Formal Semantics and the Psychology of Reasoning

Building new bridges and investigating interactions

by

Salvador Mascarenhas

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

New York University

September, 2014

Professor Philippe Schlenker

Acknowledgments

First of all I wish to thank my dissertation committee. All five members contributed greatly to the quality of the work presented here, so that it is hard to imagine what this dissertation would look like without their ideas and suggestions, their support, inspiration, and generosity toward me.

Since I took his seminar on presuppositions in the fall of 2008, Philippe Schlenker has been a major influence in my research. Most generally, and probably most importantly, Philippe's bold choices of research interests, as well as the ground breaking and careful nature of his work on those topics, played a key role in giving me the courage to dedicate my dissertation to topics not usually studied by linguists. More specifically, Philippe's contributions crucially shaped Chapters 3 and 4, and his comments on Chapter 2 made me see what aspects of that work need to be rethought in order to get the attention of linguists and philosophers.

Anna Szabolcsi has played a crucial role in my education and in my thinking also since the beginning of my graduate career at NYU. In particular, Anna helped me explore my earlier work on inquisitive semantics at a level of detail and linguistic responsibility that were new to me at the time. I also owe a lot of the professional skills I have acquired to conversations with Anna, often in scheduled appointments, but just as often in impromptu conversations. I am extremely grateful for her continued availability and interest in my work.

Chris Barker has given me intellectual and moral support more times than I can count. I thank him especially for his detailed comments on Chapter 2 and for multiple thought-provoking conversations about the topic of Chapter 4. I started interacting with Emmanuel Chemla and Kit Fine later

in my graduate career, but their influence in this dissertation was extremely important. Emmanuel has provided detailed and game-changing comments on Chapters 3 and 4, and it was at his suggestion that I pursued the topic in section 3.4. I've been privileged to discuss my research with Kit Fine on numerous occasions since I took his seminar on truth-maker semantics in the spring of 2012. The work in Chapter 2 of this dissertation owes a huge debt to Kit's recent research.

I interacted with almost the entirety of the NYU Linguistics faculty over the last several years, virtually always to my great benefit. I am especially grateful to Richie Kayne, Stephanie Harves, Alec Marantz, and Chris Collins, who played important roles in my education and as members of my QP committees. I was extremely lucky to have been surrounded by a group of active and brilliant graduate students in semantics; in chronological order: Dan Lassiter, Simon Charlow, Tim Leffel, Mike Solomon, Jeremy Kuhn, Dylan Bumford, and Linmin Zhang. Simon Charlow's influence has been particularly great, from the moment we both joined the linguistics program at NYU. Ever since I began work on the topic of this dissertation, Dylan Bumford has been an inspirational and generous interlocutor. Other NYU graduates with whom I had many formative discussions are Neil Myler, Inna Livitz, Jim Wood, and Tricia Irwin. I also want to thank Aura Holguin and Teresa Leung for saving me from my own administrative blunders on more occasions than I can remember, as well as Lisa Davidson, Alec Marantz, and Chris Barker for providing very helpful advice as DGS, respectively department chairs. Beyond NYU Linguistics, New York City is/was home to a number of great scholars with whom I was lucky to interact: Philipp Koralus, Jim Pryor, Phil Johnson-Laird, Maria Bittner, Orin Percus, Sarah Murray, Will Starr, Andreas Stokke, Gabe Greenberg, and Karen Lewis. I was also lucky to spend a semester at Institut Jean-Nicod, in Paris. Alongside Philippe Schlenker and Emmanuel Chemla, Benjamin Spector was a constant source of inspiration and constructive criticism. I also interacted very beneficially with Paul Egré, Dan Sperber, Guy Politzer, François Récanati, Heather Burnett, Guillaume Thomas, Yasu Sudo, and Jérémy Zehr.

My research program on questions, an important component of the theory developed in Chapter 2, began in Amsterdam. I am especially grateful to my masters thesis' co-advisers Jeroen Groenendijk and Dick de Jongh. I want to thank in addition Frank Veltman, Maria Aloni, Michiel van Lambalgen, Floris Roelofsen, Ivano Ciardelli, Benedikt Loewe, and Tanja Kassenaar. From Lisbon,

where I did my undergrad, I want to thank especially João Peres and Manuela Ambar. Also influential in my education were Inês Duarte, Ernesto d'Andrade, António Zilhão, Rui Marques, Ana Maria Martins, Fernando Ferreira, Cristina Morgado, and Zé Pedro Ferreira. The following scholars do not properly fall under any of the above institutional categories, but engaged me in multiple thought-changing discussions and/or provided crucial advice at different points of my graduate career: Thomas Icard, Seth Yalcin, Martin Hackl, Igor Yanovich, Craige Roberts, Gennaro Chierchia, Anastasia Giannakidou, and Daniel Rothschild.

For making life in New York great, I thank Tuuli, Andi, Tim L., Little Kevin, Bigger Kevin, Simon Y., Dickie P., Mel, Jordan, Elizabeth, Brian K., Michael, Natasha, Kit, Kat, Jane B., and Taylor. Special thanks to K&S, for showing me all the music and a lot of the fun NY had to offer.

Most of all I want to thank my wonderful parents and my brilliant brothers for the endless support throughout these many years. Finally I thank my grandmother Milu, who left us recently after a long and rich life, and whose wise advice and unwavering support helped give me the courage to try and figure out what I really care about in life and to pursue it to the best of my ability.

Salvador Mascarenhas

Oxford, UK, September 2014

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

Research in linguistic semantics in the past forty years has produced sophisticated mathematical models that represent the meanings of natural language utterances and explain how meanings relate to one another to form entailment patterns. At the same time, research on human reasoning within psychology has discovered a wealth of fallacious inference patterns, establishing that human reasoning is fallible in highly predictable ways. The psychological study of reasoning has been characterized by extensive experimental work analyzed in terms of often informal theories focusing on the processes of reasoning. Although data on reasoning are almost always collected by means of linguistically presented stimuli, psychology has not systematically sought to ground this research in interpretive insights from linguistics. On the other hand, linguistic semantics has a longstanding tradition of focusing on normatively sanctioned inference patterns, and it has so far largely ignored fallacies. The two domains of research overlap significantly, but they have progressed almost completely in parallel, with little interaction. This dissertation contributes to furthering the connection between the two fields. The work herein aims at making contributions of two kinds toward this goal.

Mapping sentences to mental representations One of the reasons for the disconnect relates to the Platonic nature of modern work in semantics. At least since Davidson (1967), most of the research done in philosophy of language and in linguistic semantics has focused on investigating mappings between linguistic expressions and truth conditions (or constitutive elements of truth con-

ditions in the case of sub-clausal expressions). Since it is essentially established that truth and truth-preservation are best explained by classical model-theory and logic, the Davidsonian program for semantics has introduced a powerful bias in model-theoretic semantics in favor of staying close to classical models whenever possible. This program has been highly successful in producing interesting and insightful research on language. Moreover, insofar as one is interested in mapping linguistic expressions to objective externalities (e.g. entities in the Platonic world in the case of sentences, the real world in the case of most theories of reference), the Platonic program is surely on the right track. It is however a different question whether this line of research holds the most promise as a program for semantics as cognitive science, where the explanatory power of such externalities is far from obvious. I am persuaded that we can do better, and will presently outline a simple program for semantics that can help bridge the gap between semantics and psychology and that preserves the formal rigor and model-theoretic approach that characterizes mainstream Platonic natural language semantics.

I propose that, instead of mapping expressions to truth, we map expressions to mental representations, whose properties can be independently elucidated using the methods of psychology.[1] This more cognitive program for semantics is certainly related to the Platonic one. The past fifty years of research in psychology have shown that mental representations do not capture truth entirely faithfully, and that the mind does not put at our disposal perfectly truth-preserving operations on mental representations. But clearly those representations and operations are in some sense *attempting* to capture truth. Tracking truth closely enough offers obvious selective advantages, and thus all things being equal we expect that evolution should reward the kind of mind that does not veer very far from the truth when representing the world and manipulating representations of the world. In this sense, we expect mental representations to be *related* to classical models of truth, and the operations on mental representations to be related to truth-preserving operations. There are many insights on the psychological program to be gained from the Platonic program. But there is little reason to expect, and no reason to assume, that the mental structures in charge of representing the world should do so

[1] Stated in these general terms, the idea is by no means novel, and finds an important precedent for example in the work of Jackendoff (1983). In fact, the specific program I will propose shortly and develop in Chapter 2 can be seen as an application of Jackendoff's general program to the specific case of interactions and overlaps between semantics and reasoning.

completely faithfully.[2]

A few words are in order to clarify further what I mean by "truth conditional" semantics and what properly falls under the Platonic heading. If by "truth conditional" I mean "classical logical," it would seem that the paragraphs above grossly misrepresent modern research in linguistic semantics. While there may well still be such a thing as a "mainstream semantics" that does not admit of non-classical representations at any level of analysis, there also exist influential frameworks that radically depart from classical logic and do not appear to display the classical bias I pointed out earlier. A prime example of such a framework is dynamic semantics (Heim, 1982; Kamp, 1981; Groenendijk and Stokhof, 1991), which eschews certain core properties of classical logic (most notably commutativity) to gain expressive power, broadening the scope of semantic analysis to domains previously (or rather, otherwise) relegated to different modules. In my view, while dynamic semantics has produced highly informative and unique insights about language, of great relevance to any program for linguistic semantics, it is still properly Platonic. This is so for two reasons. The first reason is conceptual in nature, and the less philosophically inclined reader is welcome to skip it: most dynamic semantics are still *ultimately concerned with truth* in a classical sense. These frameworks will often start by setting up dynamic (and thus non-classical) systems specially fashioned to permit the system to be sensitive (to *see*) certain distinctions between (the meanings of) linguistic expressions that are invisible to classical systems. But classical truth conditions are usually the stated ultimate goal. Some frameworks offer explicit rules for lifting and lowering

[2]Most if not all psychologists working on reasoning would agree that a theory of truth-preservation (i.e., classical logic) does not constitute a meaningful level of psychological analysis, in the sense of Marr (1982). Similarly, in my view it is highly questionable whether mappings from expressions to truth conditions classically construed constitute a meaningful level of linguistic analysis under a cognitive-science program for linguistics. A good understanding of classical logic is indispensable to anyone working on a psychological theory of human reasoning, but classical logic does not offer a theory of reasoning, not even at the computational level. Similarly, a good understanding of Platonic truth-conditional semantics can be of great use to anyone working on a psychological theory of human language, but classical truth-conditional semantics does not offer a theory of language at the computational (competence) level.

The reader might find this footnote an exercise in (currently) inconsequential philosophical pedantry, but I wish to make this point quite clearly: it is my view that the Platonic project, despite its great merits and usefulness, will not deliver a theory of human semantic competence. To argue properly for this point would take many more pages than the insight to be gained deserves, so I stick with my analogy in hope that it will help nudge the right intuitions in the reader's mind: classical-truth-centered semantics is to the human semantic faculties as classical logic is to the human reasoning faculties. Both are tightly and interestingly related to their respective human faculties, but neither offers a theory of competence of those faculties. Classical logic has emancipated itself entirely from the notion that it could be a psychological theory in any capacity, but classical-truth semantics still very often purports to be just that, using ill grounded appeals to the competence-performance distinction in an attempt to justify its disregard for some of the relevant psychological realities.

expressions from and into truth conditions (e.g. Groenendijk and Stokhof, 1990), while others define truth criteria that are meant to reduce a rich dynamic representation to a simple classical model. But what ill, one might ask, can come from considering truth criteria *in addition* to other semantic properties of expressions? The danger, as I see it, is that this concern for classical truth is not just unwarranted under a psychological program for semantics, it is also liable to lead us in the wrong direction and to steer us away from crucial insights.

The second reason why most dynamic semantics frameworks are properly Platonic projects relates to this worry of missing insights and connections. Say we disregard talk of lowering operations and of truth criteria, preserving all the way the non-classical nature of these systems. The non-classical models proposed by most of these theories are still motivated by exclusively linguistic data, and their predictions are tested by looking at what other exclusively linguistic data are or fail to be explained in their account of interpretive processes. These are two properties characteristic of the Platonic enterprise that can and must be complemented by other, more broadly *psychological* criteria for guiding theory building and theory testing. Specifically, I propose that the properties of mental representations argued for by various theories within the psychology of reasoning should constitute a primary source of data (in the form of natural inference-making behavior) and of formal constraints for natural language semantics as a cognitive science.[3]

Chapter 2 of this dissertation pursues this program, by building on the fact that certain non-truth-conditional semantics (inquisitive semantics and some of its relatives) share interesting properties with an entirely independently proposed theory of human reasoning from psychology (mental model theory). Inquisitive semantics and mental model theory both posit non-classical, non-truth-conditional representations for sentences containing disjunctions, and Chapter 2 combines the two approaches in one unified theory. In the process, independent and reciprocal justifications appear for inquisitive semantics and mental model theory. The success of mental model theory in predicting observed inference-making behavior is predicated on a theory of mental representations and the relation between sentences of natural languages and those representations. This theory of mental

[3]I do not mean to suggest that the properties of mental representations are somehow a resolved question in the psychology of reasoning. To the contrary, different schools make different claims about mental representations, and unfortunately many schools make only tacit commitments about mental representations. There is therefore room for seeking convergence between many different pairs of semantic theory and theory of reasoning.

representations bears striking connections to the models of inquisitive semantics, which has independent and strictly linguistic motivations. This work demonstrates how non-trivial convergences between linguistics and the psychology of reasoning are not only possible in principle, but can be attained in practice by looking at the space of theories already on the market in both fields, and seeing what theories can be combined in simple and insightful ways.

Exploring the line between reasoning and interpretation Chapters 3 and 4 explore a less original point of contact between semantics and psychology, but an absolutely indispensable and vastly understudied one. The psychology of reasoning overwhelmingly collects its data by means of experiments with linguistically presented stimuli. There are very interesting exceptions to this tendency, exploring reasoning about pictorially presented information for example, but the vast majority of research in the field uses language. Experimental subjects then have two tasks to perform. First they must decode the linguistic signal using their faculty of language, then they must manipulate the representations they arrive at using their faculty for reasoning. The two processes are distinct, and most likely also different. When we say "I understood the problem, but I arrived at the wrong answer," we feel that we are saying something substantive, rather than a contradiction.

In the second part of this dissertation I explore two cases of (superficially) fallacious inferential behavior for which there exist interpretation-based explanations. Chapter 3 looks at illusory inferences from disjunction, which were also studied in Chapter 2 from a reasoning-based perspective. Chapter 3 succeeds at formulating an interpretive account, but does not fully answer the question of which perspective is right (reasoning in Chapter 2, interpretation in Chapter 3). Chapter 3 concludes with a puzzle for the interpretive theories used earlier, arguing that a particular technical aspect of those theories is at odds with commonly accepted facts about psychology.

Chapter 4 focuses on the most well-known instance of fallacious reasoning, the conjunction fallacy. In it, I argue that the conjunction fallacy also has an alternative explanation in terms of interpretation. I conclude Chapter 4 with first steps toward a paradigm that can separate the predictions of reasoning-based and interpretation-based accounts of the conjunction fallacy.

Chapter 2

Question-semantics in propositional reasoning

2.1 Introduction

I introduce in this chapter a new theory of reasoning, the *erotetic theory of reasoning*, based on the idea that the relationship between questions and answers is central to both our successes and failures of reasoning.[4] The core of the proposal is the erotetic principle:

(1) **The erotetic principle**

Part I — Our natural capacity for reasoning proceeds by treating successive premises as questions and maximally strong answers to them.

Part II — Systematically asking a certain type of question as we interpret each new premise allows us to reason in a classically valid way.

What the erotetic principle comes to will be developed in formal detail in the rest of this chapter. The erotetic theory of reasoning based on this principle combines two classes of ideas from the philosophy of language and natural language semantics. The first one is the idea that we in-

[4]The material in this chapter is an abridged version of joint work with Philipp Koralus, published as Koralus and Mascarenhas (2013). I largely keep first person plural pronouns in the prose to flag the co-authored nature of the work presented here. That said, since this is an adapted version, any errors found in this chapter should be attributed entirely to me and not to my collaborator in the original paper.

terpret sentences in the context of questions, represented in the form of mental models (Koralus, 2012).[5] The second class of ideas concerns enriched notions of linguistic content, as worked out in the frameworks of inquisitive semantics (Groenendijk, 2008; Mascarenhas, 2009) and truth-maker semantics (van Fraassen, 1969; Fine, 2012).

The intuition behind Part I of the erotetic principle is that the process of interpreting premises largely reduces to the search for answers to questions posed by other premises, regardless of whether those premises superficially look like questions. This approach will be grounded by a fresh look at the linguistic meaning of premise statements. Informally for now, a reasoner will take a disjunctive premise like "John and Bill are in the garden, or else Mary is" to pose the question of which of the disjuncts is the case. If she then accepts as a second premise that "John is in the garden," she will interpret it to be as strong an answer as possible to the question in context. As luck would have it, "John is in the garden" is part of the first possible answer to the question at hand, and not the second, so the reasoner will conclude that the question in context has been answered: "John and Bill are in the garden," deriving a well-known illusory inference (Walsh and Johnson-Laird, 2004). The inference is fallacious in that we are not entitled to make it given the information provided by the premise statements, on a linguistically plausible account of the meaning of those premise statements, an issue to which I will return. This example is just one of many data points on reasoning that need to be captured and that the theory in this chapter offers an insight on.

But we cannot only be interested in systematically capturing divergences in naive reasoning from what we are entitled to conclude from given premises. We also want to account for the fact that our reasoning endowment makes correct reasoning possible, as evidenced by the simple fact that it is possible to train oneself to reason in a manner that respects classical validity. This is where Part II of the erotetic principle comes in: although our natural capacity for reasoning is not specifically aiming at classical validity, there exists a systematic strategy of using questions, made available by our natural capacities, which would guarantee that we only make inferences we are in fact entitled to make. Toward the end of the chapter, I prove that this reasoning strategy exists. In

[5] Koralus (2012) proposes a theory of interpretation where sentence meaning corresponds to instructions to build mental representations. These instructions are often ambiguous or underspecified, allowing for multiple candidate mental model representations. Koralus argues that hearers decide between candidate representations by finding that representation that is the most responsive to the *questions* the hearer is attending to.

a certain specific sense I will develop below, *questions make us rational*. By the erotetic principle, our desire for answers makes our reasoning fallible, but our ability to ask the right questions puts us back on track.

The text in this chapter is almost entirely imported from Koralus and Mascarenhas (2013). I have kept first person plural pronouns as in the original article given that this chapter is the product of joint work, but I made small modifications to a number of sections in order to reduce redundancy with the rest of the dissertation.

2.2 The erotetic theory of reasoning

2.2.1 Preliminaries

2.2.1.1 The scope of this chapter

This chapter focuses on reasoning with premises involving propositional connectives, such as expressed by the English words 'or', 'and', 'not', and certain interpretations of 'if'. Within the domain of reasoning picked out, we address both the problem of failure and the problem of success. In other words, we seek to explain within a single system both how naive reasoning diverges from what we are entitled to conclude and how it is possible for us to come to reason correctly.

For these purposes, solving the problem of success means showing that there exists a reasoning strategy using our naive reasoning capacities that is classically sound, and, under performance idealizations, classically complete. The objective is to explain part of the puzzle of how science and philosophy are possible for humans. Since science and philosophy rely on classical standards of correctness, this is the standard we need to show is achievable.

