (IT). Pour Toulmin, le raisonnement scientifique ne peut être compris que comme une activité socialement ancrée, qui dépend de

la maîtrise de différentes méthodes de représentation de l'information permettant des formes spécifiques d'inférences. Toulmin utilise ces deux notions dans ses deux ouvrages les plus importants sur la philosophie des sciences (Toulmin, 1953, 1972a), et elles jouent un rôle central dans son approche de la science basée sur l'explication. Toutefois, il ne leur fournit pas de définitions analytiques et ne les utilise pas systématiquement dans ses travaux.

Mais, quelles sont exactement ces méthodes de représentation? Si Toulmin a largement utilisé cette notion dans ses deux ouvrages les plus ambitieux sur la philosophie des sciences, il n'en a pas donné de définition analytique. En gros, les MR sont des "techniques intellectuelles" qui permettent aux scientifiques de construire et d'utiliser des modèles de phénomènes. En général, tout système symbolique standardisé que les scientifiques utilisent pour représenter des phénomènes –des diagrammes, images, formules mathématiques, programmes informatiques, etc. Toulmin suggère que le rôle crucial que les MR jouent dans la théorisation scientifique est lié à la manière dont ils compensent les limites de représentation du langage naturel. Il écrit:

> ..."representation techniques" include all those varied procedures by which scien- tists demonstrate—i.e. exhibit, rather than prove deductively—the general relations discoverable among natural objects, events and phenomena: so, comprising not only the use of mathematical formalisms, but also the drawing of graphs and diagrams, the establishment of taxonomic "trees" and classifications, the devising of computer programmes, etc. (Toulmin, 1972a, pp. 162–163)

Les caractéristiques les plus importantes des MR sont : (1) ils sont associés aux idéaux explicatifs des disciplines scientifiques ; (2) ils établissent les règles et les ressources symboliques qui constitueront les modèles spécifiques que les scientifiques utiliseront pour représenter, comprendre et raisonner les phénomènes ; (3) elles font partie des "méthodes collectives de pensée" (**?**, viii), et en tant que telles elles sont de nature "communautaire" - leur utilisation et leur "validité" dépendent de l'accord de la communauté scientifique ; (4) elles

jouent un rôle central dans la découverte ; et (5) elles sont essentielles au raisonnement scientifique car elles apportent avec elles de nouvelles technologies de l'information. Dans la suite du chapitre, je me concentrerai sur les points (4) et (5) et sur leur relation mutuelle.

En ce qui concerne la notion de "technique inférentielle", elle peut être définie approximativement comme l'ensemble des procédures qui permettent aux scientifiques de tirer des conclusions basées sur un modèle dans le contexte d'un MR particulier. Plus précisément, lorsque les scientifiques tentent d'expliquer un phénomène, ils le représentent par moyen de la construction d'un modèle à l'aide de ressources symboliques spécifiques. De nombreuses déductions que les scientifiques vont faire en utilisant ce modèle, dépendent des procédures de manipulation de ces structures symboliques. Comme j'explique dans le Chapitre 8, les IT sont importantes parce qu'elles remettent en question la vision classique du raisonnement qui s'appuie sur la thèse de la formalité (voir Toulmin,1972a, pp. 487-488 ; ou Toulmin, 1953, p. 25).

De mon analyse des idées de Toulmin, il découle une conclusion forte en accord avec mon approche pluraliste de l'inférence : le format de représentation détermine (dans un sens fort) le type de mécanismes inférentiels possibles. En d'autres termes, il n'y a pas d'inférence sans représentation.

Je soutiens cette denrière affirmation en analysant une étude de cas de l'histoire des sciences. Je montre, à partir des idées de M.Panza (Panza, 2002) et M. Blay (Blay, 1998), que l'émergence de la notion de vitesse instantanée en physique a été possible grâce au développement d'une nouvelle méthode de représentation.En gros, la physique du mouvement avant Newton était principalement basée sur une méthode de représentation géométrique héritée de Galilée. Cette méthode avait, comme contrainte implicite, l'impossibilité de comparer directement deux variables liées au mouvement. Au lieu de cela, seules les proportions (relations entre deux variables) étaient candidates à la comparaison. Cela empêchait les scientifiques de représenter directement une notion explicite de vitesse instantanée, et par conséquent, les empêchait de raisonner sur ce phénomène.

Lorsque Leibniz et Newton développèrent le calcul infinitésimal, il devint

possible de construire une méthode analytique de représentation qui était libre de cette contrainte. Ce changement dans la façon de représenter le mouvement a permis à Varignon de construire une définition fonctionnelle de la vitesse instantanée, basée sur la notion d'infinitésimaux. Il en résulte une nouvelle façon de faire des inférences sur le mouvement, qui n'était pas disponible auparavant.

Une conclusion importante de cette analyse est qu'en science, le contenu conceptuel est distribué, dans une large mesure, sur des modèles externes qui utilisent différentes ressources symboliques. En conséquence, et en raison de la relation entre les concepts, le raisonnement et la compréhension qui a été soulignée au début de cette thèse, la manipulation des modèles est une condition préalable à la possession de concepts pour de nombreuses notions scientifiques. En même temps, le pluralisme inférentiel peut être vu comme une conséquence naturelle de la diversité des formes de représentations dans la science.

References

Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Aggelopoulos, N. C. (2015). Perceptual inference. *Neuroscience & Biobehavioral Reviews*, *55*, 375–392.

Anderson, A., Belnap Jr, N. D., & Dunn, J. M. (2017). *Entailment, vol. ii: The logic of relevance and necessity*. Princeton University Press.

Anderson, J., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press.

Andrews, A. D. (1993). Mental models and tableau logic. *Behavioral and Brain Sciences*, *16*(2), 334–334.

Aristoteles, & Ross, W. D. (1965). *Aristotle's prior and posterior analytics: A revised text with introduction and commentary*. Clarendon Press.

Aydede, M. (2005). Computation and functionalism. In G. Irzik & T. Grünberg (Eds.), *Turkish studies in the history and philosophy of science* (pp. 177–204). Springer.

Baker, M. C. (2003). *Lexical categories: verbs, nouns, and adjectives*. Cambridge University Press.

Bar-Am, N. (2008). *Extensionalism: The revolution in logic*. Springer Science & Business Media.

Barsalou, L. (1987). The instability of graded structure. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 101–140). Cambridge University Press.

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and cognitive processes*, *18*(5-6), 513–562.

Barsalou, L., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg,

& A. Graesser (Eds.), *Symbols, embodiment, and meaning* (pp. 245–283). Oxford University Press.

Barsalou, L., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought* (pp. 129–163).

Barwise, J. (1986). Information and circumstance. *Notre Dame Journal of Formal Logic*, *27*(3), 324–338.

Barwise, J. (1989). *The situation in logic*. Center for the Study of Language (CSLI).

Barwise, J., & Etchemendy, J. (1996). Visual information and valid. *Logical Reasoning with Diagrams*.

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought. *Thinking & Reasoning*, *23*(2), 158–183.

Bierwisch, M., & Kiefer, F. (1969). Remarks on definitions in natural language. In F. Kiefer (Ed.), *Studies in syntax and semantics* (pp. 55–79). Springer.

Black, J., & Overton, W. F. (1990). Reasoning, logic, and thought disorders. In W. F. Overton (Ed.), *Reasoning, necessity, and logic* (pp. 255–297). Lawrence Elrbaum.

Blay, M. (1992). *La naissance de la mécanique analytique*. PUF.

Blay, M. (1998). *Reasoning with the infinite: from the closed world to the mathematical universe*. University of Chicago Press.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest studies in philosophy*, *10*, 615–678.

Block, N. (1998a). Conceptual role semantics. In E. Craig (Ed.), *Routledge encyclopedia of philosophy* (pp. 242–256). Routledge.

Block, N. (1998b). Holism, mental and semantic. In E. Craig (Ed.), *Routledge encyclopedia of philosophy*. Routledge.