Empirically, our aim is to account for what a particularly interesting set of data on propositional reasoning. The erotetic theory of reasoning developed below captures a significant catalog of empirically documented patterns of naive reasoning that diverge from classical norms, listed in Table 1 (on page 9). In addition to data from existing experimental literature, we present a novel illusory inference that no extant theory of reasoning captures.[6]

[6]This advantage seems to be largely due to the fact that the erotetic theory is more formally systematic than other

Conn.	Result	Reference	This chapter	Appendix
Not	Few list all alternatives compatible with negated conjunction	Khemlani et al. (2012)	exx. 16 & 17, p. 40	
Not	Most can list what corresponds to negated disjunction	Khemlani et al. (2012)		ex. 24, p. 133
Not	Easy to list a case that falsifies conditional	Oaksford and Stenning (1992)		ex. 25, p. 134
—	"Explosion" highly counterintuitive	Harman (1986)	ex. 21, p. 47	
Or	Disjunctive syllogism is harder than disjunctive modus ponens	Rips (1994)		ex. 26, p. 134
Or	Disjunctive syllogism easier if categorical premise comes first	García-Madruga et al. (2001)	ex. 13 & 14, p. 37	
Or	Illusory inferences from disjunction	Walsh and Johnson-Laird (2004)	ex. 10, p. 36	
Or	Control problems with disjunction	Walsh and Johnson-Laird (2004)		ex. 27, p. 135
Or	Supposition makes some problems easier	Johnson-Laird (2008)	exx. 20 & 11, p. 46	
Or	Disjunction introduction is counterintuitive	Braine et al. (1984)	ex. 19, p. 43	
Or/If	Fallacies with conditionals and disjunction	Johnson-Laird (2008)		ex. 28, p. 136
Or/If	Illusory inferences with conditional embedded in disjunction	Johnson-Laird and Savary (1999)		ex. 29, p. 137
Or/If	Control problems with conditionals embedded in disjunction	Johnson-Laird and Savary (1999)		ex. 30, p. 138
If	Modus ponens is extremely easy	Braine and Rumain (1983)	ex. 22, p. 48	
If	Modus ponens easier than modus tollens	Evans et al. (1993)	ex. 22, p. 48	
If	Affirming consequent more rapid than denying antecedent	Barrouillet et al. (2000)		ex. 32, p. 141
If	Order effects on modus tollens	Girotto et al. (1997)		ex. 33, p. 142
If	Illusions of consistency with sets of biconditional statements	Johnson-Laird et al. (2004)		ex. 34, p. 143
If	Control problems for biconditional consistency judgments	Johnson-Laird et al. (2004)		ex. 35, p. 144
If	Illusory inference from disjunction to conditional	Koralus and Mascarenhas (2013)	ex. 23, p. 50	

Table 1: Some core data on naive reasoning captured by the erotetic theory

2.2.1.2 Points of contact with other approaches

Before presenting the erotetic theory more fully, we briefly describe how it relates to some of the most influential existing approaches. The erotetic theory includes insights from what Oaksford and Chater (2007) have described as the "leading" formal psychological theories of reasoning, mental model theory (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991) and mental logic (Rips, 1994), as well as insights from the rational analysis approaches that are gaining increasing currency (Oaksford and Chater, 2007; Tenenbaum et al., 2006).

We take mental model theory, as developed by Johnson-Laird and collaborators, as a key inspiration, since it sheds light on a particularly interesting range of propositional reasoning patterns that we wish to understand. As far as we can see, no extant alternative account does a *better* job at covering these data points and others similar to it (Oberauer, 2006; Schroyens et al., 2001), regardless of various shortcomings that have been pointed out (Hodges, 1993; Stenning and van Lambalgen, 2008a). Moreover, we wish to keep mental models, made suitably precise, as a representational framework for our theory. Thus, classical mental model theory provides the relevant standard we wish to improve upon for the particular reasoning problems on which we focus in this chapter.

We take onboard the idea that we draw conclusions from premises by conjoining the mental model generated by the first premise with the mental models generated by subsequent premises, unless something prompts the reasoner to adopt a special strategy. A conclusion is taken to follow from the premises if it is represented in the resulting integrated mental model. In contrast to classical mental model theory, we will rely on a novel linguistically motivated account of how logical constructions in English are interpreted, following inquisitive semantics (Groenendijk, 2008; Mascarenhas, 2009) and truth-maker semantics (van Fraassen, 1969; Fine, 2012). A further difference from mental models is that on the erotetic theory, the process of integrating mental models supplied by premises has a very specific aim: posing and answering questions. Moreover, the erotetic theory, unlike classical mental model theory, is developed as a formal system, which means that predictions can be calculated mechanically, as they can, for example, in mental logic approaches (Rips, 1994).

---

theories of propositional reasoning that also involve mental models. This vindicates criticisms of mental model theory on the grounds that it lacks formal rigor, as made by Hodges (1993).

We also have important points of agreement with those defending rational analysis approaches to reasoning. A point in their favor, as remarked by Oaksford and Chater (2007), is that, to the extent that they capture the data, they have an explanatory advantage over theories like classical mental model theory. An explanation in terms of Bayesian updating of probability distributions is an explanation in terms of the cognitive system's computational aim in the most fundamental sense. In other words, it is an explanation in terms of *what reasoning is*, rather than an explanation in terms of how it happens to be implemented. By contrast, algorithmic implementation accounts, like standard mental model theory, can appear less explanatorily satisfying and are under threat of appearing like "Rube Goldberg" machines where we can see what the machine is doing but do not understand why things are done this way (*ibid.*). On our view, the kind of propositional reasoning we are interested in does not in itself justify a fundamentally probabilistic approach, but we share the view that it is better to provide an explanation in terms of the computational aim of the system.[7] The aim we propose for our naive reasoning capacity is answering questions.

Finally, like Stenning and van Lambalgen (2008a,b), we take the view that a theory of reasoning has to pay careful attention to how language is interpreted. We believe that this is best done by letting the interpretation component of the theory be informed by formal semantics as studied in linguistics and philosophy. The erotetic theory is uniquely well-fitted to provide formal foundations for mutually beneficial interactions between related branches of linguistic semantics, philosophy, and psychology. Johnson-Laird (Johnson-Laird and Stevenson, 1970; Johnson-Laird, 1970) pioneered the idea that successive utterances of sentences are interpreted by updating a mental representation of discourse, which was later independently proposed in a dynamic semantics turn in linguistics and philosophy (Karttunen, 1976; Heim, 1982; Kamp, 1981). These developments proceeded in parallel and with little to no interaction. The erotetic theory provides a preliminary but solid bridge to this gap, given the explicitly dynamic nature of the procedure that interprets new premises in the context of previous premises.

[7] The apparent tension between the approach in this chapter and that of probabilists like Oaksford and Chater (2007) is less acute than it might seem. First, Oaksford and Chater (2007) concede that logical reasoning is required alongside probabilistic reasoning. In particular, for the kinds of propositional reasoning problems we concentrate on in this chapter, the full power of Bayesian theory seems both unwarranted and inadequate. Second, we are entirely open to the possibility that Bayesian tools are necessary to account for certain aspects of human reasoning. As far as we can see, the two classes of theories can in principle (and perhaps *must*) be integrated.

We are convinced that time is ripe to explore a formal bridge to unify different research programs in semantics, philosophy, and the psychology of reasoning. Even though the erotetic theory is a novel proposal, it incorporates insights from many different approaches and, so we hope, will correspondingly find appeal among researchers from a variety of them.

2.2.2 The erotetic principle

We hold that default reasoning from a sequence of premises proceeds by updating a model of *discourse* with *mental models* representing the premises. Reasoners will take whatever holds in the resulting mental model to follow from those premises. In section 2.2.3 we introduce semi-formally the central components of the theory, to be fully formalized in section 2.2.4. For the present section, we concentrate on the principles whereby successive premises are combined, and explain intuitively what makes the account of reasoning in this chapter an erotetic account. Thus, the following informal definitions of mental models will suffice for our present purposes.

Mental models are the mental representations of premises. Setting aside possible contributions of background knowledge, the mental model representation of a premise is the *minimal* state of affairs compatible with that premise. Mental models are *minimal* in the sense that only the information explicitly given in the premise or specifically given by special background knowledge is represented in the mental model. Mental models are underspecified with respect to any information not given in this way (Johnson-Laird, 1983).

Mental models for premises are determined by an interpretation function from a natural language into mental models. This function is in principle sensitive to the linguistic expressions used. In particular, a connective like disjunction (English 'or', among others) will give rise to a mental model with two *alternatives*, one for each disjunct, modeling the fact that there are two minimal states of affairs compatible with a disjunctive sentence. By contrast, the mental model for a conjunction has only one alternative.

(2) Informal examples of mental model interpretations

a. John smokes. {*John smokes*}

b. John or Mary smokes. {*John smokes, Mary smokes*}

c. John and Mary smoke. {*John smokes & Mary smokes*}

Mental model discourses represent the workspace of reasoning. Mental model discourses are updated successively with the mental model interpretation of each premise, and they are the input and output categories for every operation of reasoning. Mental model discourses furthermore keep track of certain background and contextual information.

Default reasoning is the default procedure that reasoners engage in when faced with a set of premises for which they want to derive (or check) conclusions. Any theory of reasoning that has more than one reasoning operation, as ours does, must define a default reasoning procedure. Ours is as follows: reasoners update their mental model discourse with each premise in the order in which it is given, following the *erotetic principle* at each update step. Reasoners may then optionally perform very elementary operations on the resulting mental model, such as (an analog of) conjunction elimination. Whatever holds in the resulting mental model will be held by the reasoners to follow from the premises.

The erotetic principle is a distillation of the central and novel hypotheses of our account of reasoning. It is our proposed answer to the questions of (1) what the functional aim of human reasoning is and (2) how it is possible for trained human beings with unlimited time to, as it were, "emulate" sound classical reasoning, in view of the observed discrepancies between naive reasoning and classical logic. The erotetic principle has two parts: Part I holds that our natural capacity for reasoning proceeds by treating successive premises as questions and maximally strong answers to them. Part II holds that systematically asking a certain type of question as we interpret each new premise allows us to reason in a classically valid way.

In the remainder of this section we unpack the two parts of the erotetic principle in an informal and hopefully intuitive fashion. In sections 2.2.3 and 2.2.4 we show precisely how the two parts of

the erotetic principle are implemented in our proposed formal system.

2.2.2.1 Part I — Premises as questions and maximally strong answers

To illustrate the first part of the erotetic principle, consider so-called *illusory inferences from disjunction* (Johnson-Laird and Savary, 1999; Walsh and Johnson-Laird, 2004), exemplified in (3). These inferences were accepted by around 80% of subjects in a study by Walsh and Johnson-Laird (2004).[8]

(3) P_1: Either Jane is kneeling by the fire and she is looking at the window or otherwise Mark is standing at the window and he is peering into the garden.
P_2: Jane is kneeling by the fire.
C: Jane is looking at the window.

However, (3) is a fallacy. Suppose Jane is kneeling by the fire but *not* looking at the window, while Mark is standing at the window and peering into the garden. This situation makes both premises true while falsifying the conclusion.

How is (3) accounted for with the erotetic principle? First, we observe that, according to some theories of linguistic content, disjunctive sentences such as P_1 of (3) are interpreted in a way very similar to the way questions are interpreted. We present some of the independent motivation for this move in section 2.2.3.3, for now, the following heuristic will suffice.

(4) *Inquisitive postulate, freely adapted from Groenendijk (2008) and Mascarenhas (2009)*:
The interpretation of a sentence of the shape φ *or* ψ contains the question *whether* φ *or* ψ.

[8] The original sentences used by Walsh and Johnson-Laird (2004), as seen in (3), were very long and syntactically awkward. But there are good reasons to believe that Walsh and Johnson-Laird were on to a perfectly ecologically valid phenomenon. Notice that the problem can be recast with universal quantifiers doing the job of conjunction, preserving the attractive character of the fallacious inference while easing the processing load (see Chapter 3, section 3.3.3):

(i) P_1: Every boy or every girl will come to the party.
P_2: John will come to the party.
C: Bill will come to the party.

Just like its propositional counterpart (3), (i) above is fallacious, as it might well have happened that every girl came to the party while John was the only boy that did. However, the reader is likely to agree that (i) is a very attractive inference pattern.

By the erotetic principle, and in a way consonant with the inquisitive postulate in (4), the first premise of (3) is interpreted as a question. Informally: "are we in a Jane-kneeling-by-the-fire-and-looking-at-the-window situation, or in a Mark-standing-at-the-window-and-peering-into-the-garden situation?" Consequently, the reasoner that has just processed P_1 is now attempting to answer a question.

The erotetic principle takes it that having an unanswered question in attention is an uncomfortable state of affairs. First, questions induce conversational and social pressure to find answers to them. Second, questions force reasoners to keep track of two or more distinct possibilities, all of which are possible candidates for states of affairs of the actual world. On the erotetic theory of reasoning, reasoners attending to questions will try as hard as possible to dissolve the question by finding a sufficiently adequate answer to it.

Now, comes premise P_2. This premise is purely declarative, for notice that it does not contain a disjunction. Following the erotetic principle, the reasoner attempts to interpret it as an answer to the question she is attending to. She then observes that P_2 is related to the first possible answer to the question in attention, but not to the second possible answer. Together with the desire to resolve questions, this fact prompts the reasoner to overestimate the adequateness of the potential answer P_2, considering it to be a complete answer to the question. As a result, the reasoner has now discarded the second possible answer to P_1 (involving Mark), and considers it now established that the first answer was the true answer: Jane is kneeling by the fire and she is looking at the window. From here, the fallacious conclusion follows by a simple step of conjunction elimination.

2.2.2.2 Part II — Certain questions make us (classically) rational

In face of the fallacies that we are subject to in naive reasoning, it is important not to lose sight of the fact that the inferential capacities of our species are quite remarkable. Factors that are irrelevant to what we are entitled to conclude from the information we have make a difference to what conclusions are naively endorsed, but our reasoning abilities are not irretrievably lost to these factors. When pressed, even naive reasoners appear to be sensitive in principle to considerations of whether our conclusions would be guaranteed by the information we have. Fallacious inferences are not

always robust in the face of questioning.

Modern science and philosophy are possible and from this we can conclude that correct reasoning can be learned. Now, philosophers and scientists may rely on formalisms or invented reasoning strategies in order to go about their work. It is unlikely that all of those strategies rely on exactly the same principles and mechanisms found in naive reasoning. This may encourage some to think that it suffices to have a model of naive reasoning that is essentially fallacious and that captures the sort of data that we have discussed so far but that has nothing to say about the possibility of correct reasoning. However, this attitude is problematic. Our natural reasoning capacities should not leave us irretrievably in the dark about correct inference. If there is no way to use our natural capacities in a way that provides for correct reasoning, it is a puzzle how any reasoning technologies could be invented by us in order to ensure correctness.

From the perspective of modern science and philosophy, correct reasoning for the propositional case means classically correct reasoning. This emphatically does not mean that premises expressed in natural language would have to be interpreted in the way suggested by an introductory logic textbook. What this means is that given that we have settled on a certain representational content, however we arrived at it, by means of language or otherwise, classical logic tells us what must be true of the world given that this representational content is true of the world. Any scientific publication, even if it happens to be a philosophy paper describing a nonclassical logic, relies on a classical notion of what is entailed by the content we accept.

According to the erotetic principle, the functional aim of naive reasoning is not in the first instance to produce classically correct reasoning. However, it is possible to use the resources of naive reasoning in a way that is guaranteed to produce classically correct reasoning. The key is that we have to systematically raise the right questions as we reason.

Return to the fallacious inference in (3) above. How would systematically raising the right questions block this fallacious inference? What gets us into trouble is that we are asking, "Are we in a Jane-kneeling-by-fire-and-looking-at-window situation or are we in a Mark-standing-by-window-and-peering-into-garden situation?" effectively dismissing that there are other alternatives compatible with our first premise. When we then encounter the premise "Jane is kneeling by the

fire" and treat it as a maximally strong answer to our question, we are left with the fallacious conclusion that Jane is looking at the window. Now, what can we do to realize that this inference cannot be made from the information provided by the premises? Quite simply, before we try to answer the question raised by our first premise, we raise further questions about the propositional atoms mentioned in the premise. For example, we ask "Is Jane kneeling by the fire?" As we formally envisage in our system the effect of taking this question on board, it will force us to consider a case in which Jane is kneeling by the fire, Mark is standing by the window and peering into the garden, but Jane is not looking at the window. By raising these further questions, fallacious inferences can be blocked. As we prove formally, this holds in the general case. Assuming that we inquire on all propositional atoms mentioned in our premises right before updating a discourse with those premises, the erotetic theory of propositional reasoning is classically sound.

2.2.3 A bird's-eye view of the theory

In this section, we informally introduce each component of the formal system of the erotetic theory of reasoning, together with its motivation within the system and in light of the desiderata given by the erotetic principle.

2.2.3.1 Components of the formal system

A theory of mental representations We need an explicit theory of what mental representations (mental models) look like. Here, there are two desiderata that must be satisfied.

First, in order to implement the erotetic principle, mental models must represent disjunctions (and ultimately other natural language constructions, such as indefinites, see for example Mascarenhas, 2011) as question-like meanings. Happily, the linguistics literature offers an account of linguistic content that does precisely what we need. We will import the basic insight of inquisitive semantics (Groenendijk, 2008; Mascarenhas, 2009), where the interpretations of sentences are taken to both convey information and raise issues, as well as of *exact verification semantics* (Fine, 2012), a non-classical semantic framework that shares the requisite properties with inquisitive semantics. This will be discussed in detail in section 2.2.3.3.

Second, we need mental models to be fine-grained enough to distinguish between different contradictory representations. That is, we need mental models to distinguish the representation of $p \wedge \neg p$ from that of $q \wedge \neg q$. This will allow us to capture and explain the fact that reasoners do not find disjunctive syllogism to be a fully straightforward inference (Evans et al., 1993), while they are nonetheless capable of drawing other conclusions from the same premises. Disjunctive syllogism is schematized in (5).

(5) P_1: $p \vee q$

P_2: $\neg p$

C: q

How do we block (5)? When the information in P_2 is added to P_1, we distribute P_2 into the disjunction, getting a mental model representation of $(p \wedge \neg p) \vee (q \wedge \neg p)$. Intuitively, we will want anything that follows from both disjuncts to follow from the disjunction. Assume first that *ex falso* is blocked, for with *ex falso*, q would follow immediately. In the next section we will explain how *ex falso* is in fact hard to get. If the contradiction in the first disjunct $p \wedge \neg p$ is prevented from deriving q, the inference is blocked. Why not simply block *ex falso*, without committing (as we will do) to a view of content that distinguishes between contradictions? The reason is that, while we do not want $(p \wedge \neg p) \vee (q \wedge \neg p)$ to derive (at least not immediately) q, one should be able to derive *some things* from it. In particular, $\neg p$ should be derivable, since it is contained in both disjuncts. As far as we can see, assuming that contradictions are not all alike is the only way to allow for simple inferences out of premises containing contradictions, while being consistent with making *ex falso* a difficult inference pattern to accept. A further empirical reason to distinguish different contradictions representationally is that naive reasoners most likely do not realize automatically when they are entertaining contradictions (Morris and Hasson, 2010).

Mental model discourses The erotetic theory is dynamic: we take it that premises are interpreted in the order that they were given, and that in principle that order can make a difference.[9] This

[9] As an anonymous reviewer points out, it is important to remark that, while our system is dynamic in that the order in which premises are updated into discourses matters, the mental model interpretations *themselves* will in fact be static

much has been established in the reasoning literature. For example, the disjunctive syllogism in (5) becomes significantly more acceptable to naive reasoners if the order of the premises is switched (García-Madruga et al., 2001). Like any other dynamic theory, we will need some notion of a state (or context) to update with the premises. For us, this role will be played by what we call mental model discourses.

Updating via the erotetic principle The next ingredient is an update rule that implements Part I of the erotetic principle, treating certain premises as questions and others as maximally strong answers to questions in context whenever possible. Besides treating information as questions and answers our update rule also has to allow for cases in which we simply accumulate information, as when we are given successive categorical statements.

Simple deduction rule Reasoning is not just a matter of update. Once reasoners hear and process each premise, they must then be able to perform simple transformations on the resulting mental model, to check what follows. We assume that there is a rule of disjunct simplification, validating the inference $(p \wedge q) \vee r \vDash p \vee r$. This rule for disjunct simplification includes conjunction elimination as a special case, as the reader can see.

Eliminating contradictions We take it that reasoners do not immediately see anything wrong with contradictions. However, there must be a process allowing them to look at the representations they are entertaining and check whether they are consistent or not. This comes at a cost and is not part of default reasoning (to be defined shortly), but it must be a possibility if we want to account for the successes of our reasoning faculty. We will therefore define an operation that filters the mental model in discourse, going over each alternative and eliminating all those that are contradictory.

Expanding possibilities The mental models of the erotetic theory represent only what is minimally required to model a statement, and are therefore typically underspecified. We need an operation that expands the mental model under consideration through successive applications into one

meanings.

that represents every possibility with respect to some propositional atom. As discussed in section 2.2.2.2, this will be a crucial ingredient of the strategy allowing for classically sound reasoning. Accordingly, it implements Part II of the erotetic principle.

Default reasoning strategy Finally, we need to make a simple postulate describing how reasoning problems are approached by default. We propose the following strategy. When given a reasoning problem with premises $P_0, \ldots, P_n$ and conclusion C, reasoners update a blank mental model discourse with each premise, in the order the premises were given. They may then apply the simple deductive rule, targeting the conclusion C. If the resulting mental model in discourse is identical to C, then the inference is deemed valid. Otherwise, it is deemed invalid.

2.2.3.2 The sources of reasoning failures

The erotetic theory of reasoning pinpoints several sources of divergences from classically correct reasoning. To get a full grasp of how these sources conspire to produce these divergences, it will be necessary to work through the formal details we present in sections 2.2.4 and 2.2.5. However, it is worth considering a brief overview of the types of divergences from classical reasoning predicted by the erotetic theory and of some of the examples we will later work through.

Limits on numbers of alternatives Like other theorists, we subscribe to the view that, as more alternatives need to be represented simultaneously for a reasoning problem, it becomes less likely that reasoners arrive at an answer (fallacious or otherwise), where five to seven alternatives is the limit (Johnson-Laird, 1983). We will see that because of this constraint, strategic suppositions made in reasoning that reduce the need to represent multiple alternatives at a time can make extremely difficult-seeming problems tractable (ex. 20).

Failure to apply a creative reasoning strategy All other things being equal, given two similar inferences from similar premises, if one inference requires more creative applications of optional update rules beyond default updating, reasoners will be less likely to make the inference (fallacious or otherwise). For every optional step, there is an additional chance that reasoners fail to make it.

For example, the erotetic theory predicts that modus tollens is harder than modus ponens (Evans et al., 1993) for this reason, as we will see in ex. 22, since the former requires an optional operation that filters out contradictory alternatives.[10]

Default but fallacious reasoning via update Default reasoning via updating a mental model with successive premises is enough to yield fallacious inferences for various premise pairs. For example, this yields the category of fallacious inferences referred to as "illusory inferences" in Johnson-Laird and Savary (1999) and Walsh and Johnson-Laird (2004), as seen in ex. 10 and ex. 29. Default reasoning can also make contradictory statements seem consistent (Johnson-Laird et al., 2004), as in ex. 34.

Order effects brought about by Part I of the erotetic principle The update procedure that treats a new premise as an answer to a previous premise immediately eliminates alternatives in the new premise that conflict with what has been established as background in the discourse. As a result, certain inferences are easier if a categorical premise comes first. This captures that there are order effects for modus tollens but not for modus ponens (Girotto et al., 1997), as seen in ex. 33.

Need to creatively raise questions Certain inferences are valid but do not seem intuitive. For example, "explosion" inferences (e.g. "It is raining and it is not raining, therefore I am the Pope.") and disjunction introduction inferences (e.g. "It is hot outside, therefore it is hot outside or I am the President.") are highly counterintuitive (Harman, 1986; Braine et al., 1984). The erotetic theory predicts that these inferences should seem counterintuitive, because reasoning to those inferences would require gratuitously raising a question. The aim of naive reasoning is to answer questions as quickly as possible, but explosion and disjunction introduction would require one to diverge from the this aim, as can be seen in ex. 21, and ex. 19.

[10]Comparing the relative difficulty of reasoning problems along the dimensions of number and type of optional update rules required in solving them most likely only makes sense for similar inferences from similar premises. We suspect that different problems prime various possible moves one could make in reasoning to various degrees, so there may not be a useful *absolute* measure of how likely it is that a reasoner will fail to make a certain optional type of step in an arbitrary reasoning problem, that is, a measure that would apply across all types of reasoning problems.

Naive reasoning isn't all bad Beyond the special strategies that ensure correct reasoning that flow from Part II of the erotetic principle, the erotetic theory of reasoning also captures the fact that many classes of valid inferences are grasped even through naive reasoning. For example, modus ponens is predicted to be extremely easy (ex. 22) and so is conditional transitivity (e.g. "If *P* then *Q* and if *Q* then *R*, therefore if *P* then *R*."), as seen in ex. 36.

Overall, the erotetic theory captures all data listed in Table 1, and various others. We later present a novel illusory inference not yet reported in the literature. To the best of our understanding, no current alternative theory systematically captures all cases in Table 1 *including* this novel case to discussed below.

2.2.3.3 Interpreting premises as sets of exact verifiers

Despite being influenced by mental model theory, our approach differs from it in important respects, especially concerning interpretation. In many of these respects, we are closer in spirit to alternative approaches that have increasingly gained attention. Firstly, the erotetic theory shares with Bayesian approaches to reasoning (Oaksford and Chater, 2007; Tenenbaum et al., 2006) and with more recent non-monotonic approaches a rejection of the program of reducing human failures of reasoning to failures of reasoning *about classical interpretations*. Secondly, we share with the work of Stenning and van Lambalgen (2008a,b) an interest in the workings of the interpretive processes themselves (what these authors refer to as *reasoning to an interpretation*). As we explain in this section, we assume a non-classical semantics for natural language, the semantics of exact verification (van Fraassen, 1969; Fine, 2012, among others), together with an inquisitive semantics (Groenendijk, 2008; Mascarenhas, 2009) perspective on questions and declaratives.

Exact verifiers We propose interpretations that track the *exact verifiers of statements*.[11] The concept of verification (Fox, 1987; Armstrong, 2004; Fine, 2012) is concerned with assessing what it is in the world that makes a statement true. In the special case of *exact* verification, in which our

[11] More precisely, we use *minimal* exact verifiers. Minimality amounts to the assumption that there is always *one* minimal truth-maker that verifies a sentence. This assumption is warranted for the propositional case (see also footnote 12).

account is couched, the question is what *exactly* in the world makes a statement true: the meaning of a statement is modeled as the set of those things that would exactly make it true. This is quite distinct from meanings in classical logic and classical possible-world semantics, which characterize in which fully specified possible worlds a sentence is true. In classical semantics, we consider objects (possible worlds, truth-table rows) fully specified for every propositional atom mentioned in a sentence and ask whether the sentence is true or false at those points of evaluation. An exact verification semantics takes objects only as large as must be to make a particular sentence true.

Consider a disjunction like 'P or Q'. The classical truth-table analysis of connectives considers three situations where this sentence is true: both P and Q are true, P is true and Q is false, P is false and Q is true. Notice that the second and third situations include conditions irrelevant to making the disjunction true: if P is the case, that is all you need to make the disjunction true. In an exact verification semantics, a situation where P is true and nothing is said about other facts is one that exactly verifies the disjunction 'P or Q'. One where P is true and Q is false, while *compatible* with the truth of the disjunction, does not *exactly* verify it.

The truth-makers of exact verification semantics can be seen as situations, though other constructs (such as sets of possible worlds) achieve the same results in many important cases.[12] The crucial notion however is that of the verification relation and what it says about how sentences with connectives are exactly verified. In this work, we abstract away from model-theoretic con-

[12]This is true notably for the case of exact verification semantics for a propositional language. Because we can assume the existence of minimal verifiers, taking the truth-maker for a sentence P to be the situation that exactly verifies P or simply the set of all possible worlds that make P (classically) true are both viable routes. We note that the quantified case is more complex, opening various possible avenues to extend the present system, which we leave for future work. The difficulties in the quantified case are introduced by the fact that, while first order formulas are finite objects, the models that make them true may well be infinite. For sentences whose exact verifiers are situations with some infinite cardinality, it is easy to see that no one situation will be a *minimal* exact verifier. The literature offers two kinds of solutions to this issue. In the inquisitive semantics literature, Ciardelli (2009) proposes a formal constraint on models that guarantees the existence of minimal exact verifiers for the quantified case — at the cost of some expressive power necessary to model a certain class of mathematical statements. The tenability of Ciardelli's proposal for the purposes of building a theory of reasoning is thus predicated on the importance we ascribe to providing *exactly accurate* mental model representations of mathematical statements. While the philosopher will immediately discard Ciardelli's system, we suspect that the psychologist (and to some extent the linguist) might not be too concerned with assigning mental model representations that are completely faithful models of mathematical statements (thereby including infinitely-large alternatives in their mental models). Fine (2012) proposes a philosophically and mathematically sound solution: dropping the minimality requirement. The cost incurred by this move is one of simplicity of alternatives. The minimality assumption has the welcome advantage of allowing us to point to *one* situation per alternative, rather than dividing each alternative into some larger set of situations. Choosing between these alternatives (or others that there may be) will involve doing the same work for the quantified case as we do here for the propositional case, and therefore we must leave it to future research.

siderations about verification semantics, and focus instead on the interpretation of sentences of (pseudo-)English into mental models.

Naturally, to give interpretations for connectives, we need to say what their exact verifiers are. Our analysis of all connectives except for the conditional is isomorphic to verification-semantic analyses defended by other authors on entirely independent grounds (e.g. Fine, 2012). We define the syntactic space of well-formed formulas within our formalization of mental models in section 2.2.4.1, but limitations of space prevent us from giving an explicit model theory for exact verification semantics.[13]

We do have a set of semantical heuristics that we want to encourage the reader to keep in mind, when thinking of how sentences are interpreted in our account, given in (6) below. Two remarks are in order. First, we omit the conditional connective, as our account of it diverges significantly from exact verification semantics. It will be discussed in detail in section 2.2.5. Second, the reader will notice that in (6) we assume that only atoms are ever conjoined or negated. It will be clear why we make this move in section 2.2.4.1.

(6) a. A propositional atomic statement like *P* is exactly verified by the situation where *P* is true is nothing else is said about any other propositions.

b. A negative statement like *not P* is exactly verified by the situation where *P* is false and nothing else is said about any other propositions.

c. A conjunctive statement like *P and Q* is exactly verified by the situation where *P* is true and *Q* is true, and nothing else is said about any other propositions.

d. A disjunctive statement like φ *or* ψ is exactly verified by the *set of situations* that exactly verify φ or exactly verify ψ.

Linguistic motivations — exact verification and inquisitive semantics Exact verification semantics has recently been proposed by Fine (2012) as a way to address issues in the semantics of counterfactuals and the semantics of permission. For example, the permissions given by sentences

[13] We refer the reader to Fine (2012) and to the very close inquisitive semantics system of Groenendijk and Roelofsen (2010), two papers that present different but almost exactly equivalent model theories.

(7a) and (7b) below are quite distinct. Concretely, while (7b) gives you permission to take a biscuit and some ice-cream, (7a) does not.

(7) a. You may have a biscuit.

b. You may have a biscuit or a biscuit and some ice-cream.

However, in classical logic absorption is valid: $(\varphi \vee (\varphi \wedge \psi)) \leftrightarrow \varphi$. Under natural assumptions about the syntax of (7a) and (7b), it follows that in classical logic there is no way to distinguish the complements of the permission modal *may* in the two sentences. But (7a) and (7b) do not give equivalent permissions.

Exact verification provides a solution: rather than taking the meanings of the complements of *might* to be classical meanings, let them be sets of exact verifiers. In general, the exact verifiers for φ are distinct from the exact verifiers for $\varphi \vee (\varphi \wedge \psi)$. The exact verifiers for φ are all situations that exactly verify φ. The exact verifiers for $\varphi \vee (\varphi \wedge \psi)$ are all situations that verify φ as well as all situations that verify $\varphi \wedge \psi$. Clearly, the latter set contains situations that are absent from the former set. Thus, in exact verification semantics, absorption is not valid, and the complements of *might* in (7a) and (7b) are distinguishable, as desired. This, of course, does not immediately solve all issues pertaining to permission modals. The crucial claim is that there are strong reasons to suspect that we need a notion of semantic content allowing for much more fine-grained distinctions than classical semantics gives us. A promising hypothesis is that that notion of content is exact verification semantics.

Exact verification semantics shares interesting properties with independently proposed refinements of linguistic content from the field of linguistic semantics. The exact verification logic given by Fine (2012), in the propositional case, is isomorphic to the propositional inquisitive semantics given by Groenendijk and Roelofsen (2010). Inquisitive semantics, first proposed by Groenendijk (2008) and Mascarenhas (2009) and developed by Groenendijk and Roelofsen (2009) and later work with collaborators, argues that some syntactically declarative sentences of natural languages are semantically very much like questions. The staple example is disjunction.

(8) John or Mary will come to the party.

In inquisitive semantics, (8) provides both information (that it can't be the case that neither John nor Mary show up) and an *issue* (which one of John or Mary will come to the party). Our account of reasoning imports this insight, since it holds that sentences with disjunction pose questions that reasoners do their best to address as soon as possible.

Moreover, both exact verification semantics and inquisitive semantics are related to the alternative semantics (also known as Hamblin semantics) of Kratzer and Shimoyama (2002, and a wealth of later work). Concretely, these three independently proposed frameworks all agree that the semantics of *or* cannot be the traditional Boolean join. Unfortunately, we cannot do justice in this chapter to the linguistic and philosophical appeal of these non-classical approaches to linguistic content. However, once one grants their tenability as accounts of linguistic content, the proposals we make in this chapter gain welcome independent motivation.

2.2.4 The core fragment

In this section we give a rigorous implementation of the core fragment of the erotetic theory informally presented above. This core fragment contains all of the ingredients discussed in the preceding section except for the mechanism for supposition and an account of conditional premises, which we address in section 2.2.5.

2.2.4.1 Defining mental models

Mental models are structured mental representations that can be generated through the workings of perception, thought, and memory (Johnson-Laird, 1983). Mental models are also used to account for reasoning with visually presented information (Bauer and Johnson-Laird, 1993), not just for reasoning with verbal premises. Mental models have as much structure as there are distinctions we take into account in representing something (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991). In this section, we make explicit the basic formal ingredients that, to the best of our understanding,

must be assumed by any mental model theory of propositional reasoning.[14] We offer formal definitions of these basic ingredients, upon which we will build the rest of our proposal in the sections to come. First, we need to define mental model components that stand for propositional atoms and their negations.

Definition 1 (Mental model nuclei). A set $\mathscr{N}$ of mental model nuclei is a non-empty set containing the smallest representational units of the system for propositional reasoning, standing for atomic propositions:

$$\mathscr{N} = \{p, q, \ldots\}$$

Next, for propositional reasoning, we will need negations of mental model nuclei.

Definition 2 (Mental model nuclei closed under negation). Given a set $\mathscr{N}$ of mental model nuclei as per Definition 1, we define the set $\mathscr{N}^+$ as the closure of $\mathscr{N}$ under $\neg$.

$$\mathscr{N}^+ = \{p, \neg p, \neg\neg p, \neg\neg\neg p, \ldots, q, \neg q, \neg\neg q, \neg\neg\neg q, \ldots\}$$

The next step is to define combinations of mental model nuclei that inherit all of the representational commitments of the nuclei that compose them. Intuitively and informally, these can be thought of as "conjunctions" of mental model nuclei. We dub them *mental model molecules*, and form them from mental model nuclei with the operation '$\sqcup$'.

Definition 3 (Mental model molecules). Every mental model nucleus in $\mathscr{N}^+$ is also a mental model molecule. If α and β are mental model molecules, then $\alpha \sqcup \beta$ is a mental model molecule.

We can then think of a mental model as a set of mental model molecules of this sort. Following the discussion in section 2.2.3.3, we say that the *alternatives* in a mental model are all of the mental

[14] We adopt a slight departure in terminology from standard discussions of mental models. The mental model interpretation of a propositional atom or a conjunction, has been called a "mental model," while the interpretation of a disjunction, with two or more alternatives has been called "a set of mental models" (Johnson-Laird, 2008). We suggest that it is easier to provide uniform definitions if we think of all premises in propositional reasoning as supplying "mental models" to the system and updating "mental models." Rather than speaking of mental models and sets of mental models, we will speak of mental models with only one element standing for one state of affairs and mental models with multiple elements standing for multiple alternative states of affairs.

model molecules contained in it. If there is more than one alternative in the set, the mental model is "inquisitive", representing a question. If there is only one combination in the set, the mental model represents a categorical statement. We give explicit definitions shortly.

In order to define operations for reasoning with mental models, we need to be able to talk about what nuclei are shared between mental model molecules that make up different representations of alternatives, and we need to be able to ask whether a molecule is part of a representation of an alternative.

We will write '$\alpha \sqcap \beta$' to stand for the mental model molecule that two molecules α and β have in common. For example:

Example 1. (i) $(\mathsf{p} \sqcup \mathsf{q}) \sqcap (\mathsf{p} \sqcup \mathsf{r}) = \mathsf{p}$ (ii) $(\mathsf{p} \sqcup \mathsf{q} \sqcup \mathsf{r}) \sqcap (\mathsf{p} \sqcup \mathsf{q} \sqcup \mathsf{s}) = \mathsf{p} \sqcup \mathsf{q}$

We can now also formalize the situation where a mental model molecule is included in another. We use the symbol '$\sqsubseteq$' for the relation "included in or possibly equal to."

Example 2. (i) $\mathsf{p} \sqsubseteq \mathsf{p} \sqcup \mathsf{q}$ (ii) $\mathsf{p} \sqcup \mathsf{q} \sqsubseteq \mathsf{p} \sqcup \mathsf{q}$

To sum up, we propose the following algebraic structure.

Definition 4 (Molecular structures). A structure $\mathfrak{M} = \langle \mathscr{M}, \sqcup, \sqcap, 0 \rangle$ is a mental model molecular structure iff all of the following conditions hold: (1) $\mathscr{N}^+ \subseteq \mathscr{M}$. (2) $\langle \mathscr{M}, \sqcup, \sqcap \rangle$ is a lattice, that is, the operations $\sqcup$ and $\sqcap$ obey the laws of commutativity, associativity, and absorption. (3) For any α in $\mathscr{M}$, $\alpha \sqcup 0 = \alpha$, that is, 0 is the identity element for the join operation. The null nucleus 0 makes no representational commitments at all and can be thought of as corresponding to *truth* in classical logic. This structure gives rise to a partial order $\langle \mathscr{M}, \sqsubseteq \rangle$, where $\sqsubseteq$ is defined in the usual way $\alpha \sqsubseteq \beta$ iff $\alpha \sqcup \beta = \beta$ or $\alpha \sqcap \beta = \alpha$.

Notice that this structure implements the idea that mental models can distinguish between different contradictory states of affairs. In $\mathfrak{M}$ as defined above, $p \sqcup \neg p$ and $q \sqcup \neg q$ are distinct objects.

We do not require that all elements of $\mathscr{M}$ be mental model nuclei standing for atomic propositions; the set of those nuclei is merely included in $\mathscr{M}$. Thus, the system can in principle be

expanded beyond propositional reasoning while keeping these basic structural properties. Because of this openness of the system, it will be convenient to refer to the set of *atoms* of the molecular structure. Intuitively, this is the set of all those elements of the molecular structure that are not the $\sqcup$-combinations of simpler elements. Notice that negative mental model nuclei count as atoms: they are not gotten by applying the operation $\sqcup$ to simpler elements.

Definition 5 (Atoms). Given a mental model molecular structure $\mathfrak{M} = \langle \mathscr{M}, \sqcup, \sqcap, 0 \rangle$,

$$Atoms(\mathfrak{M}) = \{\alpha \in \mathscr{M} : (\neg\exists \beta \in \mathscr{M})\ \beta \neq \alpha\ \&\ \beta \neq 0\ \&\ \beta \sqsubseteq \alpha\}$$

We can now define mental models in terms of this general notion of a mental model molecular structure.

Definition 6 (Mental models). Given a mental model molecular structure $\langle \mathscr{M}, \sqcup, \sqcap, 0 \rangle$, the set $\mathbb{M}$ of mental models is the smallest set containing every subset of $\mathscr{M}$ and the absurd mental model, notated $\emptyset$, corresponding to the contradiction (*falsum*) in classical logic. Further, we will call all $\Gamma \in \mathbb{M}$ such that $|\Gamma| \leq 1$ *categorical* mental models, and all other mental models *inquisitive* mental models.

2.2.4.2 Mental model discourses

Background knowledge can influence reasoning. As far as we can see, the simplest way to account for this in our system is to say that background knowledge itself is ultimately represented in the form of mental models. In our system, background knowledge will consist of a set of mental models with two properties: (1) the reasoner considers the facts represented in this set to be sufficiently well established, and (2) the mental models in this set are easily accessible to the reasoner, meaning she is especially aware of their content. This means that background knowledge will represent both relevant facts part of a reasoner's knowledge prior to the current reasoning task *and* especially salient facts established within the current reasoning task. Background knowledge has its natural locus within mental model discourses, the workspaces of reasoning that will be successively updated with information from premises.