Boden, M. (1988). *Computer models of mind: Computational approaches in theoretical psychology*. Cambridge University Press.

Boghossian, P. (1996). Analyticity reconsidered. *Noûs*, *30*(3), 360–391.

Boghossian, P. (2014). What is inference? *Philosophical studies*, *169*(1), 1–18.

Boghossian, P. (2018). Delimiting the boudaries of inference. *Philosophical Issues*, *28*, 55–69.

Bonatti, L. (1994a). Propositional reasoning by model? *Psychological Review*, *101*, 725–733.

Bonatti, L. (1994b). Why should we abandon the mental logic hypothesis? *Cognition*, *50*(1-3), 17–39.

Bonnay, D. (2014). Logical constants, or how to use invariance in order to complete the explication of logical consequence. *Philosophy Compass*, *9*(1), 54–65.

Booth, R., Meyer, T., & Varzinczak, I. (2013). A propositional typicality logic for extending rational consequence. *Trends in belief revision and argumentation dynamics*, *48*, 123–154.

Brachman, R. J. (1977). What's in a concept. *International journal of man-machine studies*, *9*(2), 127–152.

Braddon-Mitchell, D., & Jackson, F. (2006). *The philosophy of mind and cognition*. Wiley-Blackwell.

Braine, M. (1962). Piaget on reasoning: A methodological critique and alternative proposals. *Monographs of the Society for Research in Child Development*, 41–63.

Braine, M. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological review*, *85*(1), 1—21.

Braine, M. (1990). The "natural logic" approach to reasoning. In W. F. Overton (Ed.), *Reasoning, necessity, and logic* (p. 133—157). Lawrence Elrbaum.

Braine, M., Reiser, B., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In *Psychology of learning and motivation* (pp. 313–371).

Brandom, R. (1981). Semantic paradox of material implication. *Notre dame journal of formal logic*, *22*(2), 129–132.

Brandom, R. (1998a). Action, norms, and practical reasoning. *Philosophical Perspectives*, *12*, 127–139.

Brandom, R. (1998b). *Making it explicit*. Harvard University Press.

Brandom, R. (2000). *Articulating reasons: An introduction to inferentialism*.

Harvard University Press.

Brandom, R. (2015). *From empiricism to expressivism.* Harvard University Press.

Brewka, G., Roelofsen, F., & Serafini, L. (2007). Contextual default reasoning. In *Proceedings of the 20th international joint conference on artifical intelligence* (p. 268–273). Morgan Kaufmann Publishers.

Brigandt, I. (2010). Scientific reasoning is material inference: Combining confirmation, discovery, and explanation. *International Studies in the Philosophy of Science*, *24*(1), 31–43.

Bright, A., & Feeney, A. (2014). Causal knowledge and the development of inductive reasoning. *Journal of Experimental Child Psychology*, *122*, 48–61.

Broome, J. (2013). *Rationality through reasoning.* John Wiley & Sons.

Brown, H. I. (1986). Sellars, concepts and conceptual change. *Synthese*, *68*(2), 275–307.

Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. Small, G. Cottrell, & M. Tanenhaus (Eds.), *Lexical ambiguity resolution* (pp. 477–508).

Buijsman, S. (2018). How numerals support new cognitive capacities. *Synthese*, *197*, 1–18.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2/3), 96–107.

Burnston, D. (2020). Fodor on imagistic mental representations. *Rivista internazionale di Filosofia e Psicologia*, *11*(1), 71–94.

Byrne, R. (1991). Can valid inferences be suppressed? *Cognition*, *39*(1), 71–78.

Byrne, R., & Johnson-Laird, P. N. (2009). 'if' and the problems of conditional reasoning. *Trends in Cognitive Sciences*, *13*, 0–287.

Byrnes, J. (1992). Meaningful logic: Developmental perspectives. In H. Beilin & P. B. Pufall (Eds.), *Piaget's theory: Prospects and possibilities* (pp. 163–183). Erlbaum Hillsdale.

Calzavarini, F. (2020). *Brain and the lexicon.* Springer.

Camp, E. (2007). Thinking with maps. *Philosophical perspectives*, *21*, 145–182.

Cann, R. (2011). Sense relations. In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Semantics* (p. 456-478). De Gruyter Mouton.

Carey, S. (1985). *Conceptual change in childhood.* MIT Press.

Carey, S. (2000). The origin of concepts. *Journal of Cognition and Development*, *1*(1), 37–41.

Carey, S. (2015). Why theories of concepts should not ignore the problem of acquisition. *Disputatio*, *7*(41), 113–163.

Carnap, R. (1938-55). Logical foundations of the unity of science. In O. Neurath, R. Camap, & C. Morris (Eds.), *International encyclopedia of unified science* (p. 42-62). University of Chicago Press.

Carnap, R. (1952). Meaning postulates. *Philosophical studies*, *3*(5), 65–73.

Carnap, R. (1955). Meaning and synonymy in natural languages. *Philosophical studies*, *6*(3), 33–47.

Carnap, R. (1959). The elimination of metaphysics through logical analysis of language. In A. Ayer (Ed.), *Logical positivism* (pp. 60–81). Free Press.

Carnap, R. (1971). A basic system of inductive logic, part i. In R. Carnap & R. C. Jeffrey (Eds.), *Studies in inductive logic and probability* (Vol. 1, pp. 35–165). University of California Press.

Carnap, R. (1988). *Meaning and necessity.* University of Chicago Press.

Carnap, R. (2000). *Logical syntax of language.* Routledge.

Carruthers, P., Stich, S., & Siegal, M. (2002). *The cognitive basis of science.* Cambridge University Press.

Carus, A. (2004). Sellars, carnap, and the logical space of reasons. In S. Awodey & C. Klein (Eds.), *Carnap brought home: The view from jena* (pp. 317–355).

Chapman, M. (1979). *Constructive evolution: Origins and development of piaget's thought.* Cambridge University Press.

Chater, N., & Oaksford, M. (1993). Logicism, mental models and everyday reasoning. *Mind & Language*, *8*, 72–89.

Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science*, *67*(4), 625–647.

Cheng, P., & Holyoak, K. (1985). Pragmatic reasoning schemas. *Cognitive*

Psychology, *17*, 391–416.

Cheng, P., Holyoak, K., Nisbett, R., & Oliver, L. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, *18*(3), 293–328.

Cherniak, C. (1984). Prototypicality and deductive reasoning. *Journal of Verbal Learning and Verbal Behavior*, *23*(5), 625–642.

Cherniak, C. (1986). *Minimal rationality*. MIT Press.

Chomsky, N. (1959). A review of bf skinner's verbal behavior. *Language*, *35*(1), 26–58.

Chomsky, N. (1986). *Knowledge of language*. Greenwood Publishing Group.

Chomsky, N. (2002). *On nature and language*. Cambridge University Press.

Chomsky, N. (2005). *Rules and representations*. Columbia University Press.

Chomsky, N. (2014 [1965]). *Aspects of the theory of syntax*. MIT Press.

Cimpian, A., Brandone, A., & Gelman, S. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, *34*(8), 1452–1482.

Clagett, M., & Oresme, N. (1968). *Nicole oresme and the medieval geometry of qualities and motions*. University of Wisconsin Press.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in cognitive sciences*, *10*(8), 370–374.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Clavelin, M. (1968). *La philosophie naturelle de galilée*. Librairie Armand Colin.

Coley, J., Shafto, P., Stepanova, O., & Baraff, E. (2005). Knowledge and category-based induction. In W.-K. Ahn, R. Goldstone, B. Love, A. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of douglas medin* (pp. 69–85). American Psychological Association.

Conway, P. (1995). *Aristotelian formal and material logic*. University Press of America.

Corcoran, J. (2006). Schemata: the concept of schema in the history of logic. *Bulletin of Symbolic Logic*, *12*(2), 219–240.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, *31*(3), 187 - 276.

Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences*, *107*(Supplement 2), 9007–9014.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? *Cognition*, *58*(1), 1–73.

Cruse, A. (1986). *Lexical semantics*. Cambridge University Press.

Cruse, A. (2000). *Meaning in language*. Oxford University Press.

Cruse, A. (2002). Hyponymy and its varieties. In R. Green, C. Bean, & S. H. Myaeng (Eds.), *The semantics of relationships* (pp. 3–21). Springer.

Cummings, L. (2013). *Pragmatics: A multidisciplinary perspective*. Routledge.

Cummins, R. (1992). Conceptual role semantics and the explanatory role of content. *Philosophical Studies*, *65*, 103–127.

Cuyckens, H. (1997). Prepositions in cognitive lexical semantics. *Lexikalische und grammatische Eigenschaften präpositionaler Elemente*, 63–82.

Decock, L. (2010). Quine's antimentalism in linguistics. *Logique et Analyse*, *53*, 371–385.

Decock, L., & Douven, I. (2011). Similarity after goodman. *Review of philosophy and psychology*, *2*(1), 61–75.

Decock, L., & Douven, I. (2014). What is graded membership? *Noûs*, *48*(4), 653–682.

Delgrande, J. (2011). What's in a default? thoughts on the nature and role of defaults in nonmonotonic reasoning. In G. Brewka, V. W. Marek, & M. Truszczynski (Eds.), *Nonmonotonic reasoning. essays celebrating its 30th anniversary* (pp. 89–109). College Publications.

Delgrande, J., & Schaub, T. (2000). Expressing preferences in default logic. *Artificial Intelligence*, *123*(1-2), 41–87.

De Neys, W., Schaeken, W., & D'ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory & cognition*, *30*(6), 908–920.

Dennett, D. (1969). *Content and consciousness*. Routledge.

Dennett, D. (1979). *Brainstorms series*. Harvester Press.

Dennett, D. (1984). The role of the computer metaphor in understanding the mind. In *Proc. of a symposium on computer culture* (Vol. 426, pp. 266–275).

Devadoss S., O. J. (2011). *Discrete and computational geometry*. Princeton University Press.

deVries, W. A. (2005). *Wilfrid sellars*. Acumen.

Dominowski, R. (1995). Content effects in wason's selection task. In S. Newstead & J. S. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of peter wason* (pp. 41–65). Psychology Press.

Douven, I. (2019). Putting prototypes in place. *Cognition*, *193*, 104007.

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic*, *42*(1), 137–160.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110*(3), 412–431.

Ducheyne, S. (2008). Galileo and huygens on free fall: Mathematical and methodological differences. *Dynamis*, *28*, 243–274.

Dummett, M. (1993). *The seas of language*. Clarendon Press Oxford.

Dutilh Novaes, C. (2012). *Formal languages in logic*. Cambridge University Press.

Edgington, D. (2020). Indicative conditionals. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/conditionals/.

Elffers, E. (2014). Earlier and later anti-psychologism in linguistics. In V. Kasevich, Y. A. Kleiner, & P. Sériot (Eds.), *History of linguistics 2011* (Vol. 123, p. 127-136). John Benjamins Publishing Company.

Eliasmith, C., & Bechtel, W. (2006). Symbolic versus subsymbolic. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (pp. 123–142). MacMillian.

Eliot, J. (1987). *Models of psychological space*. Springer-Verlag.

Elqayam, S., & Over, D. (2012). Probabilities, beliefs, and dual processing: the

paradigm shift in the psychology of reasoning. *Mind & Society*, *11*(1), 27–40.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning. *Thinking & Reasoning*, *19*(3-4), 249–265.

Etchemendy, J. (1983). The doctrine of logic as form. *Linguistics and Philosophy*, 319–334.

Etherington, D. W., & Reiter, R. (1983). On inheritance hierarchies with exceptions. In *Aaai* (Vol. 83, pp. 104–108).

Eva, B., & Hartmann, S. (2018). Bayesian argumentation and the value of logical validity. *Psychological review*, *125*(5), 806.

Evans, J. S. (1989). Concepts and inference. *Mind & Language*, *4*(1-2), 29–34.

Evans, J. S. (1993a). The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition*, *48*(1), 1–20.

Evans, J. S. (1993b). On rules, models and understanding. *Behavioral and Brain Sciences*, *16*(2), 345–346.

Evans, J. S. (2019). *Hypothetical thinking: Dual processes in reasoning and judgement*. Psychology Press.

Evans, J. S., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, *11*(3), 295–306.

Evans, J. S., & Feeney, A. (2004). The role of prior belief in reasoning. In J. Leighton & R. Sternberg (Eds.), *The nature of reasoning* (pp. 78–102). Cambridge University Press.

Evans, J. S., Newstead, S., & Byrne, R. (1993). *Human reasoning: The psychology of deduction*. Psychology Press.

Evans, J. S., & Over, D. (2004). *If: Supposition, pragmatics, and dual processes*. Oxford University Press.

Evans, J. S., Over, D., & Handley, S. (2005). Suppositions, extensionality, and conditionals: A critique of the mental model theory of johnson-laird and byrne (2002). *Psychological Review*, *112*, 1040–1052.

Falmagne, R. J. (1990). Language and the acquisition of logical knowledge. In W. F. Overton (Ed.), *Reasoning, necessity, and logic* (pp. 111–131). Lawrence Elrbaum.

Falmagne, R. J., & Gonsalves, J. (1995). Deductive inference. *Annual Review of Psychology*, *46*(1), 525–559.

Feeney, A. (2017). Forty years of progress on category-based inductive reasoning. In L. J. Ball & V. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 167–185). Routledge.

Feeney, A., Hayes, B., & Heit, E. (2015). From tool to theory: What recognition memory reveals about inductive reasoning. In A. Feeney & V. Thompson (Eds.), *Reasoning as memory* (p. 110-127). Psychology Press.

Feeney, A., & Heit, E. (2011). Properties of the diversity effect in category-based inductive reasoning. *Thinking & Reasoning*, *17*(2), 156–181.

Feest, U. (2005). Operationism in psychology: What the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences*, *41*(2), 131–149.

Fillmore, C. (2006). Frame semantics. In D. Geeraerts (Ed.), *Cognitive linguistics: Basic readings* (pp. 373–400). Mouton de Gruyter.

Fodor, J. (1975). *The language of thought.* Harvard University Press.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and brain sciences*, *3*(1), 63–73.

Fodor, J. (1985). Fodor's guide to mental representation. *Mind*, *94*(373), 76–100.

Fodor, J. (1987). *Psychosemantics.* MIT Press.

Fodor, J. (1988). Modules, frames, fridgeons, sleeping dogs and the music of the spheres. In Z. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial.* Ablex Publishing Corporation.

Fodor, J. (1994). Concepts: A potboiler. *Cognition*, *50*(1-3), 95–113.

Fodor, J. (2001). *The mind doesn't work that way.* MIT Press.

Fodor, J. (2008). *Lot 2.* Oxford University Press.

Fodor, J., & Lepore, E. (1991). Why meaning (probably) isn't conceptual role. *Mind and language*, *6*(4), 328–43.

Fodor, J., & Lepore, E. (1992). *Holism: A shopper's guide.* Blackwell Publishing.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture. In

S. Pinker & J. Mehler (Eds.), *Connections and symbols* (pp. 3–71). MIT Press.

Fodor, J., & Pylyshyn, Z. (2015). *Minds without meanings*. MIT Press.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition*, *54*(1), 1–71.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, *5*(10), 914–926.

Frege, G. (1948). Sense and reference. *The philosophical review*, *57*(3), 209–230.

Frege, G. (1979). *Posthumous writings*. University of Chicago Press.

Frege, G. (1980). *The foundations of arithmetic*. Northwestern University Press.