In addition, we should also keep track of what mental models are taken to be about. Presumably, the same mental models could be used to represent the universe according to Brian's beliefs, according to Brian's beliefs plus a supposition, or according to Mary's dream. Drawing conclusions not only involves transforming mental models but knowing what we take them to stand for. We model this by adding an index parameter to mental model discourses, containing information about what the mental model discourse represents. This index will also be responsible for flagging mental model discourses that carry uncanceled suppositions, analogous to assumptions to be canceled by implication introduction in natural deduction systems. The index will in fact be idle for most of the formal system except for discourses with suppositions so, in order to reduce clutter in our formulas, we omit the index from the discussion in section 2.2.4 and reintroduce it in section 2.2.5, where we discuss the supposition mechanism and conditional sentences. Formally, this means that the definitions of operations on mental model discourses that follow are in fact abbreviations for operations that take as input and output mental model discourses with indexes in the rightmost position. Every operation that we abbreviate in this way should be seen as standing for a definition just like it, except that the index parameter from the input discourse occurs as the index in the output discourse as well.

Definition 7 (Mental model discourses). A mental model discourse is strictly a triple $\mathfrak{D} = \langle \Gamma, B, i \rangle$, where Γ is a mental model, B (for background) is a set of established mental models, and i is an index flagging what the discourse is about. For the rest of this section, we omit the idle index and abbreviate all mental model discourses as pairs $\langle \Gamma, B \rangle$.

2.2.4.3 Mental model conjunction

Mental model conjunction combines the information and the questions in two mental models. Given two models Γ and Δ, mental model conjunction looks at each pair that can be formed from one element of Γ together with one element from Δ, combines the two elements via '$\sqcup$', and collects all of the results. This procedure returns the empty set if either one of Γ or Δ is empty.

Definition 8 (Mental model conjunction). For Γ and Δ mental models:

$$\Gamma \times \Delta = \begin{cases} \{\gamma \sqcup \delta : \gamma \in \Gamma \ \& \ \delta \in \Delta\} & \text{if } \Gamma \neq \emptyset \text{ and } \Delta \neq \emptyset \\ \emptyset & \text{otherwise} \end{cases}$$

Example 3. (i) $\{\mathsf{a},\mathsf{b}\} \times \{\mathsf{c},\mathsf{d}\} = \{\mathsf{a}\sqcup\mathsf{c}, \mathsf{a}\sqcup\mathsf{d}, \mathsf{b}\sqcup\mathsf{c}, \mathsf{b}\sqcup\mathsf{d}\}$ (ii) $\{\mathsf{a}\} \times \{\mathsf{b}\} = \{\mathsf{a}\sqcup\mathsf{b}\}$.

Mental model conjunction is the most elementary way in which premises can be combined, but it does nothing to implement the erotetic principle. This will be the job of mental model *update*.

2.2.4.4 Update for the erotetic theory of reasoning

The update procedure in the erotetic theory has two components. One implements the erotetic principle, attempting to interpret the new premise as an answer to the question in discourse — we call it Q-update. The second component, C-update, embeds whatever new information and questions are provided by the premise under consideration into the discourse, checking to see if there are additions to be made to the set of background information.

Definition 9 (Q-update). The Q-update of a mental model discourse $\langle \Gamma, B \rangle$ with a mental model Δ is defined as follows.[15]

$$\langle \Gamma, B \rangle [\Delta]^Q = \langle \Gamma - \{\gamma \in \Gamma : \left(\bigsqcap \Delta\right) \sqcap \gamma = 0\}, B \rangle$$

Q-update leaves only those alternatives in the question posed by Γ that include a mental model molecule that is shared by all alternatives in Δ. In other words, Q-update leaves only those alternatives in the question posed by Γ that involve some conjunct c such that c could be obtained from each alternative in Δ by the equivalent of a conjunction-reduction inference. This is how we implement the idea that Q-updating means taking the new information Δ as the strongest possible answer to the question at hand. Here, taking the new information as the "strongest possible" answer means that we do not require *all* conjuncts in the remaining alternatives to match something explicitly

[15] It is understood that $\bigsqcap\{\delta_1, \ldots, \delta_n\} = \delta_1 \sqcap \cdots \sqcap \delta_n$.

represented in the answer. In the best case for this way of answering our question, only one of the alternatives in our question shares a conjunct with all alternatives in a putative answer. In this case, we take it that our question has been narrowed down to that one alternative, and hence answered. If a putative answer has nothing in common with our question but we nevertheless treat it as an answer, we end up with no remaining alternatives, and Q-update returns the empty set.[16]

Example 4. $\langle\{a \sqcup b, c \sqcup d, e \sqcup f\}, B\rangle[\{a\}]^Q = \langle\{a \sqcup b\}, B\rangle$

Example 5. $\langle\{a \sqcup b, c \sqcup d\}, B\rangle[\{a \sqcup e, a \sqcup f\}]^Q = \langle\{a \sqcup b\}, B\rangle$

Example 6. $\langle\{a \sqcup b, c \sqcup d\}, B\rangle[\{e \sqcup f, g\}]^Q = \langle\emptyset, B\rangle$ The new premise cannot be interpreted as an answer to the question in discourse, since it is completely orthogonal to the discourse. Q-update returns the absurd mental model.

C-update does two things. First, it builds a new mental model by mental-model-conjoining Γ with that subset of Δ whose elements do not contradict any element of the established background B. This implements the idea that reasoners are especially mindful, and find it easy, not to contradict information in the background. The reader should bear in mind that this ease in identifying contradictions is a characteristic of the established background: it does not extend to the mental model in discourse, that is Γ. This will be important to derive some of the effects of different orderings of premises. Second, C-update updates the background B by adding to it individual mental models for any mental model atom that Δ establishes.

Definition 10 (C-update). Let a mental model discourse $\mathfrak{D} = \langle\Gamma, B, i\rangle$ and a mental model Δ be given. The C-update of $\mathfrak{D}$ with Δ is defined as follows.

$$\langle\Gamma, B\rangle[\Delta]^C = \langle\Gamma', B^*(\Gamma', \Delta)\rangle$$

[16]The attentive reader may wonder why we did not choose a nearby alternative definition of Q-update that would leave all of those γ in Γ that have something in common with *at least one* alternative in Δ (this would simply amount to replacing $\sqcap$ with $\sqcup$ in the formal definition of Q-update). This nearby version of Q-update would yield a version of Update (see Definition 11 below) that ultimately makes the reasoner even more credulous in treating material in successive premises as answers to questions posed by previous premises. As it happens, together without our account of conditionals in section 2.2.5.2 this would predict that from 'if P then Q' and 'if Q then R', we would be strongly tempted to fallaciously infer 'P and Q and R'. We take it that there is no temptation of this sort.

Where $\Gamma' = \Gamma \times \{\delta \in \Delta : (\neg\exists\beta \in B)\ (\forall b \in \{\delta\} \times \beta)\ \textsc{Contr}(b)\}$. The function $B^*(\alpha,\beta)$ is defined as follows (recall that $\mathfrak{M}$ is the molecular structure underlying $\mathfrak{D}$, as per Definition 4).[17]

$$B^*(\alpha,\beta) = \begin{cases} B \cup \{\{p\} : p \in \mathit{Atoms}(\mathfrak{M})\ \&\ (\exists a \in \alpha)\ p \sqsubseteq a\} & \text{if } \alpha \text{ is categorical} \\ B \cup \{\{p\} : p \in \mathit{Atoms}(\mathfrak{M})\ \&\ (\exists b \in \beta)\ p \sqsubseteq b\} & \text{if } \alpha \text{ is inquisitive, } \beta \text{ categorical} \\ B & \text{otherwise} \end{cases}$$

CONTR is a function from mental model molecules into truth values, returning **true** whenever its argument is a contradiction and **false** otherwise. Formally:

$$\textsc{Contr}(\alpha) = \begin{cases} \textbf{true} & \text{if } (\exists a \sqsubseteq \alpha)\ \neg a \sqsubseteq \alpha \\ \textbf{false} & \text{otherwise} \end{cases}$$

We make further use of the functions CONTR and $B^*(\alpha,\beta)$ in the definitions that follow.

Example 7. Taking a *tabula rasa* mental model discourse, $\langle\{0\},\emptyset\rangle$, with no background information and a non-committal mental model as a starting point for a C-update:

$$\langle\{0\},\emptyset\rangle[\{\mathsf{a}\sqcup\mathsf{b}\}]^{\mathsf{C}} = \langle\{\mathsf{a}\sqcup\mathsf{b}\},\{\{\mathsf{a}\},\{\mathsf{b}\}\}\rangle$$

Notice how the discourse is updated with $\mathsf{a}\sqcup\mathsf{b}$, while the background set is updated with two categorical mental models $\{\mathsf{a}\}$ and $\{\mathsf{b}\}$. These two mental models now count as established facts, and from this point onward reasoners will be especially good at detecting when new premises contradict them.

Example 8. Suppose we have a background $\mathsf{B} = \{\{\neg\mathsf{a}\}\}$ establishing that what is represented by a is not the case:

$$\langle\{\mathsf{d}\sqcup\mathsf{e},\mathsf{g}\},\{\{\neg\mathsf{a}\}\}\rangle[\{\mathsf{a}\sqcup\mathsf{b},\mathsf{f}\}]^{\mathsf{C}} = \langle\{\mathsf{d}\sqcup\mathsf{e},\mathsf{g}\}\times\{\mathsf{f}\},\{\{\neg\mathsf{a}\}\}\rangle = \langle\{\mathsf{d}\sqcup\mathsf{e}\sqcup\mathsf{f},\mathsf{g}\sqcup\mathsf{f}\},\{\{\neg\mathsf{a}\}\}\rangle$$

[17] This definition of $*$, the background-updating function, adds to the background not only facts established by the new mental model Δ, but also any categorical facts in the *result* of updating the discourse with the new model Δ. The intuition is that categorical facts receive special attention, not only when they are heard but also when they are derived.

Notice that the first alternative of the new premise, $\{\mathsf{a} \sqcup \mathsf{b}\}$, was discarded. It contradicted a model previously established in the background, namely $\{\neg \mathsf{a}\}$.

Example 9. $\langle \{\mathsf{a} \sqcup \mathsf{b}, \mathsf{c}\}, \emptyset \rangle [\{\mathsf{d}\}]^{\mathsf{C}} = \langle \{\mathsf{a} \sqcup \mathsf{b}, \mathsf{c}\} \times \{\mathsf{d}\}, \{\{\mathsf{d}\}\} \rangle = \langle \{\mathsf{a} \sqcup \mathsf{b} \sqcup \mathsf{d}, \mathsf{c} \sqcup \mathsf{d}\}, \{\{\mathsf{d}\}\} \rangle$

We can now define the general update procedure in terms of Q-update and C-update. The procedure begins with a Q-update. If Q-update returns a discourse with a non-empty mental model, that means the new premise was successfully interpreted as an answer to the question in discourse. The update procedure then performs a C-update with the same new premise, for the new premise might also provide new information beyond answering the question. On the other hand, if the initial Q-update returns a discourse with an empty mental model, this means that the new premise cannot be interpreted as an answer to the question. In this case, the update procedure just performs a C-update with the new premise, adding whatever new information it provides to the discourse.

Definition 11 (Mental model discourse update). The result of updating a mental model discourse $\langle \Gamma, B \rangle$ with a mental model Δ is defined as follows.

$$\langle \Gamma, B \rangle [\Delta]^{\mathrm{Up}} = \begin{cases} \langle \Gamma, B \rangle [\Delta]^{C} & \text{if } \langle \Gamma, B \rangle [\Delta]^{Q} = \langle \emptyset, B' \rangle \\ \langle \Gamma, B \rangle [\Delta]^{Q} [\Delta]^{C} & \text{otherwise} \end{cases}$$

With mental model discourse update, we define the default procedure for reasoning from successive premises.

Definition 12 (Default reasoning by update). By default, reasoners take it that what holds in the mental model that results from successively updating their mental model discourse with given premises can be inferred from those premises.

To put the notion of default reasoning to use, we need to define a basic fragment of interpretation.

Definition 13 (Basic fragment of interpretation into mental models). We define a function $\|\cdot\|^{\mathfrak{D}}$

from sentences S of a language and a mental model discourse $\mathfrak{D}$ into mental models as follows.

$$\|p\|^{\mathfrak{D}} = \{p\}, \text{ for atoms } p$$

$$\|\varphi \textit{ or } \psi\|^{\mathfrak{D}} = \|\varphi\|^{\mathfrak{D}} \cup \|\psi\|^{\mathfrak{D}}$$

$$\|\varphi \textit{ and } \psi\|^{\mathfrak{D}} = \|\varphi\|^{\mathfrak{D}} \times \|\psi\|^{\mathfrak{D}}$$

$$\|\textit{not } \varphi\|^{\mathfrak{D}} = \textsc{Neg}(\|\varphi\|^{\mathfrak{D}})$$

Notice that the interpretation function is parametrized to a mental model discourse, so as to allow for interpretation itself to access elements of the context where it is interpreted. While the discourse parameter is idle in the four clauses of Definition 13, it will be crucial in the interpretation of the conditional, to be addressed in section 2.2.5.

The interpretation clause for *not* in Definition 13 uses a function that takes a mental model and returns its negation. Recall from our definition of mental models that negation proper ($\neg$) applies only to *mental model nuclei*. A definition of *external* negation that is to apply to a *mental model* should therefore return a mental model that is recognizably the negation of the original model, but that uses negation proper only at level of atoms. We accomplish this with the following function.[18,19]

[18]This procedure is complex, but it is more natural than it may seem. The reader may find it helpful to consider the following analogy between mental models and the formulas of a standard propositional language. Mental model molecules can be seen as conjunctions of propositional atoms and their negations, while mental models themselves, since they represent alternatives, can be seen as disjunctions of such conjunctions. In other words, mental models can be mapped straightforwardly into propositional formulas in disjunctive normal form. Negating a mental model is just as complex a procedure as negating a propositional formula in disjunctive normal form and then converting that negated formula into a disjunctive normal form of its own. This involves not just successive applications of DeMorgan's laws, to push negation into the level of atoms, but also running a normal form algorithm on this negated formula; that is, applying distributivity the number of times necessary to produce a disjunctive normal form. While this is only an analogy, it is a perspicuous way of construing the operation defined above: we take each molecule in the original mental model, reverse the polarity of each nucleus that occurs in it, collect each of these nuclei in a mental model (a disjunction) and conjoin all of the mental models thus formed. Because of the way mental model conjunction ($\times$) is defined, the result is guaranteed to be in (the mental model analog of) disjunctive normal form.

[19]While people find it easy to list the possibilities compatible with a negated *disjunction*, they fail at listing all possibilities compatible with the negation of a *conjunction* (Khemlani et al., 2012). This is captured by Definition 14, for the negation of a simple disjunction returns a mental model with only one alternative, while the negation of a simple conjunction returns a mental model with as many alternatives as there were conjuncts.

(i) a. It is not the case that A or B or C. $\textsc{Neg}(\{a,b,c\}) = \{\neg a\} \times \{\neg b\} \times \{\neg c\} = \{a \sqcup b \sqcup c\}$

b. It is not the case that A and B and C. $\textsc{Neg}(\{a \sqcup b \sqcup c\}) = \{\neg a, \neg b, \neg c\}$

Definition 14 (External negation). NEG is a function from mental models to mental models. For Γ a mental model, notice that $\Gamma = \{\alpha_0, \ldots, \alpha_n\}$ and for each $\alpha_i \in \Gamma$ we have that $\alpha_i = \bigsqcup\{a_{i0}, \ldots, a_{im_i}\}$, for $m_i + 1$ the number of mental model nuclei in α_i. Now,

$$\text{NEG}(\Gamma) = \text{NEG}(\{\alpha_0, \ldots, \alpha_n\}) = \{\neg a_{00}, \ldots, \neg a_{0m_0}\} \times \cdots \times \{\neg a_{n0}, \ldots, \neg a_{nm_n}\} .$$

We now have enough machinery to consider how the erotetic theory of reasoning captures disjunctive illusory inferences and disjunctive syllogism.

Example 10. (*Illusory inference from disjunction*) Walsh and Johnson-Laird (2004)

P_1	Either Jane is kneeling by the fire and she is looking at the TV or else Mark is standing at the window and he is peering into the garden.	$\{k \sqcup l, s \sqcup p\}$
P_2	Jane is kneeling by the fire.	$\{k\}$
C	Jane is looking at the TV.	$\{l\}$

Assuming we begin with a non-committal *tabula rasa* discourse $\langle\{0\}, \emptyset\rangle$, the default reasoning procedure predicts the following sequence of updates:

$$\langle\{0\}, \emptyset\rangle[\{k \sqcup l, s \sqcup p\}]^{\text{Up}} = \langle\{k \sqcup l, s \sqcup p\}, \emptyset\rangle$$
$$[\{k\}]^{\text{Up}} = \langle\{k \sqcup l\}, \{\{k\}\}\rangle$$

The second update is the interesting one. Q-update succeeds at interpreting the model $\{k\}$ as an answer to the question posed by the first premise: we are in a $k \sqcup l$ situation. Thus, the fallacious conclusion $\{l\}$ is "included" in the mental model that results from updating a mental model discourse with the premises. To obtain the mental model $\{l\}$ as a separate conclusion, we need to define a further operation.

2.2.4.5 Simplifying and concluding from alternatives: molecular reduction

Any theory of reasoning with mental models needs to specify operations that allow us to extract something that is established in a mental model and represent it as its own mental model. Similarly, we should be able to reduce individual alternatives so we can conclude simpler disjunctions from more complex ones. Both moves are made possible in the erotetic theory by molecular reduction.

Definition 15 (Molecular reduction). For α a mental model molecule,

$$\langle \Gamma, B \rangle [\alpha]^{\mathrm{MR}} = \begin{cases} \langle \Gamma', B' \rangle = \langle (\Gamma - \{\gamma \in \Gamma : \alpha \sqsubseteq \gamma\}) \cup \{\alpha\}, B^*(\Gamma', \Gamma') \rangle & \text{if } (\exists \gamma \in \Gamma)\ \alpha \sqsubseteq \gamma \\ \textit{undefined} & \text{otherwise} \end{cases}$$

In words, MR removes from Γ every alternative that contains the molecule α that is the target of the reduction, and then adds to Γ that molecule α. MR is undefined if its argument occurs nowhere in the mental model in discourse.

Example 11. We can now complete the last step of Example 10 with an application of MR, providing the conclusion mental model $\{\mathsf{I}\}$.

$$\langle \{\mathsf{k} \sqcup \mathsf{l}\}, \{\{\mathsf{k}\}\} \rangle [\mathsf{l}]^{\mathsf{MR}} = \langle \{\mathsf{l}\}, \{\{\mathsf{k}\}, \{\mathsf{l}\}\} \rangle$$

Example 12. Beside 'pulling out a conjunct' from a categorical mental model, MR also allows us to 'reduce' alternatives.

$$\langle \{\mathsf{a} \sqcup \mathsf{b}, \mathsf{a} \sqcup \mathsf{c}, \mathsf{d}\}, \emptyset \rangle [\mathsf{a}]^{\mathsf{MR}} = \langle \{\mathsf{a}, \mathsf{d}\}, \emptyset \rangle$$

Disjunctive syllogism, though valid, is in fact endorsed less often than the fallacy in Example 10 (García-Madruga et al., 2001). We will shortly explain how the erotetic theory accounts for this difficulty. First, it is worth considering that disjunctive syllogism is subject to an ordering effect: if the categorical premise is already established when the disjunctive premise is encountered, disjunctive syllogism becomes easier (García-Madruga et al., 2001).

Example 13. (*Disjunctive syllogism with categorical premise before disjunctive premise*)

P_1	John won't come to the party.	$\{\neg \mathsf{j}\}$
P_2	John will come to the party, or else Mary will.	$\{\mathsf{j}, \mathsf{m}\}$
C	Mary will come to the party.	$\{\mathsf{m}\}$

Because the first premise is categorical, it is added to the background set upon the first update. The reasoner is now especially attentive to anything that might contradict this established fact in any of the following premises. This can be seen in the second update: the j alternative to model $\{j, m\}$ contradicts the background, so it is discarded.

$$\langle\{0\},\emptyset\rangle[\{\neg j\}]^{\mathsf{Up}} = \langle\{\neg j\},\{\{\neg j\}\}\rangle$$
$$[\{j, m\}]^{\mathsf{Up}} = \langle\{\neg j \sqcup m\},\{\{\neg j\}\}\rangle$$
$$[m]^{\mathsf{MR}} = \langle\{m\},\{\{\neg j\},\{m\}\}\rangle$$

While disjunctive syllogism with a backgrounded categorical premise is predicted to be just as easy as the illusory inference from disjunction, it becomes harder in the canonical premise order, as seen in the next example.

Example 14. (*Canonical disjunctive syllogism*)

P_1	John will come to the party, or else Mary will.	$\{j, m\}$
P_2	John won't come to the party.	$\{\neg j\}$
Conc.	Mary will come to the party.	$\{m\}$

$$\langle\{0\},\emptyset\rangle[\{j, m\}]^{\mathsf{Up}} = \langle\{j, m\},\emptyset\rangle$$
$$[\{\neg j\}]^{\mathsf{Up}} = \langle\{j \sqcup \neg j, m \sqcup \neg j\},\{\{\neg j\}\}\rangle$$

Notice that updating with the premises alone does not yield a categorical conclusion, since we are left with two alternatives. Moreover, applying molecular reduction to m will not help, as we will still have two alternatives in the resulting model.

Unlike in the case of disjunctive syllogism with a backgrounded categorical premise as in example 10, the update procedure does not by itself eliminate the contradictory alternative. This is because the reasoner had already processed the disjunctive premise when she heard the categorical one. Eliminating the contradictory alternative here is a separate step that requires an additional 'filter' operation $[\,]^{\mathsf{F}}$, to be defined below. Using the filter operation and molecular

reduction from before, we get disjunctive syllogism in the canonical case:

$$\langle\{\mathsf{j}\sqcup\neg\mathsf{j},\mathsf{m}\sqcup\neg\mathsf{j}\},\{\{\neg\mathsf{j}\}\}\rangle[\,]^{\mathsf{F}}=\langle\{\mathsf{m}\sqcup\neg\mathsf{j}\},\{\{\neg\mathsf{j}\}\}\rangle$$
$$[\mathsf{m}]^{\mathsf{MR}}=\langle\{\mathsf{m}\},\{\{\neg\mathsf{j}\},\{\mathsf{m}\}\}\rangle$$

2.2.4.6 Removing contradictions and double negations: the filter operation

The lesson from disjunctive syllogism is the following. If our *successes* of reasoning are as much a part of the explananda as our failures, as we argued in section 2.2.2.2, then the erotetic theory needs a way to eliminate contradictory alternatives from mental models. Although naive reasoners most likely do not realize automatically when they are entertaining contradictions (Morris and Hasson, 2010), we take it that a reasoner can go over the mental model in discourse and filter out each contradictory molecule, incurring a processing cost. We call this the Filter operation, and in addition it eliminates double negations from particular atoms such as $\neg\neg p$. Much like realizing that you have been considering a contradiction is by no means a trivial step, realizing that $\neg\neg p$ represents the same proposition as p requires an application of Filter.

Definition 16 (Filter). For Γ a mental model,

$$\langle\Gamma,B\rangle[\,]^{\mathrm{F}}=\langle\Gamma',B^{*}(\Gamma',\Gamma')\rangle\ ,$$

where $\Gamma'=\{\textsc{dne}(\gamma):\gamma\in\Gamma\ \&\ \neg\textsc{Contr}(\gamma)\}$, and $B^{*}(\Gamma',\Gamma')$ is as in Definition 10. The function DNE (for double negation elimination) is inductively defined as follows.

$$\textsc{dne}(a)=\begin{cases}b & \text{if } a=\neg\neg b \text{ for some } b\in \mathit{Atoms}(\mathfrak{M})\\ a & \text{otherwise}\end{cases}$$
$$\textsc{dne}(\alpha)=\bigsqcup\{\textsc{dne}(a):a\in\mathit{Atoms}(\mathfrak{M})\ \&\ a\sqsubseteq\alpha\}$$

Notice that a single application of the DNE function does *not* ensure that we only find nuclei like p and $\neg p$ in a molecule. Indeed, to eliminate $2n$ negations from an atom in a molecule, n applications of DNE will be required. This is an intentional feature of the Filter operation: we take it that the more iterations of negation there are in an atom the more costly it is to find its simplest classically equivalent expression. Since application of optional reasoning operations such as Filter is costly, the greater difficulty of such cases is modelled by the fact that Filter may have to be applied multiple times.

Example 15. $\langle\{\mathsf{a}\sqcup\neg\neg\neg\mathsf{a}\},\emptyset\rangle[\,]^{\mathsf{F}}[\,]^{\mathsf{F}}=\langle\{\mathsf{a}\sqcup\neg\mathsf{a}\},\{\{\mathsf{a}\},\{\neg\mathsf{a}\}\}\rangle[\,]^{\mathsf{F}}=\langle\emptyset,\{\{\mathsf{a}\},\{\neg\mathsf{a}\}\}\rangle$

2.2.4.7 Fleshing out alternatives: the inquire operation

In the erotetic theory, mental models can contain molecules that subsume more than one possibility. For example, people find it difficult to list all the possibilities compatible with the negation of a conjunction (Khemlani et al., 2012).

Example 16. (*Negation of conjunction*)

It is not the case that A and B and C. $\textsc{Neg}\{\mathsf{a}\sqcup\mathsf{b}\sqcup\mathsf{c}\}$

In cases like this, evaluating the negation of a conjunction yields as many alternatives as there are atomic propositions in the conjunction: $\textsc{Neg}(\{\mathsf{a}\sqcup\mathsf{b}\sqcup\mathsf{c}\})=\{\neg\mathsf{a},\neg\mathsf{b},\neg\mathsf{c}\}$

In this case, $\neg a$, $\neg b$, and $\neg c$ each subsume multiple *classical* alternatives (i.e., fully specified, *à la* possible worlds): $\neg a$ subsumes $\neg a\sqcup b\sqcup c$, $\neg a\sqcup\neg b\sqcup c$, and so forth. We define an operation that does the job of expanding mental model molecules tracking exact truth makers by explicitly considering possibilities that those mental model molecules were tacit about. Our proposed operation in fact does nothing but allow us to ask a certain type of *questions*.[20] Inquiring in our sense upon each of the atoms will yield a more complete set of explicit alternatives. This will ultimately provide the foundation for part 2 of the erotetic principle — asking certain questions makes us classically rational.

[20]For those with an interest in classical mental model theory: here our erotetic framework allows us to do away entirely with the problematic notion of *mental model footnotes* as used by Johnson-Laird and collaborators.

Inquiring on p can be thought of as asking, "what possible alternatives are there with respect to p and its negation?" We implement this as a C-update with $\{p, \neg p\}$ followed by an application of filter. The inquire operation may be applied freely, though a choice to apply it constitutes a creative reasoning step, as discussed in section 2.2.3.2, and is therefore costly.

Definition 17 (Inquire). For any mental model Δ,

$$\langle \Gamma, B \rangle [\Delta]^{\text{Inq}} = \langle \Gamma, B \rangle [\Delta \cup \textsc{Neg}(\Delta)]^{\text{C}} [\,]^{\text{F}}$$

Successive applications of inquire on singleton mental models for all nuclei in a mental model will yield the full set of "classical" alternatives in which a premise is true, that is alternatives corresponding to each true entry on a classical truth table. Return to the example of the negation of a conjunction.

Example 17. (*Negation of a conjunction expanded to fully explicit alternatives*). To get the full range of alternatives compatible with the negation of the conjunction, we have to inquire on each conjunct. Taking the result from Example 16 as a starting point:

$$\begin{aligned}
\langle \{\neg \mathsf{a}, \neg \mathsf{b}, \neg \mathsf{c}\}, \emptyset \rangle [\{\mathsf{a}\}]^{\text{Inq}} &= \langle \{\neg \mathsf{a}, \neg \mathsf{b} \sqcup \mathsf{a}, \neg \mathsf{b} \sqcup \neg \mathsf{a}, \neg \mathsf{c} \sqcup \mathsf{a}, \neg \mathsf{c} \sqcup \neg \mathsf{a}\}, \emptyset \rangle \\
[\{\mathsf{b}\}]^{\text{Inq}} &= \langle \{\neg \mathsf{a} \sqcup \mathsf{b}, \neg \mathsf{a} \sqcup \neg \mathsf{b}, \neg \mathsf{b} \sqcup \mathsf{a}, \neg \mathsf{c} \sqcup \mathsf{a} \sqcup \mathsf{b}, \neg \mathsf{c} \sqcup \neg \mathsf{a} \sqcup \mathsf{b}, \\
&\qquad \neg \mathsf{c} \sqcup \mathsf{a} \sqcup \neg \mathsf{b}, \neg \mathsf{c} \sqcup \neg \mathsf{a} \sqcup \neg \mathsf{b}\}, \emptyset \rangle \\
[\{\mathsf{c}\}]^{\text{Inq}} &= \langle \{\neg \mathsf{a} \sqcup \mathsf{b} \sqcup \mathsf{c}, \neg \mathsf{a} \sqcup \mathsf{b} \sqcup \neg \mathsf{c}, \neg \mathsf{a} \sqcup \neg \mathsf{b} \sqcup \mathsf{c}, \neg \mathsf{a} \sqcup \neg \mathsf{b} \sqcup \neg \mathsf{c}, \\
&\qquad \neg \mathsf{b} \sqcup \mathsf{a} \sqcup \mathsf{c}, \neg \mathsf{b} \sqcup \mathsf{a} \sqcup \neg \mathsf{c}, \neg \mathsf{c} \sqcup \mathsf{a} \sqcup \mathsf{b}\}, \emptyset \rangle
\end{aligned}$$

Khemlani et al. (2012) found that none of their participants were consistently able to produce all the alternatives compatible with the negation of conjunctions. On the present account, this is no surprise: Reasoners have to creatively apply inquire and deal with an exploding number of alternatives.

If we apply inquiry systematically, reasoning, as described by the erotetic theory, will respect classical validity. With the inquire operation, there is a well-defined mechanism on the erotetic the-

ory that makes questions about what is not already explicitly considered lead to better reasoning.[21] The following theorem is proved in Appendix A.

(9) **Soundness for classical logic in the erotetic theory** — Assuming that we inquire on all propositional atoms mentioned in the premises right before updating a discourse with those premises, conclusions produced by the erotetic theory of propositional reasoning are classically valid.

Example 18. Systematic inquiry blocks the disjunctive illusory inference discussed in Example 10. The surest way to block fallacies is to inquire on every atom mentioned in a premise before updating with that premise. For simplicity, we assume here that the reasoner did this only before updating with the second premise, for that will suffice to block this fallacy.

P_1 Either Jane is kneeling by the fire and she is looking at the TV or else Mark is standing at the window and he is peering into the garden. $\{k \sqcup l, s \sqcup p\}$

P_2 Jane is kneeling by the fire. $\{k\}$

C Jane is looking at the TV. $\{l\}$

$$\langle\{0\},\emptyset\rangle[\{k \sqcup l, s \sqcup p\}]^{\mathsf{Up}} = \langle\{k \sqcup l, s \sqcup p\},\emptyset\rangle$$
$$[\{k\}]^{\mathsf{Inq}} = \langle\{k \sqcup l, s \sqcup p \sqcup k, s \sqcup p \sqcup \neg k\},\emptyset\rangle$$
$$[\{k\}]^{\mathsf{Up}} = \langle\{k \sqcup l, s \sqcup p \sqcup k\},\{\{k\}\}\rangle$$

Notice that there is now no way to apply Molecular Reduction and get the fallacious conclusion $\{l\}$. Indeed, reasoners should find that *no* obvious conclusion follows to the extent that they realize that the tempting but fallacious conclusion does not follow.

Consider another example of inquire as a promoter of classical reasoning in the case of disjunction introduction. This pattern is very counterintuitive. One study found that only 52% of participants judged the conclusion to follow from the premise in such cases. Those who did say that

[21] There is some empirical data supporting the notion that inquiring on what is not already explicitly represented can lead to improved reasoning. Baron noted that while asking subjects why a conclusion is correct does not yield improved accuracy, asking subjects why it might be incorrect does reduce errors (Baron, 1993, citing Anderson, 1982, Arkes et al., 1988, Hoch, 1985, Koriat et al., 1980). This sits well with the view presented here. Once could interpret the question, "why might C be incorrect?" as prompting inquiry on propositional atoms one has not made fully explicit yet, while asking about correctness might just tend to prompt a redo of reasoning steps already made.

the conclusion follows reported that this inference was very difficult for them to make (Braine et al., 1984). In this system, arriving at disjunction introduction is possible, but the required reasoning strategy is far from obvious. It is immediately clear that disjunction introduction is not obvious, since default reasoning via update cannot yield the conclusion. A non-default chosen reasoning strategy is required to reach the conclusion. As far as we can see, the easiest way to obtain *There is an ace or a queen* from *There is an ace* involves a strategic use of inquire followed by two applications of molecular reduction.

Example 19.

P_1 There is an ace. $\{\mathsf{a}\}$

C There is an ace or a queen. $\{\mathsf{a},\mathsf{q}\}$

$$\langle\{0\},\emptyset\rangle[\{\mathsf{a}\}]^{\mathsf{Up}} = \langle\{\mathsf{a}\},\{\{\mathsf{a}\}\}\rangle$$
$$[\{\mathsf{q}\}]^{\mathsf{Inq}} = \langle\{\mathsf{a}\sqcup\mathsf{q},\mathsf{a}\sqcup\neg\mathsf{q}\},\{\{\mathsf{a}\}\}\rangle$$
$$[\mathsf{q}]^{\mathsf{MR}} = \langle\{\mathsf{q},\mathsf{a}\sqcup\neg\mathsf{q}\},\{\{\mathsf{a}\}\}\rangle$$
$$[\mathsf{a}]^{\mathsf{MR}} = \langle\{\mathsf{q},\mathsf{a}\},\{\{\mathsf{a}\}\}\rangle$$

The erotetic theory explains the counterintuitive nature of disjunction introduction as follows: reasoning is characteristically aiming at answering questions posed by premises. Disjunction introduction requires departing from this characteristic aim of reasoning by raising an orthogonal question 'from out of nowhere'. If the characteristic aim of the system is to answer questions, we would expect resistance to raising further questions not forced by premises.

2.2.5 Beyond the core fragment: Supposition and the conditional

Evaluating what follows from multiple complex statements can seem intractable. As Johnson-Laird (2008) observes, it is very hard to see what, if anything, follows from the premises below.

P_1 The broadcast is on network TV or it is on the radio, or both. $\{n, r, n \sqcup r\}$

P_2 The broadcast is not on the radio or it is on cable TV, or both. $\{\neg r, c, c \sqcup \neg r\}$

Johnson-Laird (2008) has pointed out that if we begin this sort of reasoning problem with a strategically chosen supposition (that the broadcast is not on network TV, in the case at hand), the difficulty is markedly reduced. We provide a treatment of supposition within the erotetic theory that allows us to capture this. To our knowledge, this is the first systematic account of supposition within the mental model paradigm.

2.2.5.1 Supposing

Just as we can have a mental model discourse that represents what we take the world to be like, we can build a mental model discourse that represents what the universe is like given certain suppositions. Recall from Definition 7 (on page 30) that we encode this information with an index. We have so far been omitting this index, because it was entirely idle and our operations were completely conservative with respect to it. We now reintroduce the index parameter and expand it to include structure that keeps track of suppositions and what the suppositions are about. The basic operation of making a supposition S with respect to some mental model discourse $\langle \Gamma, B, i \rangle$ changes the index of the mental model discourse to a triple $\langle B, i, S \rangle$, including the background of the original mental model discourse, the index of the original mental model discourse, and the supposition mental model S. The extra structure is necessary because the index of the suppositional mental model discourse needs to keep track of what is established independently of the supposition, thus requiring us to keep track of the original B and i. Ultimately, we want to make sure that whatever we conclude about the world after going through the exercise of making suppositions is not sneaking in something merely supposed as established fact. Supposition of S creates a mental model discourse $\langle \Gamma, B, \langle B, i, S \rangle \rangle$ and then updates it with S. By updating with S we can see what follows from the supposition and the mental model already in discourse.

What we have in mind in defining a supposition operation are suppositions that, for all we know, might be true. The use for reasoning with such suppositions, as we argue with Johnson-Laird, emerges from cognitive capacity limitations. This kind of suppositional reasoning is quite different from and much simpler than counterfactual reasoning with suppositions that have been established to be false. Counterfactual suppositional reasoning would require additional procedures to decide

what conflicting facts to drop for the sake of reasoning. The kind of suppositional reasoning we are interested in is much simpler. Thus, we define our supposition operation in a way that makes it undefined if what is to be supposed is already established as false in B, in recognition of the fact that a much more cognitively involved mechanism, beyond the scope of this work, has to be triggered for counterfactual suppositions.

Definition 18 (Suppose).

$$\langle \Gamma, B, i \rangle [S]^{\mathrm{Sup}} = \begin{cases} \langle \Gamma, B, \langle B, i, S \rangle \rangle [S]^{\mathrm{Up}} & \text{if } (\neg \exists \beta \in B)\ (\forall b \in S \times \beta)\ \textsc{Contr}(b) \\ \textit{undefined} & \text{otherwise} \end{cases}$$

Once we have made a supposition, we can rely on the usual mental model update procedures to see what follows on the assumption of the supposition. To make this useful for reasoning, we also need an operation that allows us to discard the supposition and draw non-suppositional conclusions based on what we learned by considering what follows on the supposition. Intuitively, the envisaged operation should allow us to conclude that either we are in a situation in which the result of pursuing our supposition holds, or we are in a situation in which the supposition is false. The operation that allows us to draw this conclusion removes the special supposition index and returns us to a mental model discourse with the original index and background (where the latter will be augmented by whatever has been definitely established by the whole exercise). We shall call this operation "depose," since it can be seen as undoing the effects of supposing.[22]

Definition 19 (Depose). Let $\langle \Gamma, B', \langle B, i, S \rangle \rangle$ be a mental model with suppositions. We define

$$\langle \Gamma, B', \langle B, i, S \rangle \rangle [\,]^{\mathrm{Dep}} = \langle \Gamma \cup \textsc{Neg}(S), B^*(\Gamma \cup \textsc{Neg}(S), \Gamma \cup \textsc{Neg}(S)), i \rangle$$

The idea here is that deposing the supposition characteristically yields two alternatives. In one

[22] To make sense of the intuitive notion that we can make a supposition in reasoning and then make a further supposition, we can let suppositional discourses have the shape $\langle \Gamma, B', \langle B, i, \Sigma \rangle \rangle$, where Σ is a set of mental models, rather than just one mental model. Definition 18 would be adapted accordingly in the obvious fashion, and one would define a supplementary operation "suppose further": $\langle \Gamma, B', \langle B, i, \Sigma \rangle \rangle [\Delta]^{\mathrm{SF}} = \langle \Gamma, B', \langle B, i, \Sigma \cup \{\Delta\} \rangle \rangle [\Delta]^{\mathrm{Up}}$. The depose operation (Definition 19) would also have to adapted to handle multiple suppositions, thus: $\langle \Gamma, B', \langle B, i, \Sigma \rangle \rangle [\,]^{\mathrm{Dep}} = \langle \Gamma', B'', i \rangle = \langle \Gamma \cup \textsc{Neg}(S_0) \cup \cdots \cup \textsc{Neg}(S_n), B^*(\Gamma', \Gamma'), i \rangle$, for $\Sigma = \{S_0, \ldots, S_n\}$.

alternative, we have the result of our suppositional exercise. In the other alternative, the supposition is false. As far as we can see, this is the simplest way to meet the minimal conceptual requirement for supposition within the core fragment defined earlier. We can now consider how a strategically employed supposition may simplify a reasoning problem.