French, R., Mareschal, D., Mermillod, M., & Quinn, P. (2004). The role of bottom-up processing in perceptual categorization by 3-to 4-month-old infants. *Journal of experimental psychology: General*, *133*(3), 382.

Frigg, R., & Hartmann, S. (2020). Models in Science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/models-science/.

Galilei, G. (1954 [1632]). *Dialogues concerning two new sciences*. Dover Publications.

Galotti, K., Baron, J., & Sabini, J. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General*, *115*(1), 16–25.

Gärdenfors, P. (1992). The role of expectations in reasoning. In M. Masuch & L. Pólos (Eds.), *Knowledge representation and reasoning under uncertainty* (pp. 1–16). Springer.

Gärdenfors, P. (1993). The emergence of meaning. *Linguistics and Philosophy*, *16*(3), 285–309.

Gardenfors, P. (1994). How logic emerges from the dynamics of information peter gardenfors. In A. Visser & J. Eijck (Eds.), *Logic and information flow* (pp. 49–77). MIT Press.

Gärdenfors, P. (1997). Symbolic, conceptual and subconceptual representations. In V. Cantoni, V. Di Gesù, A. Setti, & D. Tegolo (Eds.), *Human and machine perception* (pp. 255–270). Springer.

Gärdenfors, P. (2000). *Conceptual spaces.* MIT Press.

Gärdenfors, P. (2008). Reasoning in conceptual spaces. In J. Adler & L. Rips (Eds.), *Reasoning: studies of human inference and its foundations* (pp. 302–320). Cambridge University Press.

Gärdenfors, P. (2014). *The geometry of meaning.* MIT Press.

Gärdenfors, P. (2015). The geometry of preposition meanings. *Baltic International Yearbook of Cognition, Logic and Communication*, *10*(1), 2–33.

Gärdenfors, P. (2020). Events and causal mappings modeled in conceptual spaces. *Frontiers in Psychology*, *11*, 1-10.

Gärdenfors, P., Jost, J., & Warglien, M. (2018). From actions to effects: Three constraints on event mappings. *Frontiers in psychology*, *9*, 1391.

Gärdenfors, P., & Makinson, D. (1994). Nonmonotonic inference based on expectations. *Artificial Intelligence*, *65*(2), 197–245.

Gärdenfors, P., & Warglien, M. (2012). Using conceptual spaces to model actions and events. *Journal of semantics*, *29*(4), 487–519.

Gaudet, E. (2006). *Quine on meaning: The indeterminacy of translation.* A&C Black.

Geeraerts, D. (2010). *Theories of lexical semantics.* Oxford University Press.

Geeraerts, D., & Cuyckens, H. (2007). *The oxford handbook of cognitive linguistics.* Oxford University Press.

Gelman, R. (1990). First principles organize attention to and learning about relevant data. *Cognitive science*, *14*(1), 79–106.

Gelman, S. (2009). Learning from others: Children's construction of concepts. *Annual review of psychology*, *60*, 115–140.

Gelman, S., Leslie, S.-J., Was, A. M., & Koch, C. M. (2015). Children's interpretations of general quantifiers, specific quantifiers and generics. *Language, cognition and neuroscience*, *30*(4), 448–461.

Giere, R. N. (1999). Using models to represent reality. In *Model-based reasoning in scientific discovery* (pp. 41–57). Springer.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of science*, *71*(5), 742–752.

Giere, R. N. (2010). *Scientific perspectivism.* University of Chicago Press.

Gigerenzer, G., & Goldstein, D. (1996). Mind as computer: Birth of a metaphor. *Creativity Research Journal*, *9*, 131–144.

Giordano, L., Gliozzi, V., Olivetti, N., & Pozzato, G. L. (2008). Reasoning about typicality in preferential description logics. In S. Hölldobler, C. Lutz, & H. Wansing (Eds.), *European workshop on logics in artificial intelligence* (pp. 192–205). Springer.

Giusti, E. (1994). Il filosofo geometra. matematica e filosofia naturale in galileo. *Nuncius*, *9*(2), 485–498.

Goldfarb, W. (2001). Frege's conception of logic. In J. Floyd & S. Shieh (Eds.), *Future pasts: The analytic tradition in twentieth-century philosophy* (pp. 25–41). Oxford University Press.

Golding, E. (1981). The effect of past experience on problem-solving. In *Bulletin of the british psychological society* (Vol. 34, pp. 186–186).

Goldstone, R., Medin, D., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition*, *25*(2), 237–255.

Gómez-Torrente, M. (2002). The problem of logical constants. *Bulletin of Symbolic Logic*, *8*(1), 1–37.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (p. 437–446). Bobbs-Merrill.

Greenberg, M., & Harman, G. (2005). Conceptual role semantics. In E. Lepore & B. Smith (Eds.), *Oxford handbook of philosophy of language.* Oxford University Press.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 41–58). Academic Press.

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

Guicciardini, N. (2013). Mathematics and the new sciences. In J. Z. Buchwald & R. Fox (Eds.), *The oxford handbook of the history of physics* (pp. 226–264).

Hacking, I. (1994). Styles of scientific thinking or reasoning. In K. Gavroglu, J. Christianidis, & E. Nicolaidis (Eds.), *Trends in the historiography of science* (pp. 31–48). Springer.

Hacking, I. (2009). *Scientific reason.* National Taiwan University Press Taipei.

Hahn, U. (2020). Argument quality in real world argumentation. *Trends in Cognitive Sciences*, *34*, 363–374.

Hampton, J. (2001). The role of similarity in natural categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 13–28). Oxford University Press.

Hampton, J. (2012). Thinking intuitively: The rich (and at times illogical) world of concepts. *Current directions in psychological science*, *21*(6), 398–402.

Hampton, J., & Cannon, I. (2004). Category-based induction: An effect of conclusion typicality. *Memory & Cognition*, *32*(2), 235–243.

Hanna, R. (1991). How ideas became meanings: Locke and the foundations of semantic theory. *The Review of Metaphysics*, *44*, 775–805.

Hanna, R. (2004). *Kant and the foundations of analytic philosophy.* Oxford University Press.

Harman, G. (1967). Quine on meaning and existence, i. the death of meaning. *The Review of Metaphysics*, 124–151.

Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*(2), 242–256.

Harman, G. (1984). Logic and reasoning. In H. Leblanc, E. Mendelson, & A. Orenstein (Eds.), *Foundations: Logic, language, and mathematics* (pp. 107–127). Springer.

Harman, G. (1986). *Change in view: Principles of reasoning.* MIT Press.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 1–52). Cambridge University Press.

Hatfield, G. (2002). Perception as unconscious inference. In D. Heyer & R. Mausfeld (Eds.), *Perception and the physical world* (pp. 113–143). John Wiley & Sons.

Haugeland, J. (1987). An overview of the frame problem. In K. Ford &

Z. Pylyshyn (Eds.), *The robot's dilemma* (pp. 77–94). Ablex.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT Press.

Hayes, B., Fritz, K., & Heit, E. (2013). The relationship between memory and inductive reasoning: Does it develop? *Developmental Psychology*, *49*(5), 848.

Hayes, B., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *WIREs Cognitive science*, *1*(2), 278–292.

Hayes, P. (1977). In defense of logic. In *Proceedings of the 5th ijcai'77* (p. 559–565). Morgan Kaufmann.

Hayes, P. (1988). What the frame problem is and isn't. In Z. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial* (pp. 123–137). Ablex Publishing Corporation.

Heit, E. (1997). Features of similarity and category-based induction. In *An interdisciplinary workshop on similarity and categorisation (simcat)*.

Heit, E. (1998). A bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*(4), 569–592.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology*, *20*(2), 411.

Hendricks, S. (2006). The frame problem and theories of belief. *Philosophical studies*, *129*(2), 317–333.

Henle, M. (1962). On the relation between logic and thinking. *Psychological review*, *69*(4), 366.

Henschen, L. (1987). Inference. In S. Shapiro (Ed.), *Encyclopedia of artificial intelligence* (pp. 418–419). John Wiley and Sons.

Hlobil, U. (2014). Against boghossian, wright and broome on inference. *Philosophical Studies*, *167*(2), 419–429.

Hlobil, U. (2019). Inferring by attaching force. *Australasian Journal of Philosophy*, *97*(4), 701–714.

Hodges, W. (1993). The logical content of theories of deduction. *Behavioral*

and Brain Sciences, *16*(2), 353–354.

Hoffman, D. (1983). The interpretation of visual illusions. *Scientific American*, *249*(6), 154–163.

Hoffman, D. (2005). Visual illusions and perception. *Yearbook of Science and Technology. McGraw-Hill*.

Hoffman, D., & Richards, W. (1984). Parts of recognition. *Cognition*, *18*(1-3), 65–96.

Hoffman, D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic bulletin & review*, *22*(6), 1480–1506.

Horowitz, M. (1967). Visual imagery and cognitive organization. *American Journal of Psychiatry*, *123*(8), 938–946.

Horst, S. (1999a). Symbols and computation a critique of the computational theory of mind. *Minds and Machines*, *9*(3), 347–381.

Horst, S. (1999b). *Symbols, computation, and intentionality*. University of California Press.

Horty, J. F. (2012). *Reasons as defaults*. Oxford University Press.

Hout, M., Papesh, M., & Goldinger, S. (2013). Multidimensional scaling. *WIREs: Cognitive Science*, *4*(1), 93–103.

Hume, D. (1894). *An enquiry concerning the human understanding: And an enquiry concerning the principles of morals*. Clarendon Press.

Hutchins, E. (2010). Cognitive ecology. *Topics in cognitive science*, *2*(4), 705–715.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. Psychology Press.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. Harper and Row.

Jackendoff, R. (1976). Toward an explanatory semantic representation. *Linguistic inquiry*, *7*(1), 89–150.

Jackendoff, R. (1981). On katz's autonomous semantics. *Language*, 425–435.

Jackendoff, R. (1992). *Semantic structures* (Vol. 18). MIT Press.

Jackendoff, R. (2002). *Foundations of language*. Oxford University Press.

Jackendoff, R. (2017). In defense of theory. *Cognitive science*, *41*, 185–212.

Jackendoff, R., & Landau, B. (1993). "what" and "where" in spatial language and spatial cognition. *Behavioral and brain sciences*, *16*(2), 217–265.

Jackson, F. (1997). Mental causation without the language of thought. In M. L. DallaChiara, K. Doets, D. Mundici, & J. V. Benthem (Eds.), *Structures and norms in science* (pp. 303–318). Springer.

Jäger, G. (2007). The evolution of convex categories. *Linguistics and Philosophy*, *30*(5), 551–564.

Johannesson, M. (2002). *Geometric models of similarity* (Vol. 90). Lund University.

Johnson-Laird, P. N. (1981). Comprehension as the construction of mental models. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *295*(1077), 353–374.

Johnson-Laird, P. N. (1982). Formal semantics and the psychology of meaning. In E. Saarinen & S. Peters (Eds.), *Processes, beliefs, and questions* (pp. 1–68). Springer.

Johnson-Laird, P. N. (1983). *Mental models*. Harvard University Press.

Johnson-Laird, P. N. (2002). Peirce, logic diagrams, and the elementary operations of reasoning. *Thinking & Reasoning*, *8*(1), 69–95.

Johnson-Laird, P. N. (2010a). Against logical form. *Psychologica Belgica*, *50*(3-4).

Johnson-Laird, P. N. (2010b). Deductive reasoning. *WIREs Cognitive Science*, *1*(1), 8–17.

Johnson-Laird, P. N. (2012). Inference with mental models. In J. Holyoak & R. G. Morrison (Eds.), *The oxford handbook of thinking and reasoning* (pp. 134–145). Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. (1991). *Deduction.* Lawrence Erlbaum Associate.

Johnson-Laird, P. N., Byrne, R., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological review*, *99*(3), 418.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni,

J.-P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychological review*, *106*(1), 62.

Jones, D. (2010). Human kinship, from conceptual structure to grammar. *Behavioral and Brain Sciences*, *33*(5), 367.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and brain sciences*, *34*(4), 169.

Jönsson, M. L. (2014). Semantic holism and language learning. *Journal of Philosophical Logic*, *43*(4), 725–759.

Katz, J. (1975). Logic and language: An examination of recent criticism of intensionalism. In K. Gunderson (Ed.), *Language, mind, and knowledge* (p. 36-130). University of Minnesota Press, Minneapolis.

Katz, J. (1992). The new intensionalism. *Mind*, *101*(404), 689–719.

Katz, J. (2004). *Sense, reference, and philosophy*. Oxford University Press.

Keil, F. (1979). *Semantic and cognitive developmen*. Harvard University Press.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Keßler, C., Raubal, M., & Janowicz, K. (2007). The effect of context on semantic similarity measurement. In *Otm confederated international conferences" on the move to meaningful internet systems"* (pp. 1274–1284).

Kiefer, F. (1988). Linguistic, conceptual and encyclopedic knowledge. In *Proceedings of the 3rd euralex international congress. budapest: Akadémiai kiadó* (pp. 1–10).

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological review*, *107*(4), 852.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.

Kremer, M. (2010). Representation or inference: must we choose? should we? In B. Weiss & J. Wanderer (Eds.), *Reading brandom* (pp. 237–256). Routledge.

Lakemeyer, G., & Nebel, B. (1994). Foundations of knowledge representation and reasoning. In G. Lakemeyer & B. Nebel (Eds.), *Foundations of*

knowledge representation and reasoning (pp. 1–12). Springer.

Lakkof, G. (2017). Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 63–100). Cambridge University Press.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by.* University of Chicago press.

Lampert, M. (2009). *Attention and recombinance.* Peter Lang.

Land, S. K. (1974). *From signs to propositions.* Longman.

Landy, D., Allen, C., & Zednik, C. (2014). A perceptual account of symbolic reasoning. *Frontiers in psychology*, *5*, 1-10.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.

Langacker, R. W. (2000). *Grammar and conceptualization* (Vol. 14). Walter de Gruyter.

Lange, M. (2000). *Natural laws in scientific practice.* Oxford University Press.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, *11*(1), 65–100.

Lascarides, A., & Asher, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and philosophy*, *16*(5), 437–493.

Lehmann, D., & Magidor, M. (1992). What does a conditional knowledge base entail? *Artificial intelligence*, *55*(1), 1–60.

Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1), 1–47.

Leslie, S.-J., & Gelman, S. (2012). Quantified statements are recalled as generics. *Cognitive psychology*, *64*(3), 186–214.

Levinson, S. (2003). *Space in language and cognition: Explorations in cognitive diversity.* Cambridge University Press.

Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to philosophy of mind* (p. 412–431). Blackwell Publishers.

Lewis, M., & Lawry, J. (2016). Hierarchical conceptual spaces for concept combination. *Artificial Intelligence*, *237*, 204–227.

Lieto, A., Chella, A., & Frixione, M. (2017). Conceptual spaces for cognitive architectures. *Biologically inspired cognitive architectures*, *19*, 1–9.

Lieto, A., Minieri, A., Piana, A., & Radicioni, D. P. (2015). A knowledge-based system for prototypical reasoning. *Connection Science*, *27*(2), 137–152.

Lieto, A., & Pozzato, G. (2019). A description logic framework for commonsense conceptual combination integrating typicality, probabilities and cognitive heuristics. *Journal of Experimental & Theoretical Artificial Intelligence*, *32*(5), 1–36.

Lieto, A., & Pozzato, G. L. (2018). A description logic of typicality for conceptual combination. In M. Ceci, N. Japkowicz, J. Liu, G. Papadopoulos, & Z. Raś (Eds.), *Foundations of intelligent systems* (pp. 189–199).