Example 20. (*Difficult reasoning problem simplified through use of supposition*). Instead of simply updating a mental model discourse with the premises in the example at the beginning of section 6, we can make a supposition and update with the premises in light of it. Using this strategy, we never need to consider a large number of alternatives at a time. First, we suppose $\{\neg n\}$, then we update with the premises and finally depose.

$$\begin{aligned}
\langle\{0\},\emptyset,i\rangle[\{\neg n\}]^{\mathsf{Sup}} &= \langle\{\neg n\},\{\{\neg n\}\},\langle\emptyset,i,\{\neg n\}\rangle\rangle \\
[\{n,r,n\sqcup r\}]^{\mathsf{Up}} &= \langle\{\neg n\sqcup r\},\{\{\neg n\},\{r\}\},\langle\emptyset,i,\{\neg n\}\rangle\rangle \\
[\{\neg r,c,c\sqcup\neg r\}]^{\mathsf{Up}} &= \langle\{\neg n\sqcup r\sqcup c\},\{\{\neg n\},\{r\},\{c\}\},\langle\emptyset,i,\{\neg n\}\rangle\rangle \\
[\,]^{\mathsf{Dep}} &= \langle\{\neg n\sqcup r\sqcup c\}\cup \textsc{Neg}(\{\neg n\}),\emptyset,i\rangle \\
&= \langle\{\neg n\sqcup r\sqcup c,\neg\neg n\},\emptyset,i\rangle \\
[\,]^{\mathsf{F}} &= \langle\{\neg n\sqcup r\sqcup c,n\},\emptyset,i\rangle
\end{aligned}$$

With supposition, the erotetic theory of reasoning is now powerful enough to provide classical completeness (see Appendix A). One crucial inference pattern in classical logic is *ex falso*, or the principle of explosion: from a contradiction, any proposition follows. This inference is far from obvious to naive reasoners, as pointed out by Harman (1986), and as recognized by anyone who has ever taught an introductory logic course. Though completeness guarantees that explosion holds for our system, it is an instance of a reasoning pattern that is far from obvious and requires ingenuity on the part of the reasoner. On the erotetic theory, the aim of reasoning is to answer questions, not raise them without prompt. To get explosion, we would have to diverge from this default aim and make our mental model more "inquisitive" without prompt, in this case via a supposition operation.

Example 21. From *P and not P* any *Q* follows. Updating with a contradiction directly will not immediately result in the absurd model, but an application of Filter to the result of updating with a contradiction will. Updating with internally consistent but contradictory premises (say, $\{\mathsf{p}\}$ and $\{\neg\mathsf{p}\}$) however will directly produce the absurd model, given that C-update always checks to see if a new premise is consistent with the set B.

From the absurd mental model, we can get any arbitrary model Γ via supposition on $\textsc{Neg}(\Gamma)$. We illustrate this for q below.

$$\begin{aligned}\langle\emptyset,\emptyset,\mathsf{i}\rangle[\{\neg\mathsf{q}\}]^{\text{Sup}} &= \langle\emptyset,\emptyset,\langle\emptyset,\mathsf{i},\{\neg\mathsf{q}\}\rangle\rangle\\ [\,]^{\text{Dep}} &= \langle\emptyset\cup\textsc{Neg}(\{\neg\mathsf{q}\}),\{\{\neg\neg\mathsf{q}\}\},\mathsf{i}\rangle\\ &= \langle\{\neg\neg\mathsf{q}\},\{\{\neg\neg\mathsf{q}\}\},\mathsf{i}\rangle\\ [\,]^{\mathsf{F}} &= \langle\{\mathsf{q}\},\{\{\neg\neg\mathsf{q},\mathsf{q}\}\},\mathsf{i}\rangle\end{aligned}$$

2.2.5.2 A semantics for the indicative conditional in the erotetic theory

We postulated operations to explain how supposition can simplify reasoning. We can now use the same operations to construct an analysis of the indicative conditional that captures the intuition that the process of supposition is crucial to its meaning (Braine and O'Brien, 1991). Our analysis is similar to a dynamic analysis of the indicative conditional recently proposed on linguistic grounds by Starr (*forthcoming*; see also Mackie, 1973, for some similarities). This convergence between an account of the conditional motivated on linguistic grounds and the present account motivated by patterns in reasoning is surely to be welcomed.

Definition 20 (Conditional as supposition).

$$\|\textit{if } \varphi,\psi\|^{\mathfrak{D}} = \Gamma, \text{ such that } \mathfrak{D}[\|\varphi\|^{\mathfrak{D}}]^{\text{Sup}}[\|\psi\|^{\mathfrak{D}}]^{\text{Up}}[\,]^{\text{Dep}} = \langle\Gamma,B,i\rangle$$

The idea behind this analysis of the conditional is that 'if p then q' encodes the following instruction to the interpreter: "Update your mental model discourse with the the mental model you get from supposing p, updating with q, and undoing the supposition." This analysis yields

$\{p \sqcup q, \neg p\}$ for 'if p then q'. This result can be expanded to $\{p \sqcup q, \neg p \sqcup q, \neg p \sqcup \neg q\}$ via inquiring on q. This corresponds to the full set of alternatives in a classical truth table.

We can now consider how the erotetic theory of reasoning captures patterns of conditional reasoning.

Example 22. (*Modus ponens is easier than modus tollens*). Adults find MP extremely easy (Braine and Rumain, 1983). Modus tollens is harder than modus ponens (Evans et al., 1993; Barrouillet et al., 2000). On the erotetic theory, modus tollens in the canonical order of premises requires an application of contradiction filter.

Modus Ponens

P_1	If the card is long then the number is even.	$\{\mathsf{l} \sqcup \mathsf{e}, \neg \mathsf{l}\}$
P_2	The card is long.	$\{\mathsf{l}\}$
C	The number is even.	$\{\mathsf{e}\}$

$$\langle\{0\},\emptyset,\mathsf{i}\rangle[\{\mathsf{l} \sqcup \mathsf{e}, \neg \mathsf{l}\}]^{\mathsf{Up}} = \langle\{\mathsf{l} \sqcup \mathsf{e}, \neg \mathsf{l}\},\emptyset,\mathsf{i}\rangle$$
$$[\{\mathsf{e}\}]^{\mathsf{Up}} = \langle\{\mathsf{l} \sqcup \mathsf{e}\},\{\{\mathsf{l}\},\{\mathsf{e}\}\},\mathsf{i}\rangle$$
$$[\mathsf{e}]^{\mathsf{MR}} = \langle\{\mathsf{e}\},\{\{\mathsf{l}\},\{\mathsf{e}\}\},\mathsf{i}\rangle$$

Modus Tollens

P_1	If the card is long then the number is even.	$\{\mathsf{l} \sqcup \mathsf{e}, \neg \mathsf{l}\}$
P_2	The number is not even.	$\{\neg \mathsf{e}\}$
C	The card is not long.	$\{\neg \mathsf{l}\}$

$$\langle\{0\},\emptyset,\mathsf{i}\rangle[\{\mathsf{l} \sqcup \mathsf{e}, \neg \mathsf{l}\}]^{\mathsf{Up}} = \langle\{\mathsf{l} \sqcup \mathsf{e}, \neg \mathsf{l}\},\emptyset,\mathsf{i}\rangle$$
$$[\{\neg \mathsf{e}\}]^{\mathsf{Up}} = \langle\{\mathsf{l} \sqcup \mathsf{e} \sqcup \neg \mathsf{e}, \neg \mathsf{l} \sqcup \neg \mathsf{e}\},\{\{\neg \mathsf{e}\}\},\mathsf{i}\rangle$$
$$[\,]^{\mathsf{F}} = \langle\{\neg \mathsf{l} \sqcup \neg \mathsf{e}\},\{\{\neg \mathsf{e}\},\{\neg \mathsf{l}\}\},\mathsf{i}\rangle$$
$$[\neg \mathsf{l}]^{\mathsf{MR}} = \langle\{\neg \mathsf{l}\},\{\{\neg \mathsf{e}\},\{\neg \mathsf{l}\}\},\mathsf{i}\rangle$$

In a way entirely parallel to the case of disjunctive syllogism, the present account also captures that modus tollens becomes easier if the negative premise is encountered before the conditional (Girotto et al., 1997).[23]

In defining the supposition operation, we noted that we were ruling out suppositions that are already established to be false since this would involve a different and more sophisticated kind of reasoning. Since we have used this definition of supposition to define a semantics for "if", we immediately get a linguistic upshot. It has been noted in the linguistics and psychology literature that the use of an indicative conditional is highly infelicitous if the antecedent is established to be false (Stalnaker, 1975; von Fintel, 1999; Gillies, 2009; Starr, forthcoming).

(10) Bob never danced. #If Bob danced, Leland danced.

As on Starr's analysis, the present account of conditionals makes the update procedure determined by the conditional undefined in this case, accounting for the infelicity.

From a psychological perspective, the fact that the conditional is undefined if the antecedent has been established to be false accounts for an interesting empirical observation. There is an asymmetry between the tasks of listing alternatives compatible with a conditional and the task of evaluating whether a conditional is true at various given scenarios. The difference has to do with how false antecedents are treated (Barrouillet et al., 2008). People have no trouble listing cases in which the antecedent is false if they are given a conditional and are asked to list alternatives that conform to the conditional (Barrouillet and Lecas, 1998). However, if subjects are given scenarios and then asked if a conditional is true in those scenarios, they will omit false-antecedent cases (Evans et al., 1993).

Philosophers as well as psychologists have pointed out that it is important to capture the fact that so-called "material antecedent" inferences are not intuitive to people (Starr, forthcoming; Oaksford and Chater, 2007), in other words examples like the following seem wrong:

[23]The issue of what we look to for potential falsifiers of a conditional comes up in the famous Wason card selection task (Wason, 1968) that much of the literature on reasoning with conditionals focuses on. Considerably more space would be needed to bring this task within the domain of the erotetic theory, since it involves quantification (a much harder fragment to cover with the degree of precision we aspire to here), as well as a more complex task for the participants than the "what, if anything, follows?" questions we are primarily addressing.

(11) Bob did not dance. Therefore, if Bob danced then he will become President.

Our proposal blocks these inferences. A conditional conclusion is not interpretable if the antecedent is taken by the hearer to be false; the update procedure is undefined in this case. Probabilistic approaches like Oaksford and Chater (2007) would also block this inference, but, in virtue of having a probability-based semantics for "if ... then", they also block various other inferences, which Starr (forthcoming) has argued should not be blocked for relevant interpretations of "if ... then". The issues are delicate and most likely there are multiple possible interpretations for various "if ... then" sentences that would yield different inference patterns. For example, there plausibly are "habitual" interpretations that may be captured by something like Stenning and van Lambalgen's nonmonotonic logic proposal (Stenning and van Lambalgen, 2008a), like the following:

(12) If the match is struck, it lights.

From this, we would not conclude that if the match is struck and it is wet, it will light. On the present account, this would have to be concluded. However, it is not obvious that the interpretation of "if ... then" at issue in this example is the same that is at issue in the examples we have been primarily concerned with. One might argue that in (12) we are giving the conditional a "habitual" interpretation that is absent in the other cases. The potential for multiple interpretations here means that there is room for multiple complementary theories. That said, there are two interesting benefits for the present proposal on reasoning with conditionals that we will now turn to.

One advantage of our proposal comes from empirical data on novel illusory inference from a disjunctive premise to a conditional conclusion that previous accounts do not seem to capture.

Example 23.

P_1 John and Mary will come to the party, or Bill and Sue will. $\{j \sqcup m, b \sqcup s\}$

C If John comes to the party, then Mary will come as well. $\{j \sqcup m, \neg j\}$

We propose that reasoners would take it that they can conclude the conditional if by supposing the antecedent and updating with the premise, they can arrive at the conditional. It is not surprising that evaluating whether a conditional conclusion follows would prime reasoners to use supposition, since the operations involved in making suppositions are integral to the linguistic

meaning of the indicative conditional on our analysis (NB: the same reasoning strategy also yields non-fallacious conditional transitivity inferences, see supplementary example 36).

$$\langle\{0\},\emptyset,i\rangle[\{j\sqcup m, b\sqcup s\}]^{\text{Up}} = \langle\{j\sqcup m, b\sqcup s\},\emptyset,i\rangle$$
$$[\{j\}]^{\text{Sup}} = \langle\{j\sqcup m\},\{\{j\},\{m\}\},\langle\emptyset,i,\{\{j\}\}\rangle\rangle$$
$$[m]^{\text{MR}} = \langle\{m\},\{\{j\},\{m\}\},\langle\emptyset,i,\{\{j\}\}\rangle\rangle$$
$$[\,]^{\text{Dep}} = \langle\{j\sqcup m,\neg j\},\emptyset,i\rangle$$

In an experiment similar to that of Johnson-Laird and Savary (1999) using an online survey conducted, we asked participants yes/no questions about whether various inferences followed from sets of statements. We found that all of our 20 participants endorsed the illusory inference from the disjunctive premise to the conditional conclusion. Overall performance on control problems was significantly better. The subset of 13 participants with perfect performance on all control problems still uniformly endorsed the fallacious inference. All illusory and control problems in the experiment as well as proportions of responses can be found in supplementary example 31.

A further advantage of our proposal is that it captures a linguistic connection between questions and conditionals. Perhaps unsurprisingly, given the importance questions play in our theory, the present analysis is the only one beside Starr (forthcoming) — the latter a purely semantic analysis without its own motivation in empirical reasoning data — that can account for the observation that *if*-clauses can both introduce a conditional and serve as the content clause of a question (Harman, 1979; Haiman, 1978). Consider:

(13) a. If Jack danced, the music must have been good.
b. John asked if Jack danced.

The question entertained in (13b) corresponds to the set of alternatives $\{d, \neg d\}$, following the standard Hamblin (1958) account of the semantics of questions discussed earlier. So what is the contribution of the *if*-clause? We can simply take it that in cases like (13b), where there is no *then*-clause,

the relevant argument in the conditional-as-supposition analysis is the null nucleus $\{0\}$.

$$\|\mathit{if}\ \varphi\|^{\mathfrak{D}} = \Gamma, \text{ such that } \mathfrak{D}[\|\varphi\|^{\mathfrak{D}}]^{\text{Sup}}[\{0\}]^{\text{Up}}[\,]^{\text{Dep}} = \langle \Gamma, B, i \rangle$$

Now, updating with the null nucleus does not change anything, so we can leave it out. Alternatively, we could attribute the entire update operation sandwiched between the suppose and depose procedures to the linguistic contribution of the then-clause, which would just make the update operation "disappear" if there is no then-clause. Both avenues yield the following result:

Definition 21 (If). $\|\mathit{if}\ \varphi\|^{\mathfrak{D}} = \Gamma, \text{ such that } \mathfrak{D}[\|\varphi\|^{\mathfrak{D}}]^{\text{Sup}}[\,]^{\text{Dep}} = \langle \Gamma, B, i \rangle$

Thus, $\|\text{if Jack danced}\|^{\mathfrak{D}} = \{d, \neg d\}$, and it falls out of our account of the conditional that *if* without a *then*-clause contributes a question.

With the core fragment and the supposition operator, we now have enough machinery to prove that the erotetic theory of reasoning allows for a reasoning strategy that is classically sound and complete. This is done in Appendix A.

2.2.6 Conclusion: Questions make us rational

We have proposed a new theory of reasoning based on the erotetic principle in (1).

(1) **The erotetic principle**

Part I — Our natural capacity for reasoning proceeds by treating successive premises as questions and maximally strong answers to them.

Part II — Systematically asking a certain type of question as we interpret each new premise allows us to reason in a classically valid way.

We showed how the erotetic theory of reasoning derives a large number of naive reasoning patterns described in the literature based on a dynamic update procedure that implements the principle in (1). The theory accomplishes this while resorting to interpretations independently motivated in the linguistic and philosophical literatures.

We argued that the erotetic theory of reasoning accounts addresses two of the key problems for a theory of reasoning: The first problem was to account for the various systematic divergences from standards of classical correctness in naive reasoning that have been observed experimentally. The second concerned how our natural reasoning capacity can make it possible to learn how to reason correctly by classical standards.

With the formal framework of the erotetic theory, we can explore more seriously the potential for connections between semantics as studied in linguistics and philosophy and the empirical psychology of reasoning. As noted above, dynamic approaches to interpretation were independently proposed in psychology (Johnson-Laird and Stevenson, 1970) on the one hand, and in linguistics (Karttunen, 1976; Heim, 1982) and philosophy (Kamp, 1981) on the other, but have heretofore been developed in parallel without any significant connection between psychology and linguistics/philosophy. It is also an interesting development that the natural analysis of suppositional reasoning on the erotetic theory yielded an account of the semantics of the indicative conditional that is similar in important ways to an account recently defended on purely linguistic and philosophical grounds (Starr, forthcoming). A formally rigorous approach to reasoning with mental models as presented with the erotetic theory makes it possible to find independent motivation for various operators that can be used for such accounts. For example, the machinery used to analyze the conditional was independently motivated by the need to make sense of how we can use suppositions to aid reasoning that does not involve conditional premises. Moreover, within a theory of reasoning, which is not limited to reasoning with verbal premises (Bauer and Johnson-Laird, 1993), the very notion of a representation of discourse that is independent from linguistic meaning is motivated.

The key proposal of the erotetic theory of reasoning is that the peculiar patterns of naive reasoning are due to the default system for reasoning treating successive premises as questions and answers. Our natural capacity for reasoning does not have its peculiar features because it aims at quasi-approximations of classically valid reasoning. Rather, the system aims at answering questions. What is remarkable is that this endowment provides resources that support the discovery of a reasoning strategy that in fact allows us to reason with classical validity. Questions, in a particular way we made precise in this chapter, make us rational.

Chapter 3

Scalar implicature in propositional reasoning

3.1 Introduction

Psychological accounts of human inference-making behavior for the most part consider that behavior to be shedding a direct light on our general-purpose reasoning faculty.[24] Mental model theory (Johnson-Laird, 1983), heuristics and biases (Tversky and Kahneman, 1974), and probabilistic theories of reasoning (Oaksford and Chater, 2007) are all examples of this kind of program for the study of reasoning. So is the erotetic theory of reasoning just presented in Chapter 2.

In this chapter, I outline an alternative source for at least some of our failures of reasoning: an interpretation-based account of some reasoning failures. On an interpretation-based view, "failures of reasoning" is a misnomer, at least for a representative class of attractive but fallacious inference patterns. Rather than being the product of classically unsound general-purpose reasoning mechanisms, some of the "mistakes" we make in fact arise from more complex interpretive processes

[24]There are exceptions to this tendency. The ones I am familiar with are cases where specific classes of well-known inference-making behavior (often simply specific data points) are explained away with recourse to interpretive mechanisms. For example, in the literature on the conjunction fallacy (discussed in some depth in Chapter 4), Dulany and Hilton (1991) rather carefully consider the contribution of pragmatic processes to the fallacy, while Hertwig and Gigerenzer (1999) consider even more narrowly semantic alternative accounts of the phenomenon. What is lacking, to the best of my knowledge, is a framework for the study of reasoning that considers the possible effects of interpretive processes *first*, in an entirely systematic fashion.

than meet the eye. Specifically, I show how *illusory inferences from disjunction* (Johnson-Laird and Savary, 1999; Walsh and Johnson-Laird, 2004), can be accounted for if we assume a classically sound reasoning module that acts upon the pragmatically strengthened meaning of premises, rather than on their literal meaning. Schematically, the illusory inference from disjunction is a conclusion of b from the premises $(a \wedge b) \vee (c \wedge d)$ and a. This inference is fallacious, as it is proved invalid at a world where a, c, and d are true but b is false.

The account I give here is in sharp contrast with accounts of the same phenomena from psychology, which assume simple interpretive processes feeding a reasoning module specially tailored to give rise to the observed non-classical inference patterns. The interpretation-based account also makes strong predictions, distinct from the predictions of its reasoning-based competitors from psychology. These clear predictions allow for a novel experimental paradigm to be sketched, which aims at comparing the two classes of theories. I show how in principle we will be able to test these predictions and decide which kind of account (reasoning-based or interpretation-based) is right for which classes of fallacies.

Compelling fallacies and repugnant validities in an interpretation-based approach

It is useful to distinguish between two broad ways in which human reasoning can fail. First, we can fall prey to *compelling fallacies*. Compelling fallacies are classically invalid inference patterns that reasoners often accept. To begin with a simple and well-known example, consider the fallacy of affirming the consequent, accepted by approximately 75% of subjects in a study by Barrouillet et al. (2000).

(14)
P_1: If the card is long then the number is even.
P_2: The number is even.
Concl.: The card is long.