Lifschitz, V., Morgenstern, L., & Plaisted, D. (2008). Knowledge representation and classical logic. In *Handbook of knowledge representation* (Vol. 3, pp. 3–88). Elsevier.

Lin, P.-J., Schwanenflugel, P., & Wisenbaker, J. (1990). Category typicality, cultural familiarity, and the development of category knowledge. *Developmental Psychology*, *26*(5), 805—813.

Lindstromberg, S. (2010). *English prepositions explained.* John Benjamins Publishing.

Lloyd, R. (1993). Cognitive processes and cartographic maps. *Advances in psychology*, *96*, 141–169.

Locke, J. (1979). *An essay concerning human understanding* (P. Nidditch, Trans.). Clarendon Press.

López, A., Gelman, S., Gutheil, G., & Smith, E. (1992). The development of category-based induction. *Child development*, *63*(5), 1070–1090.

Lormand, E. (1990). Framing the frame problem. *Synthese*, *82*(3), 353–374.

Lowe, E. (1993). Rationality, deduction and mental models. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 211–230). Taylor & Frances/Routledge.

Lyons, J. (1996). *Linguistic semantics.* Cambridge University Press.

MacFarlane, J. (2000). *What does it mean to say that logic is formal?* (Unpublished doctoral dissertation). University of Pittsburgh.

MacFarlane, J. (2002). Frege, kant, and the logic in logicism. *The philosophical review*, *111*(1), 25–65.

MacFarlane, J. (2010). Pragmatism and inferentialism. In B. Weiss & J. Wanderer (Eds.), *Reading brandom. on making it explicit* (pp. 81–95). Routledge.

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

Macnamara, J. (1986). *A border dispute: The place of logic in psychology*. MIT Press.

Maddox, T. (1992). Perceptual and decisional separability. In G. Ashby (Ed.), *Multidimensional models of perception and cognition* (p. 147–180). Lawrence Erlbaum Associates.

Magnani, L. (2002). Epistemic mediators and model-based discovery in science. In L. Magnani & N. J. Nersessian (Eds.), *Model-based reasoning* (pp. 305–329). Springer.

Magnani, L. (2004). Reasoning through doing. *Journal of Applied Logic*, *2*(4), 439–450.

Magnani, L. (2011). *Abduction, reason and science*. Springer Science & Business Media.

Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in cognitive sciences*, *15*(3), 97–103.

Mandler, J. M., Bauer, P. J., & McDonough, L. (1991). Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology*, *23*(2), 263–298.

Manktelow, K., & Evans, J. S. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology*, *70*(4), 477–488.

Manktelow, K., & Over, D. (1990). Deontic thought and the selection task. In N. Wetherick, K. Gilhooly, K. Gilhooly, M. Keane, R. Logic, & G. Erdos (Eds.), *Lines of thinking* (pp. 91–114). Jons Wiley & Sons.

Manktelow, K., & Over, D. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, *39*, 1–105.

Marchetti, G. (2015). Attentional semantics: An overview. In G. Marchetti, G. Benedetti, & A. Alharbi (Eds.), *The attentional basis of meaning*

(p. 33-76). Nova Science Publishers.

Marconi, D. (1997). *Lexical competence*. MIT Press.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & cognition*, *17*(1), 11–17.

Marras, A. (1973). On sellars' linguistic theory of conceptual activity. *Canadian Journal of Philosophy*, *2*(4), 471–483.

McGinn, C. (1989). *Mental content*. Blackwell.

McRae, K. (2004). Semantic memory: Some insights from feature-based connectionist attractor networks. *The psychology of learning and motivation: Advances in research and theory*, *45*, 41–86.

Medin, D., Coley, J., Storms, G., & Hayes, B. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517–532.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254-278.

Medin, D., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual review of psychology*, *51*(1), 121–147.

Mercier, H. (2012). Looking for arguments. *Argumentation*, *26*(3), 305–324.

Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences. In J. S. Evans & K. Frankish (Eds.), *In two minds: dual process and beyond* (p. 149—170). Oxford University Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57-74.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(1), 89–115.

Mezzadri, D. (2018). Logic, judgment, and inference. *Journal of the History of Philosophy*, *56*, 727–746.

Minsky, M. (1974). A framework for representing knowledge. *Artificial Intelligence*, *12*.

Minsky, M. (1991). Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI magazine*, *12*(2), 34–34.

Moktefi, A., & Shin, S.-J. (2013). *Visual reasoning with diagrams.* Springer Science & Business Media.

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. *Ideas in context*, *52*, 10–37.

Murphy, G. (2004). *The big book of concepts.* MIT Press.

Murphy, L. (2003). *Semantic relations and the lexicon.* Cambridge University Press.

Nagy, W., & Gentner, D. (1990). Semantic constraints on lexical categories. *Language and Cognitive Processes*, *5*(3), 169–201.

Nersessian, N. J. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. J, Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 5–22). Springer.

Nersessian, N. J. (2006). Model-based reasoning in distributed cognitive systems. *Philosophy of science*, *73*(5), 699–709.

Nersessian, N. J. (2010). *Creating scientific concepts.* MIT Press.

Newton, I. (1999 [1687]). *The principia: mathematical principles of natural philosophy.* University of California Press.

Niiniluoto, I. (1987). *Truthlikeness.* Reidel.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*(4), 647–670.

Norton, J. D. (2010). There are no universal rules for induction. *Philosophy of Science*, *77*(5), 765–777.

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology*, *115*(1), 39–61.

Oaksford, M., & Chater, N. (1989). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning.* Psychology Press.

Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind & Language*, *6*(1), 1–38.

Oaksford, M., & Chater, N. (2003). Optimal data selection. *Psychonomic Bulletin & Review*, *10*(2), 289–318.

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

Oaksford, M., & Chater, N. (2009). Précis of bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*(1), 69–84.

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, *53*(3), 238–283.

O'Brien, D. P., Braine, M., & Yang, Y. (1994). Propositional reasoning by mental models? simple to refute in principle and in practice. *Psychological Review*, *101*(4), 711.

Osherson, D. (1975a). *Logical abilities in children*. Lawrence Erlbaum.

Osherson, D. (1975b). Logic and models of logical thinking. In R. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 81–92). Erlbaum Hillsdale.

Osherson, D., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Osta-Vélez, M. (2019). The enigma of reason. *Philosophical Psychology*, *32*(6), 995–999.

Osta-Vélez, M. (2019). Methods of representation as inferential devices. *Journal for General Philosophy of Science*, *50*(2), 231–245.

Osta-Vélez, M., & Gärdenfors, P. (n.d.). *Nonmonotonic reasoning, expectation orderings, and conceptual spaces*. (submitted)

Osta-Vélez, M., & Gärdenfors, P. (2020a). Category-based induction in conceptual spaces. *Journal of Mathematical Psychology*, *90*, 102357.

Overton, W. F. (1990). Competence and procedures. In W. F. Overton (Ed.), *Reasoning, necessity, and logic* (pp. 1–32). Lawrence Elrbaum.

Owen, D. (1999). *Hume's reason*. Oxford University Press.

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press.

Palmerino, C. R. (2010). The geometrization of motion. *Early Science and Medicine*, *15*(4-5), 410–447.

Palmieri, P. (2003). Mental models in galileo's early mathematization of nature. *Studies in History and Philosophy of Science Part A*, *34*(2), 229–264.

Panza, M. (2002). Mathematisation of the science of motion and the birth of analytical mechanics: A historiographical note. In P. Cerrai, P. Freguglia,

& C. Pellegrini (Eds.), *The application of mathematics to the sciences of nature* (pp. 253–271). Springer.

Paradis, C. (2003). Is the notion of linguistic competence relevant in cognitive linguistics. *Annual Review of Cognitive Linguistics*, *1*(1), 207–231.

Parsons, D. (2016). *Theories of intensionality: A critical survey.* Springer.

Partee, B. H. (2014). A brief history of the syntax-semantics interface in western formal linguistics. *Semantics-Syntax Interface*, *1*(1), 1–20.

Peacocke, C. (1992). *A study of concepts.* MIT Press.

Peacocke, C. (1999). Computation as involving content: A response to egan. *Mind & Language*, *14*(2), 195–202.

Peirce, C. S. (1868). Some consequences of four incapacities. *The Journal of Speculative Philosophy*, *2*(3), 140–157.

Piaget, J. (1947). *La psychologie de l'intelligence.* Armand Collin.

Piaget, J. (1949). *Traité de logique: essai de logique opératoire.* Armand Collin.

Piaget, J. (1956). Les stades du développement intellectuel de l'enfant et de l'adolescent. In M. Osterrieth et al. (Eds.), *Le problèmes des stades en psychologie de l'enfant* (pp. 33–99). PUF.

Piaget, J. (1957). *Logic and psychology.* Basic Books.

Piaget, J., & Garcia, R. (1990). *Toward a logic of meanings.* Lawrence Erlbaum.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, *37*(1), 1–38.

Pietroski, P. M. (2017). Semantic internalism. In J. Mcgilvray (Ed.), *The cambridge companion to chomsky* (pp. 196–216). Cambridge University Press.

Pollard, P., & Evans, J. S. (1987). Content and context effects in reasoning. *The American journal of psychology*, *100*(1), 41–60.

Prinz, J. J. (2004). *Furnishing the mind.* MIT Press.

Proffitt, J. B., Coley, J., & Medin, D. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 811.

Putnam, H. (1975). The meaning of "meaning". In K. Gunderson (Ed.), *Language, mind, and knowledge* (pp. 131–193). University of Minnesota

Press.

Pylyshyn, Z. W. (1981). The imagery debate. *Psychological Review*, *88*(1), 16.

Pylyshyn, Z. W. (2002). Mental imagery. *Behavioral and brain sciences*, *25*(2), 157.

Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology*, *63*(9), 1716–1739.

Quillan, M. R. (1966). *Semantic memory* (Tech. Rep.). Cambridge, MA.: Bolt Beranek and Newman INC.

Quillian, M. R. (1967). Word concepts. *Behavioral science*, *12*(5), 410–430.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, *60*, 20–43.

Quine, W. V. O. (1956). Quantifiers and propositional attitudes. *the Journal of Philosophy*, *53*(5), 177–187.

Quine, W. V. O. (1969a). Natural kinds. In N. Rescher (Ed.), *Essays in honor of carl g. hempel* (pp. 5–23). Springer.

Quine, W. V. O. (1969b). *Ontological relativity and other essays*. Columbia University Press.

Quine, W. V. O. (1974). *The roots of reference*. Open Court.

Quine, W. V. O. (1981). Five milestones of empiricism. In *Theories and things* (pp. 70–71). Harvard University Press.

Quine, W. V. O. (1986). *Philosophy of logic*. Harvard University Press.

Quine, W. V. O. (2013). *Word and object*. MIT Press.

Read, D. (2013). Reconstructing the proto-polynesian terminology: Kinship terminologies as evolving logical structures. *Kinship systems: Change and reconstruction*, 59–91.

Read, D., Fischer, M., & Lehman, F. (2014). The cultural grounding of kinship. *L'Homme. Revue française d'anthropologie*(210), 63–89.

Read, S. (1988). *Relevant logic*. Basil Blackwell.

Read, S. (1994). Formal and material consequence. *Journal of Philosophical Logic*, *23*(3), 247–265.

Rehder, B. (2006). When similarity and causality compete in category-based

property generalization. *Memory & Cognition*, *34*(1), 3–16.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories. *Journal of Experimental Psychology: General*, *130*(3), 323.

Reiter, R., & Criscuolo, G. (1981). On interacting defaults. In *Ijcai* (Vol. 81, pp. 270–276). Morgan Kaufmann Publishers.

Rescher, N. (1959). The distinction between predicate intension and extension. *Revue philosophique de Louvain*, *57*, 623–636.

Rescorla, M. (2012). Are computational transitions sensitive to semantics? *Australasian Journal of Philosophy*, *90*(4), 703–721.

Rescorla, M. (2018). Maps in the head. In K. Andrews & J. Beck (Eds.), *The routledge handbook of philosophy of animal minds* (p. 34-45). Routledge.

Rips, L. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*(6), 665–681.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge University Press.

Rips, L. (1994). *The psychology of proof.* MIT Press.

Rogers, T., & McClelland, J. (2004). *Semantic cognition.* MIT Press.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Rosch, E. (1971). "focal" color areas and the development of color names. *Developmental psychology*, *4*(3), 447–455.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and acquisition of language* (pp. 111–144). Elsevier.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Rosch, E. (1983). Prototype classification and logical classification. In E. Scholnick (Ed.), *New trends in conceptual representation: Challenges to piaget's theory* (pp. 73–86). Lawrence Erlbaum Associates.

Rosenberg, J. F. (1968). Wittgenstein's theory of language as picture. *American*

Philosophical Quarterly, *5*(1), 18–30.

Roux, S. (2010). Forms of mathematization (14th-17th centuries). *Early Science and Medicine*, *15*(4-5), 319–337.

Ruphy, S. (2011). From hacking's plurality of styles of scientific reasoning to "foliated" pluralism. *Philosophy of science*, *78*(5), 1212–1222.

Ryle, G. (1954). *Dilemmas: the tarner lectures 1953*. Cambridge University Press.

Sagi, G. (2018). Logicality and meaning. *The Review of Symbolic Logic*, *11*(1), 133–159.

Schaefer, R. (2016). Brandom's account of reasoning. *Journal of Philosophical Research*, *41*, 129-150.

Schemmel, M. (2008). *The english galileo*. Springer.

Scheutz, M. (1999). The ontological status of representations. In A. Riegler, M. Peschl, & A. von Stein (Eds.), *Understanding representation in the cognitive sciences* (pp. 33–38). Springer.

Schockaert, S., & Prade, H. (2013). Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. *Artificial Intelligence*, *202*, 86–131.

Schwanenflugel, P. J., & Rey, M. (1986). The relationship between category typicality and concept familiarity. *Memory & Cognition*, *14*(2), 150–163.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and brain Sciences*, *21*(1), 1–17.

Searle, J. R. (1990). Cognitive science and the computer metaphor. In B. Göranzon & M. Florin (Eds.), *Artifical intelligence, culture and language* (pp. 23–34). Springer.

Sellars, W. (1948). Concepts as involving laws and inconceivable without them. *Philosophy of Science*, *15*(4), 287–315.

Sellars, W. (1950). Language, rules and behavior. In S. Hook (Ed.), *John dewey: Philosopher of science and freedom* (pp. 129–155).

Sellars, W. (1953). Inference and meaning. *Mind*, *62*(247), 313–338.

Sellars, W. (1958). Counterfactuals, dispositions, and the causal modalities. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories, and the*

mind-body problem (p. 225-308). University of Minnesota Press.

Sellars, W. (1974). Meaning as functional classification. *Synthese*, *27*(3-4), 417–437.

Sellars, W. (1991). *Science, perception and reality.* Ridgeview.

Sellés, M. A. (2006). Infinitesimals in the foundations of newton's mechanics. *Historia Mathematica*, *33*(2), 210–223.

Shafto, P., Coley, J. D., & Vitkin, A. (2007). Availability in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 114–136). Cambridge University Press.

Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society*, *114*, 123–144.

Shea, N. (2018). *Representation in cognitive science.* Oxford University Press.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Sievert, D. (1989). Another look at wittgenstein on color exclusion. *Synthese*, *78*, 291–318.

Simon, H. (1978). On the forms of mental representation. In W. Savage (Ed.), *Perception and cognition* (pp. 3–18). University of Minnesota Press.