Under the assumption that P_1 is interpreted as a simple conditional, rather than a biconditional, the inference in (14) is fallacious. It is consistent with both premises that the card not be long, and therefore the conclusion does not follow. Notice that this assumption about interpretation is

crucial: if P_1 is interpreted as a biconditional, then the inference does follow. Insofar as we can independently motivate the hypothesis that P_1 is in fact interpreted biconditionally, the inference pattern in (14) lends itself to a simple explanation consonant with an interpretation-based program for failures of reasoning. Unsurprisingly, the fallacy of affirming the consequent has in fact received analyses in the spirit this program. For example, Horn (2000) discusses several historic accounts of the interpretive move from 'if' to 'if and only if' in terms of pragmatic strengthening. Using insights from recent work in formal pragmatics, the account can be simplified as follows.

There is ample evidence for the existence of exhaustification or strengthening operations in natural languages.[25] These operations perform a job very close to what lexical items like English 'only' do.[26] For example, in question answering, it is well established that term answers are most naturally interpreted exhaustively. In what follows, SMALL CAPS indicate focused material.

(15) Q: Which professors does John like?
A: Mary.
Interpretation: "John only likes MARY."

The pragmatically strengthened interpretation of an utterance can also be viewed as the result of an exhaustification operation, applying to the literal meaning of the utterance. As with the case of term answers, a paraphrase with *only* can help illustrate this intuition:

(16) a. John likes some of his professors.
Implicature: John does not like all of his professors.
b. John only likes SOME of his professors.
Entailment: John does not like all of his professors.

[25] I will remain largely tacit about the issue of whether exhaustification/strengthening operations are performed by (covert) *operators* part of the some level of syntactic representation, or whether these operations can be seen as the workings of a global strengthening mechanism with no corresponding operators in the syntax. The issue will in fact turn out to be immaterial for the reasoning data discussed here, though in principle it may bear on extensions of the interpretation-based theory to other reasoning data.

[26] There are important differences. Most notably, while 'only' *presupposes* its prejacent (the proposition expressed by the sentence without 'only'), exhaustification/strengthening operations merely *assert* it. Since I move from this informal illustration with 'only' to a fully spelled-out neo-Gricean account of pragmatic strengthening shortly, it is safe to ignore this fact about 'only' for now.

Suppose now that an exhaustification operation analogous to *only* applies to the conditional premise P_1 of 14, returning its pragmatically strengthened interpretation. The exhaustive interpretation of (17a) would be as in (17b). Conjoining the literal meaning (17a) with the strengthened (17b) then gets us the interpretation in (17c) for the first premise of (14).

(17) a. If the card is long then the number is even.

b. *Only* if the card is long is the number even.

c. If, and only if, the card is long is the number even.

If (17c), rather than (17a), is the interpretation of the first premise of (14), then (14) is in fact classically valid. There is no fallacy, no failure of reasoning.[27]

This account of affirming the consequent is a clear and important precursor to the approach I give later in this chapter. Unfortunately, to the best of my knowledge, work from within linguistic semantics on this topic is scarce and has focused entirely on a small number of fallacies discussed in philosophy and rhetoric, mostly related to inferences involving conditionals, whose existence scholars have been aware of since classical antiquity.[28] Clearly, a viable interpretation-based program for the study of reasoning failures must extend to the more sophisticated reasoning data discussed today in the psychology of reasoning.

The sizable body of reasoning research conducted by psychologists since the 1960's has produced examples of fallacious inferences far more complex than the example of affirming the consequent in (17). This chapter focuses on the illusory inference from disjunction, exemplified in (18), and accepted by approximately 80% of subjects in an experiment by Walsh and Johnson-Laird (2004).

[27]There is, of course, still room for a normative perspective on human inference making. But interpretation-based accounts of "failures of reasoning" in this spirit suggest a shift in focus for normative enterprises. If this view is right, then at least *some* fallacies are the result of misunderstandings in interpretation, rather than mistakes in general-purpose reasoning. Reasoners are accommodating the existence of communicative intentions and therefore computing implicatures in contexts that *normatively* ought not to be considered communicative contexts. Concretely, subjects in a standard reasoning experiment assume that the linguistically presented reasoning problems they are given are the utterances of some speaker with communicative intentions, making it legitimate for subjects to calculate implicatures for reasoning problems.

[28]Semanticists have generally not studied systematically relevant data from psychology, but the converse is also largely true: work on fallacies from psychology tends to assume that interpretive processes are either self-evident or inconsequential. A notable exception is the work of Stenning and van Lambalgen (2008a), who combine the methodologies and data of psychology with a deep understanding of non-classical logics and a serious concern for issues of interpretation.

(18) P_1: Either Jane is kneeling by the fire and she is looking at the window or otherwise Mark is standing at the window and he is peering into the garden.

P_2: Jane is kneeling by the fire.

Concl.: Jane is looking at the window.

The inference in (18) is fallacious: suppose Jane is kneeling by the fire but *not* looking at the window, while Mark is both standing at the window and peering into the garden. This situation would model both premises but falsify the conclusion.

It is useful to abstract away from the specific content given in (18) and consider the schema behind this inference, given in a standard propositional language.[29] The second premise and the conclusion have trivial propositional analyses, but P_1 might be interpreted as an inclusive or an exclusive disjunction. The two options are given as P_1 and P_1' in (19).

(19) P_1: $(a \wedge b) \vee (c \wedge d)$ $\quad$ P_1': $(a \wedge b \wedge \neg(c \wedge d)) \vee (c \wedge d \wedge \neg(a \wedge b))$

P_2: a

Concl.: b

Notice that either choice of P_1 or P_1' as the interpretation arrived at by subjects in Walsh and Johnson-Laird's (2004) study results in an invalid inference pattern. Both inferences schematized in (19) are falsified at a model making a, c, and d true, but b false. Thus, and unlike the case of affirming the consequent in (18), the illusory inference from disjunction lacks an obvious interpretation-based account. This inference pattern will be the focus of section 3.3.2, where I show that in fact such an account follows surprisingly from largely uncontroversial assumptions about scalar implicature.

Compelling fallacies like the two just exemplified must be distinguished from another kind of failure of reasoning: *repugnant validities*. Repugnant validities are classically valid inference

[29] Walsh and Johnson-Laird (2004) took care to ensure that the attractiveness of the pattern in (18) was not merely due to the close internal affinity within the contents of each disjunct—each disjunct being about the same person. Walsh and Johnson-Laird tested every permutation of the atomic (propositional) sentences in P_1 of (18), in particular the variant of (18) where each disjunct of P_1 contains a conjunction whose conjuncts are about different actors. For example: "Either Jane is kneeling by the fire and Mark is standing at the window or otherwise Jane is looking at the TV and Mark is peering into the garden." They also tested examples with four distinct actors. They found a gradual decline in acceptance rates as the number of actors was increased (from only one actor to four distinct actors). However, the difference was only reliable between the one-actor problems (which I do not discuss in this chapter) and the four-actor problems.

patterns that reasoners often reject. For example, disjunction introduction, exemplified in (20), is notoriously hard to accept (Braine et al., 1984).

(20) P_1: The card is long.
Concl.: The card is long or the number is even.

Perhaps more surprisingly, *modus tollens* (21) has an appreciably lower acceptability rate than *modus ponens* (Evans et al., 1993; Girotto et al., 1997).

(21) P_1: If the card is long then the number is even.
P_2: The number is not even.
Concl.: The card is not long.

A complete interpretation-based account of reasoning failures must either extend naturally to repugnant validities or provide a principled explanation for its inability to account for them. While this chapter is exclusively about compelling fallacies, I conclude this section with a remark on the repugnant validity in (20).

The most promising interpretation-based route to explaining compelling fallacies is the phenomenon of scalar implicature, but repugnant validities might require explanations using other elements of pragmatics. For example, resistance to disjunction introduction (20) receives a rather straightforward account purely in terms of primary implicatures (see section 3.2.2 for a discussion of primary implicatures and their relation to scalar implicatures). By virtue of being a disjunction, the conclusion of (20) prompts an ignorance implicature to the effect that the "speaker" of the sentence is not in a position to assert either of the disjuncts. In particular, the "speaker" is not in a position to assert that the card is long. However, the proposition that the card is long is in fact asserted in the premise of (20). Consequently, the sequence Premise–Conclusion in (20) would be extremely infelicitous in a conversational context, given the primary implicatures of the conclusion.

3.2 Toward an interpretation-based account of reasoning failures

In this section, I define basic desiderata for interpretation-based accounts, focusing on the role of implicatures.

3.2.1 Central components of an interpretation-based account

In the most general sense, an interpretation-based theory of our failures of reasoning has the following three central ingredients.

1. Commitments with respect to *literal* linguistic content — presumably a unidimensional classical semantics, though nothing in the theory will hinge on this null hypothesis.

2. A mechanism for enriching (strengthening) the literal content in a way that

   (a) assigns to each premise the interpretation required to get the observed reasoning patterns as a product of a classically sound reasoning module,

   (b) can be independently motivated as a plausible interpretation of the premises by purely linguistic criteria, and

   (c) introduces no mischief into extant accounts of enriched, non-literal meaning.

3. Basic commitments about reasoning processes—how do reasoners go about checking whether something follows from a set of premises?

In many if not most cases, and certainly for the data points discussed in this chapter, the crux of any interpretation-based proposal is ingredient two above: the mechanism for strengthening content.[30] I propose what I take to be the most natural move: to build a theory of reasoning failures based on independently motivated accounts of scalar implicatures. In an *implicature*-based theory, component two above is most naturally concretized as in (22).

[30] Other classes of data may be susceptible to interpretation-based accounts where all or most of the "action" happens at the very first stage of interpretation, concerned with literal content. The interpretation-based analysis of the conjunction fallacy in Chapter 4 of this dissertation crucially involves both the literal and the enriched interpretive stages.

(22) **Ingredient 2 of an interpretation-based theory, in the special case of a *(scalar) implicature*-based theory:**

A theory of implicature that

a. assigns to each premise the interpretation required to get the observed reasoning patterns,

b. can provide compelling independent evidence that the interpretations obtained for a. are implicatures in the appropriate sense, and

c. introduces no mischief into the account of implicatures that do not (obviously) bear on matters of reasoning failures.

These essential constraints on an implicature-based theory carve out a sizable space of possibilities. For example, nothing above helps us decide whether to opt for a theory of implicature operating at the global level (e.g. Sauerland, 2004) or one with enrichment operators that can be embedded in syntactic structure (Chierchia, 2004). In what follows, I remain neutral about debates within the literature on implicature that do not seem (at the moment) to bear on the tenability of these theories as accounts of some reasoning failures.

3.2.2 Sketching an implicature-based account

Since Grice's seminal work on implicature, we take it that conversational implicatures arise from processes of pragmatic reasoning on the part of the hearer of an utterance. It is customary to distinguish *primary implicatures* from *secondary implicatures*. Primary implicatures are conclusions about what the speaker *is not in a position to assert*; they are weak conclusions about the speaker's epistemic attitudes. Secondary implicatures, typically seen as deriving from primary implicatures via a strengthening procedure, are conclusions about what the speaker *believes to be false*.

Scalar implicatures are a certain kind of quantity implicatures (both primary and secondary) in which the hearer compares the speaker's utterance S to a certain class of statements that the speaker could have made but chose not to: those statements that result from substituting elements of S with members of their *scales*. The idea that (at least some) lexical items come with lexically

stipulated scales was introduced by Horn (1972) to address the *symmetry problem* with Grice's original formulation of the process that calculates quantity implicatures. The traditional view of how scalar implicatures are calculated is as follows. When interpreting an utterance S, a hearer will

1. Compute the alternatives to S, by replacing scalar lexical items in S with elements of their scales.

2. Collect those propositions φ that are (1) alternatives to S and (2) stronger than S (that is, $\varphi \vDash S$ but $S \nvDash \varphi$). Call this set SA_S (for it is the set of stronger alternatives to S).

3. Compute *primary implicatures*: for each proposition $\varphi \in SA_S$, "the speaker does not believe (i.e. is not in a position to assert) φ."

4. Compute *secondary implicatures*: assume that the speaker is opinionated, that is, for every (relevant) φ the speaker either believes φ or its negation. It follows by disjunctive syllogism that the primary implicatures can be strengthened from the form "the speaker does not believe φ" to the form "the speaker believes that φ is false."

For reasons we need not go into, Horn scales undergenerate, leaving out certain primary implicatures even in very simple cases, and certain secondary implicatures in more complex cases. Instead of appealing to Horn scales in step 1. above, I import the fundamentals of the theory of syntactically generated alternatives given by Katzir (2007) and Fox and Katzir (2011).

Katzir (2007) and Fox and Katzir (2011) propose to incorporate an appeal to judgments of complexity of the allowed substitutions. The intuition is that, while we want to allow for p to be an alternative to $p \vee q$, we do not want $p \vee q$ to be an alternative to p.[31] On this view, the difference between the two cases is that in the former we consider an alternative no more complex than the original sentence, while in the latter we consider an alternative more complex than the original sentence. If we somehow uniformly prevent more complex substitutions from being considered

[31] Sauerland (2004) offers an alternative solution that maintains the notion of Horn scales. He proposes to add binary operators L and R as scale-mates of 'or' and 'and', where $\varphi \; L \; \psi = \varphi$ and $\varphi \; R \; \psi = \psi$. It is difficult to deny that this is an inelegant solution, as there appears to be no good reason to say that the lexical items L and R thus defined are part of the lexicon of *any* natural language. Katzir's (2007) theory of alternatives altogether avoids these difficulties.

alternatives, we avoid this problem. First, we must define a relation between syntactic structures, as in (23).

(23) For two syntactic structures S and S', we say that S' is *no more complex* than S, just in case S' can be derived from S by successive replacements of sub-constituents of S with elements of the *substitution source* for S.

Second, I define the substitution source for S as follows.[32]

(24) For S a syntactic structure, the substitution source for S in C is the union of:

a. the lexicon, and

b. the sub-constituents of S.

I introduce one final modification of the classical procedure described above. Following Sauerland (2004), I take it that the strengthening procedure from primary implicatures to secondary implicatures must obey the constraint in (25).

(25) No secondary implicature of a statement S can contradict the literal meaning of S or the primary implicatures of S.

Formally, (25) corresponds to (26):

(26) Where S is a statement and SA_S is the set of alternatives of S that give rise to primary implicatures, the secondary implicatures of S are the negations of propositions $\varphi \in SA_S$ such that $(\forall \psi \in SA_S)\ \neg\varphi \wedge S \not\vDash \psi$.

[32] Fox and Katzir (2011) in fact relativize the notions of complexity and of substitution sources to a context C. This is because they have a third contributor to the substitution source 24, namely the set of salient constituents in a context C. The motivation for this addition comes from data such as (i).

(i) It was warm yesterday, and it is a little bit more than warm today.

Consider the communicative content of the first conjunct: it implicates that it wasn't a little bit more than warm yesterday. On anyone's theory of implicature, this result depends on "It was a little bit more than warm yesterday" being an alternative of the first conjunct. In Fox and Katzir's account of the source of alternatives, this is an instance of substitution of a constituent of the first conjunct ('warm') with a constituent coming not from the lexicon (24a) or the subconstituents of the first conjunct (24b), but in fact from the second conjunct ('a little bit more than warm'), presumably a salient constituent in the context of (i). For the cases I consider in this chapter, salient constituents that are not sub-constituents will play no role.

3.2.3 Synthesis

The theory of scalar implicature that I adopt here can be described as the following procedure.

1. Compute the alternatives to S that are at most as complex as S (definition in (23)).
2. Collect those alternatives φ that are (1) alternatives to S and (2) strictly stronger than S. Call this set SA_S.
3. Compute primary implicatures: for each $\varphi \in SA_S$, "the speaker does not believe that φ."
4. Compute secondary implicatures: for each $\varphi \in SA_S$ such that the negation of φ does not contradict the literal meaning of S or any of the primary implicatures of S, conclude (that the speaker believes) that φ is false.
5. Call the conjunction of the literal meaning of S together with all of its secondary implicatures the strengthened (exhaustive) meaning of S.

There are well-known issues in the literature on implicature that I did not discuss in this section. Most notably, I did not compare this neo-Gricean approach, which works at a global level, with localist approaches (such as the one proposed by Chierchia, 2004), which locate the mechanisms that generate scalar implicatures in the semantics or the syntax, allowing for embedded implicatures. I also did not discuss the matter of whether "strictly stronger," in step 2. above, is the right notion, rather than "not weaker" (Spector, 2007). These omissions were intentional: as far as the examples discussed in this chapter, the goal of deriving reasoning data in an implicature-based account offers no criterion by which to judge these competing hypotheses on theories of implicature. That being the case, I assume a conservative account of scalar implicature while observing that the results I show here translate into equivalent results in a localist theory of scalar implicature.

3.2.4 Reasoning with implicatures

The last piece of the account is an explicit proposal for how implicatures feed into reasoning processes. In order for an implicature-based theory to bear on data collected via reasoning tasks, we first need a postulate along the following lines.

(27) Even when interpreting sentences in the absence of a speaker, as in a piece of paper in the context of an experiment, reasoners accomodate the existence of some abstract speaker, the author of the sentences under evaluation.

Finally, we need to make working assumptions about how pragmatically strengthened meanings figure into reasoning.

(28) **Reasoning in the implicature-based account**

Given a sequence of premises $P_1, \ldots, P_n$ and a conclusion C, begin by calculating the strengthened meaning of each premise, getting the sequence $P_1^+, \ldots, P_n^+$. Then, check to see if the conclusion C follows *classically* from $P_1^+, \ldots, P_n^+$.

The second step of (28) glosses over the issue of what this "checking" procedure consists of. In particular, is it model-theoretic checking, perhaps similar in spirit to mental models accounts, or is it a proof-theoretic process of trying to find a derivation of the conclusion from the premises? As far as I can see, the question is orthogonal to the account of the fallacies given in this chapter. It suffices therefore to assume that, however the general-purpose reasoning module works, it targets classically valid reasoning, succeeding if the conclusion follows classically from the premises, and failing otherwise.

One final remark is in order. Since this interpretation-based approach builds on theories of scalar implicature, it is natural to be concerned that the line between reasoning and interpretation has been dangerously blurred. After all, scalar implicatures as described above arise from a process of Gricean *reasoning*. In what sense then is this approach properly *interpretation*-based? The following demarcation line should help. I consider to be reasoning-based any account of fallacies that makes crucial use of non-classical behavior of the *general-purpose* reasoning mechanisms. The mental models approach given in Chapter 2 and in section 3.3.1 below falls squarely in this category, given the peculiar way in which the *conjoin* operation is defined. By contrast, I call interpretation-based any theory of fallacies that makes no assumptions about the general-purpose reasoning mechanisms, but rather locates the root of fallacies in language-specific processes. Gricean strengthening

procedures as described in this section are language-specific, and can therefore be distinguished from general-purpose reasoning. Accordingly, the predictions of reasoning-based accounts differ from those of interpretation-based accounts, at least in principle. For example, reasoning-based accounts, all things being equal, predict that the same errors will be made with logically equivalent linguistic statements. Interpretation-based accounts make no such prediction, among other reasons, because (otherwise) logically equivalent statements can carry very different implicatures. Predictions will also differ with respect to reasoning about premises not given in a linguistic form (e.g. reasoning about pictorially given information). I return to this issue in section 3.3.3, where I discuss concrete predictions of the interpretation-based account given here.

3.3 An interpretation-based account of illusory inferences from disjunction

3.3.1 A reasoning-based account: mental model theory as a point of departure

The only extant account of human reasoning that predicts the illusory inference from disjunction in (18) is mental model theory (Johnson-Laird, 1983, and a wealth of work by Johnson-Laird and collaborators). While a complete presentation of mental model theory would take us far afield, it is important to see a brief sketch of the account. This will provide a point of departure for the interpretation-based theory.

(18)
P_1: Either Jane is kneeling by the fire and she is looking at the window or otherwise Mark is standing at the window and he is peering into the garden.
P_2: Jane is kneeling by the fire.
Concl.: Jane is looking at the window.