Singh, S., & Karwayun, R. (2010). A comparative study of inference engines. In *2010 seventh international conference on information technology: New generations* (pp. 53–57).

Sloman, S. (1993). Feature-based induction. *Cognitive psychology*, *25*(2), 231–280.

Sloman, S. (1998a). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33.

Sloman, S. (1998b). Categorical inference is not a tree: The myth of inheritance hierarchies. *Biologically inspired cognitive architectures*, *35*(1), 1–33.

Sloman, S., & Lagnado, D. (2005). The problem of induction. In K. Holyoak & R. Morrison (Eds.), *The cambridge handbook of thinking and reasoning* (pp. 95–116). Cambridge University Press.

Smith, E., Shoben, E., & Rips, L. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review*, *81*(3), 214-241.

Smith, N. J. (2009). Frege's judgement stroke and the conception of logic as the study of inference not consequence. *Philosophy Compass*, *4*(4), 639–665.

Smolensky, P. (2012). Symbolic functions from neural computation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *370*(1971), 3543–3569.

Sowa, J. (1991). Toward the expressive power of natural language. In J. Sowa (Ed.), *Principles of semantic networks* (pp. 157–189). Elsevier.

Sowa, J. (1999). *Knowledge representation*. Brooks/Cole Publishing Co.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Sperber, D. (2001). In defense of massive modularity. In E. Dupoux (Ed.), *Language, brain and cognitive development: Essays in honor of jacques mehler* (pp. 47–57). MIT Press.

Sperber, D., & Wilson, D. (1986). *Relevance*. Harvard University Press.

Staffel, J. (2013). Can there be reasoning with degrees of belief? *Synthese*, *190*(16), 3535–3551.

Stalnaker, R. (1981). Anti-essentialism. *Midwest Studies of Philosophy*, *4*, 343–355.

Stalnaker, R. C. (1968). A theory of conditionals. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 41–55). Springer.

Stenning, K., & van Lambalgen, M. (2011). Reasoning, logic, and psychology. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 555–567.

Storjohann, P. (2016). Sense relations. In N. Riemer (Ed.), (pp. 248–265). Routledge London.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of science*, *71*(5), 767–779.

Sutherland, S. L., & Cimpian, A. (2017). Inductive generalization relies on category representations. *Psychonomic bulletin & review*, *24*(2), 632–636.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449–508.

Swoyer, C. (1995). Leibniz on intension and extension. *Noûs*, *29*(1), 96–114.

Talmy, L. (1975). Figure and ground in complex sentences. In *Annual meeting of the berkeley linguistics society* (Vol. 1, pp. 419–430).

Talmy, L. (1983). How language structures space. In H. Pick & L. Acredolo (Eds.), *Spatial orientation* (pp. 225–282). Springer.

Talmy, L. (2007). Attention phenomena. In D. Geeraerts & H. Cuyckens (Eds.), *The oxford handbook of cognitive linguistics.* Oxford University Press.

Tenenbaum, J., Kemp, C., & Shafto, P. (2007). Theory-based bayesian models of inductive reasoning. inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 167–204). Cambridge University Press.

Thagard, P. (1984). Frames, knowledge, and inference. *Synthese*, *61*(2), 233–259.

Thagard, P. (1992). *Conceptual revolutions.* Princeton University Press.

Thompson, V. A. (1996). Reasoning from false premises: The role of soundness in making logical deductions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *50*(3), 315.

Toulmin, S. (1953). *An introduction to philosophy of science.* Hutchinson University Press.

Toulmin, S. (1961). *Foresight and understanding.* Hutchinson & CO.

Toulmin, S. (1971). From logical systems to conceptual populations. In R. C. Buck & R. S. Cohen (Eds.), *Psa 1970* (pp. 552–564). Springer.

Toulmin, S. (1972a). *Human understanding.* Clarendon Press.

Toulmin, S. (1972b). Rationality and scientific discovery. In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1972, pp. 387–406). D. Reidel Publishing.

Toulmin, S. (1974). Scientific strategies and historical change. In R. J. Seeger & R. S. Cohen (Eds.), *Philosophical foundations of science* (pp. 401–414). Springer.

Toulmin, S. (2003). *The uses of argument.* Cambridge University Press.

Touretzky, D. S. (1984). Implicit ordering of defaults in inheritance systems.

In *Aaai-84 proceedings* (pp. 322–325). Morgan Kaufmann Publishers.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

Tyler, A., & Evans, V. (2003). *The semantics of english prepositions*. Cambridge University Press.

Ungerer, F., & Schmid, H.-J. (2006). *An introduction to cognitive linguistics*. Pearson-Longman.

Valaris, M. (2017). What reasoning might be. *Synthese*, *194*(6), 2007–2024.

van der Henst, J.-B. (2000). Mental model theory and pragmatics. *Behavioral and Brain Sciences*, *23*, 283–284.

Van Heijenoort, J. (1967). Logic as calculus and logic as language. *Synthese*, *17*, 324–330.

von der Emde, G. (2004). Distance and shape: perception of the 3-dimensional world by weakly electric fish. *Journal of Physiology*, *98*(1-3), 67–80.

Von Eckardt, B. (1995). *What is cognitive science?* MIT Press.

Von Eckardt, B. (2005). Connectionism and the propositional attitudes. In C. Erneling & D. Johnson (Eds.), (p. 225). Oxford University Press.

Vorms, M. (2011). Representing with imaginary models: Formats matter. *Studies in History and Philosophy of Science Part A*, *42*(2), 287–295.

Warglien, M., Gärdenfors, P., & Westera, M. (2012). Event structure, conceptual spaces and the semantics of verbs. *Theoretical linguistics*, *38*(3-4), 159–193.

Wartofsky, M. (1983). From genetic epistemology to historical epistemology: Kant, marx, and piaget. In L. S. Liben (Ed.), *Piaget and the foundations of knowledge* (pp. 1–17). Lawrence Erlbaum.

Wartofsky, M. (1987). Epistemology historicized. In A. Shimony & D. Nails (Eds.), *Naturalistic epistemology* (pp. 357–374). Springer.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental*

psychology, *20*(3), 273–281.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly journal of experimental psychology*, *23*(1), 63–71.

Waxman, S. R., & Leddon, E. M. (2011). Early word-learning and conceptual development. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (p. 102–126). Wiley-Blackwell.

Weiskopf, D. A. (2009). Atomism, pluralism, and conceptual content. *Philosophy and Phenomenological Research*, *79*(1), 131–163.

Westphal, K. R. (2015). Conventionalism and the impoverishment of the space of reasons: Carnap, quine and sellars. *Journal for the History of Analytical Philosophy*, *3*(8), 1–63.

Wikforss, Å. (2008). Semantic externalism and psychological externalism. *Philosophy Compass*, *3*(1), 158–181.

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Wilson, R. A., & Keil, F. C. (2001). *The mit encyclopedia of the cognitive sciences*. MIT Press.

Wittgenstein, L. (2001 [1921]). *Tractatus logico-philosophicus* (D. Pears & B. McGuinness, Trans.). Routledge.

Woods, W. A. (1987). Knowledge representation: What's important about it? In N. Cercone & G. McCalla (Eds.), *The knowledge frontier* (pp. 44–79). Springer.

Yang, J., & Long, C. (2020). Common and distinctive cognitive processes between categorization and category-based induction: Evidence from event-related potentials. *Brain Research*, 147134.

Yee, E., Jones, M. N., & McRae, K. (2018). Semantic memory. In J. Wixted & S. Thompson-Schill (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 3, pp. 1–38). Wiley Online Library.

Zimmermann, T. E. (1999). Meaning postulates and the model-theoretic approach to natural language semantics. *Linguistics and Philosophy*, *52*, 529–561.

Zwarts, J., & Gärdenfors, P. (2016). Locative and directional prepositions in conceptual spaces: The role of polar convexity. *Journal of Logic, Language and Information*, *25*(1), 109–138.