On mental model theory, reasoners build mental representations (mental models) that verify each of the premises. Following what is known as the *principle of truth*, reasoners construct models that make *only* overtly stated material true, remaining tacit about anything else. Furthermore, certain connectives, such as disjunction, are represented as sets of alternative mental models, one for each

disjunct. Recall the schema for the illusory inference from disjunction in (29). Since the issue of inclusive or exclusive disjunction is imaterial, I choose the simpler inclusive formulation for ease of exposition.

(29) P_1: $(a \wedge b) \vee (c \wedge d)$

P_2: a

Concl.: b

Upon processing P_1, reasoners will construct two alternative mental models, one for each disjunct. These models represent only what is overtly stated in each disjunct. P_2 receives a straightforward interpretation as a single mental model:

(30)

P_1:	Alternative mental model 1	a model of a and of b, tacit about c and d
	Alternative mental model 2	a model of c and of d, tacit about a and b
P_2: a		a model of a, tacit about b, c, and d

Next, the reasoning module performs a *conjoin* operation on these premises, combining the mental models. This operation is crucially non-classical. The following is one way to describe it, inspired by the corresponding operation in Koralus and Mascarenhas's (2013) rethinking of mental model theory (see Chapter 2 of this dissertation). Because the model for P_2 is entailed by alternative mental model 1 for P_1, but not alternative mental model 2, reasoners ignore the second alternative model for P_1. Since alternative mental model 2 is no longer being attended to, the conjoined interpretation of the premises consists of only one mental model: a model of a and of b, tacit about everything else. The fallacious conclusion, b, follows immediately from a simple reasoning operation amounting to conjunction elimination.

Mental model theory can account for a large body of data on frequencies of acceptance and rejection of inferences at least as well as its competitors.[33] Because of the dynamic nature of its *conjoin* operation, mental model theory provides immediate accounts of many effects of the ordering of premises on the attractiveness of fallacious conclusions (see for example Girotto et al.,

[33]For example, Oberauer (2006) compares mental model theory to three of its main competitors with respect to fine-grained data on inferences involving conditionals. His study concludes that a mild modification of mental model theory fares better than its competitors at capturing the observed patterns of acceptance and rejection.

1997). Moreover, mental model theory was the only theory of reasoning that predicted the illusory inferences discussed in this chapter (Johnson-Laird and Savary, 1999; Walsh and Johnson-Laird, 2004).

Despite being a major simplification of the mental model account, this sketch illustrates two important problems with the theory. First, mental model theory has been criticized by logicians and philosophers for not being formal enough (Hodges, 1993). This introduces a problem of generality, as the theory is not precise enough to make predictions for novel inference patterns of arbitrary complexity. Second, some elements of the theory lack independent motivation (Koralus and Mascarenhas, 2013). In particular, mental model theory posits linguistically implausible interpretations for some connectives, as well as an operation for combining premises that is largely *ad hoc*.[34]

3.3.2 The interpretation-based account

We are now in a position to give a complete interpretation-based account of illusory inferences from disjunction. To make an in-depth presentation of the proposal possible, I begin with a slight variant of the original inference, with one fewer atomic proposition. This will make the total number of formal alternatives to consider more manageable. As we will see, even this simplification gives rise to a rather rich set of alternatives. The (presumed) fallacious inference pattern I discuss in detail in this section is the following.

(31)
$$\begin{array}{rl} P_1\text{:} & (a \wedge b) \vee c \\ P_2\text{:} & a \\ \text{Concl.:} & b \end{array}$$

The schema in (31) is still fallacious, and it seems to include everything from the original inference pattern that is relevant. Interestingly, the exact schema in P_1 of (31) (but not the more complex P_1

[34] For example, the conditional connective *if*, in mental model theory, is interpreted as a mental model of the conjunction of antecedent and consequent, along with a *mental model footnote* indicating that other alternative mental models exist that are not being attended to. Put perhaps more intuitively, a conditional "if a then b" is effectively interpreted as a disjunction "a and b, or else other alternatives not worth considering for the moment." From this interpretation, the fallacy of affirming the consequent follows along the same lines as the illusory inference from disjunction above: when hearing the second premise b, reasoners discard the mental model footnote (the second disjunct in my informal explanation of the interpretation of conditionals) and consider only the first model for the conditional, a model of a and of b. From here, the antecedent a follows immediately.

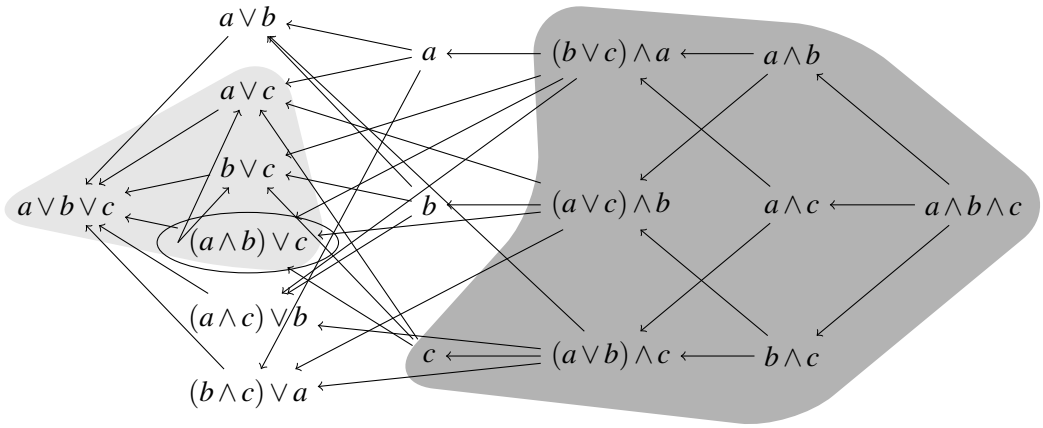


Figure 1: All formal alternatives, up to classical equivalences, for the source $(a \wedge b) \vee c$ (circled in the figure). Arrows between alternatives indicate entailment (transitivity and reflexivity are not represented). The lightly shaded alternatives on the left are weaker than or equivalent to $(a \wedge b) \vee c$. The darker alternatives on the right are strictly stronger than $(a \wedge b) \vee c$.

of (18)) was discussed by Spector (2007) and shown to have the same implicature I will derive. The proof I give in this section differs from Spector's in several respects, but these are largely more expository than substantive.[35]

According to the implicature-based theory given in section 3.2.2, we begin by calculating the enriched meanings of each premise. These are the interpretations that will feed into the reasoning component of the theory. Since P_2 is (for our purposes) atomic and contains no scalar items, I take it that its enrichment P_2^+ is identical to P_2. The interesting case is therefore P_1.

The first step to calculate the strengthened meaning of P_1 is to calculate its set of formal alternatives. It is given in Figure 1. Each expression in this set of alternatives is the result of a licensed substitution according to the adopted theory of formal alternatives. This set is complete, up to certain equivalences we need not worry about.[36]

[35] Spector's (2007) objectives were entirely unrelated to mine. He showed a very general correspondence result connecting neo-Gricean approaches to localist accounts of implicature. The few substantive differences between our proofs of the strengthened meaning of P_1 of (31) are due to the fact that Spector had to make some assumptions which I needn't make, given our different purposes.

[36] The more technically minded reader will be interested in an informal proof of the completeness of this set. Explaining the procedure I followed will hopefully suffice. First, I considered every substitution of the connectives in the original formula: $(a \wedge b) \vee c$, $(a \vee b) \wedge c$, $a \wedge b \wedge c$, $a \vee b \vee c$. The next step was to consider substitutions and deletions of the propositional atoms for each of these four alternatives. Substitutions of individual atoms for the last two (only $\wedge$s and only $\vee$s), given commutativity and idempotence, will yield formulas equivalent to deletions of atoms, so we can disregard these substitutions. Deletions from these formulas result in all simplex disjunctions and conjunctions possible with the set of atoms $\{a, b, c\}$, as well as each individual atom. All three possible conjunctions and all three possible disjunctions

Will be a primary implicature and also a sec. implic.	but not a sec. implic.,	because it would entail
$(a \vee b) \wedge c$	c	$(a \vee c) \wedge b$
$a \wedge c$	$(b \vee c) \wedge a$	c
$b \wedge c$	$(a \vee c) \wedge b$	c
$a \wedge b \wedge c$	$a \wedge b$	c

Table 2: Alternatives that will give rise to primary implicatures. The first and second columns contain all eight alternatives that give rise to primary implicatures. Alternatives in the first column will also give rise to secondary implicatures. Those in the second column will not, due to the entailment displayed in the third column.

Next, we calculate primary implicatures for those alternatives that are strictly stronger than P_1.[37] There are eight such alternatives, given in Table 2. Table 2 indicates already which of these alternatives will *also* give rise to secondary implicatures and, for the ones that will not, what the relevant alternatives are that will block their strengthening into secondary implicatures. The predicted primary implicatures are therefore propositions of the form "the speaker is not in a position to assert φ," for each φ in Table 2.

Now for secondary implicatures. For each of the eight distinct alternatives in Table 2, we ask whether we can negate that alternative without contradicting the literal meaning P_1 or any of the

are included in the set of alternatives, as well as all three individual atoms. Consider now the formulas $(a \wedge b) \vee c$ and $(a \vee b) \wedge c$. Deletions from these formulas will result in simplex conjunctions and disjunctions, all of which are already in our list. Substitutions are more interesting in this case. Any substitutions that use an atom more than once (e.g. $(a \wedge b) \vee a$) will be equivalent by absorption to an atom, so we can disregard them. We are therefore left with only two substitution variants for each of $(a \wedge b) \vee c$ and $(a \vee b) \wedge c$, corresponding to "reshufflings" of the propositional atoms. As the reader can verify, all four of these reshufflings are represented in Figure 1.

[37] We could have adopted the view that the relevant alternatives are not just the ones stronger than the literal meaning, but rather all alternatives not weaker than the literal meaning. The set of alternatives giving rise to primary implicatures becomes larger (including unshaded alternatives in Figure 1), therefore predicting a few additional primary implicatures. Interestingly, even with this larger set of primary implicatures, the secondary implicatures will be the same as with the smaller set of primary implicatures gotten by looking only at stronger alternatives. Consequently, my results and analysis of the illusory inference from disjunction are preserved under this alternative view of implicature. I give here a proof sketch. First we need to show that there are no new secondary implicatures. This amounts to proving that each of the negations of the new alternatives, when conjoined with the source of the alternatives, entails some other alternative in Figure 1. The new alternatives are $a \vee b$, $(a \wedge c) \vee b$, $(b \wedge c) \vee a$, a, and b. Each of these, when negated, entails $\neg a$ or $\neg b$, and therefore, conjoined with the source of the alternatives $(a \wedge b) \vee c$, each will entail c, a minimal alternative in Figure 1. Second, we must show that the secondary implicatures we had before will be preserved when we broaden the set of primary implicatures. By (32) and (33), this is equivalent to proving that every element of $\{\{(\neg a \wedge \neg b) \vee \neg c, (a \wedge b) \vee c, \varphi\} : \varphi \in \{a \vee b, (a \wedge c) \vee b, (b \wedge c) \vee a, c\}\}$ is a consistent set of formulas. For $\varphi = c$, the set of formulas is satisfied by a model making c true and both a and b false. For the other cases of φ, the corresponding sets of formulas are satisfied by a model making a and b true and c false.

Since the data discussed so far provide no basis on which to choose between these two theories of strengthening, I opted for the more traditional view of looking only at strictly stronger alternatives. The non-weaker view will be relevant to the discussion of quantified variants of the illusory inference, in section 3.3.3.

primary implicatures. Formally, for S the source of the alternatives and SA_S the set of alternatives stronger than S (the alternatives in Table 2), we collect all formulas $\varphi \in SA_S$ such that

$$(\forall \psi \in SA_S)\ \neg\varphi \wedge S \nvDash \psi\ .$$

If an alternative ψ is not entailed by a (potential) secondary implicature together with the literal meaning S, then alternatives ψ' that are at least as strong as ψ will also not be entailed. This reduces significantly the set of primary implicatures that we need to consider.

(32) For $\textsc{Bot}(\Phi)$ a function returning all elements of Φ that do not asymmetrically entail any other elements of Φ,[38] any $\varphi \in SA_S$ will give rise to a secondary implicature just in case

$$(\forall \psi \in \textsc{Bot}(SA_S))\ \neg\varphi \wedge S \nvDash \psi\ .$$

Moreover, notice that, if an alternative φ satisfies the condition in (32), so will any alternatives stronger than φ. This fact will also simplify matters: if we can show that a very weak alternative will give rise to a secondary implicature, we will also have shown that any alternatives stronger than it will give rise to secondary implicatures.

(33) $(\forall \psi \in \textsc{Bot}(SA_S))\ \neg\varphi \wedge S \nvDash \psi \Rightarrow (\forall \psi \in \textsc{Bot}(SA_S))\ \neg\varphi' \wedge S \nvDash \psi$, for $\varphi' \vDash \varphi$

We are now in a position to see that Table 2 draws the correct line between the alternatives that give rise to secondary implicatures and those that do not. First, alternative $(a \vee b) \wedge c$ satisfies the condition.[39] It suffices to show that its negation is consistent with the literal meaning $(a \wedge b) \vee c$ and each of the three weakest alternatives, namely c, $(a \vee c) \wedge b$, and $(b \vee c) \wedge a$. In (34), I list the

[38] Formally, for Φ a set of formulas, $\textsc{Bot}(\Phi) = \{\varphi \in \Phi : (\forall \varphi' \in \Phi)\ \varphi \rightarrow \varphi' \Rightarrow \varphi \leftrightarrow \varphi'\}$.

[39] Fox (2007, ft. 35) points out a case where an alternative is problematic that has been derived by two steps of substitution, one yielding a stronger formula and the other a weaker one. The crucial alternative for deriving the required implicature for the illusory inference, namely $(a \vee b) \wedge c$, is of this "zigzagging" kind. It is however very unclear whether Fox's concern carries over to the case discussed here. First, the problematic alternative discussed by Fox contains quantifiers, which are not present in the crucial alternative discussed here. Second, every problematic case of a zigzagging alternative that I am familiar with concerns an alternative non-weaker (not stronger) than the literal meaning. The crucial alternative for my discussion of the illusory inference is in fact stronger than the literal meaning. Finally, as far as I can see, including this alternative has no pernicious effects.

three sets of formulas that must be consistent for $(a \vee b) \wedge c$ to give rise to a secondary implicature, together with a model of each of those sets of formulas, proving their consistency.

(34) The following sets of formulas are consistent

$\{(\neg a \wedge \neg b) \vee \neg c, (a \wedge b) \vee c, c\}$	$\neg a, \neg b, c$
$\{(\neg a \wedge \neg b) \vee \neg c, (a \wedge b) \vee c, (a \vee c) \wedge b\}$	$a, b, \neg c$
$\{(\neg a \wedge \neg b) \vee \neg c, (a \wedge b) \vee c, (b \vee c) \wedge a\}$	$a, b, \neg c$

By (34) and (33), we establish that the alternatives in the first column of Table 2 will indeed give rise to secondary implicatures. We must now show that these are *all* the alternatives that will.

Consider c. From $\neg c$ and the literal meaning it would follow that $a \wedge b$ and therefore that $(a \vee c) \wedge b$, which is a member of $\textsc{Bot}(SA_S)$. Therefore c, by (32), will not give rise to a secondary implicature. A similar reasoning applies to each alternative in column two of Table 2: each of these alternatives, if strengthened, would entail the member of $\textsc{Bot}(SA_S)$ indicated in the third column. We are left with the following set of secondary implicatures, corresponding to the negations of the formulas in the first column of Table 2.

(35) $\{\neg((a \vee b) \wedge c), \neg(a \wedge c), \neg(b \wedge c), \neg(a \wedge b \wedge c)\}$

Among these, the first secondary implicature $\neg((a \vee b) \wedge c)$, equivalently $(\neg a \wedge \neg b) \vee \neg c$, entails the remaining three secondary implicatures, as the reader can easily check. We can therefore safely disregard the weaker secondary implicatures: in the presence of the stronger secondary implicature, they add nothing to the strengthened meaning of the premise P_1.

Finally, we calculate the strengthened meaning P_1^+ of P_1, by conjoining the literal meaning P_1 with the secondary implicature $(\neg a \wedge \neg b) \vee \neg c$:

$$((a \wedge b) \vee c) \wedge ((\neg a \wedge \neg b) \vee \neg c)\ .$$

By distributivity of the second conjunct into the first, this is equivalent to

$$(a \wedge b \wedge ((\neg a \wedge \neg b) \vee \neg c)) \vee (c \wedge ((\neg a \wedge \neg b) \vee \neg c)) \,,$$

which is in turn equivalent to the final strengthened interpretation P_1^+ in (36).

(36) $\quad (a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b)$

The reader can likely already see that (36) will do the required job. Under our simple assumptions about the reasoning component, the (putative) illusory inference in (31) will be judged valid if (37) is classically valid.

(37)

P_1^+:	$(a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b)$
P_2^+:	a
Concl.:	b

The pattern in (37) is classically valid.[40] As with the simpler case of affirming the consequent, briefly discussed in section 3.1, there is no fallacy to speak of.

This result extends to the original (unsimplified) illusory inference from disjunction, schematized in (38).

(38)

P_1:	$(a \wedge b) \vee (c \wedge d)$
P_2:	a
Concl.:	b

Simply let c in the demonstration above stand for $c \wedge d$ in P_1 of (38). That the inference in (38) is valid for the strengthened interpretation of its premises in (39) follows as a corollary of the discussion above.

(39)

P_1^+:	$(a \wedge b \wedge \neg(c \wedge d)) \vee (c \wedge d \wedge \neg a \wedge \neg b)$
P_2^+:	a

[40] Proof: the second disjunct of P_1^+ together with P_2^+ is a contradiction, so the first disjunct of P_1^+ must be true. Then b follows immediately.

In fact, something even stronger can be shown. The strengthened meaning of P_1 of (38) is in fact the stronger (40).

(40) P_1^+: $(a \wedge b \wedge \neg c \wedge \neg d) \vee (c \wedge d \wedge \neg a \wedge \neg b)$

Since this result is not strictly needed to derive the original illusory inference from disjunction, I radically abbreviate the proof of this claim.[41] The following two stronger alternatives will be included in the alternatives for P_1 of (38).

(41) a. $(a \vee b) \wedge (c \wedge d)$

b. $(a \wedge b) \wedge (c \vee d)$

Both (41a) and (41b) will turn out to satisfy the condition in (32), and will therefore give rise to secondary implicatures. When conjoined with the literal meaning P_1 of (38), the negations of (41a) and (41b) give a formula equivalent to (40).

3.3.3 Discussion and empirical predictions

The appeal of the implicature-based theory is undeniable. It turns out that a conservative theory of implicature, together with entirely classical assumptions about reasoning, predicts the illusory inferences from disjunction that have been taken to provide confirmation for mental model approaches (Walsh and Johnson-Laird, 2004). Insofar as the account extends to a larger subset of the data that reasoning-based theories account for, while maintaining a core notion of interpretive enrichment that is independently motivated with strictly linguistic arguments, the reasoning-based theories developed in psychology have a viable competitor, likely to be a more parsimonious alternative.

But meta-theoretical criteria such as parsimony are not the only tools at our disposal, when trying to decide between a reasoning-based account or an interpretation-based account of compelling fallacies. The interpretation-based account I give in this chapter makes a prediction not shared by reasoning-based accounts. If reasoners' acceptance of a certain fallacy hinges on the reasoning

[41] This example has one more atomic proposition than the simplified example considered above, and therefore the complete set of alternatives is appreciably larger and very hard to represent in a useful manner. This is because the output of Katzir's (2007) algorithm for generating alternatives increases exponentially as the number of syntactic nodes of the source increases. This property of the system is discussed in detail later in this chapter, in section 3.4.

module operating on the strengthened meaning of the premises, then acceptance rates should decrease significantly if we manipulate the context, syntactic or pragmatic, in ways that will reliably block the crucial implicatures.

Specifically, it is well known that a clause S will trigger different implicatures in upward entailing contexts than in downward entailing contexts. Indeed, the account of the illusory inference from disjunction given above relies on the crucial premise (P_1) being a matrix clause. If however this premise were presented in a downward entailing context, such as an *if*-clause, the required scalar implicature would not be predicted to arise (or at least it would be predicted to arise less often), and reasoners should be less likely to accept the inference as valid.

I propose a new experimental paradigm to test this prediction of the interpretation-based account. The trick is to convert standard reasoning problems (42) into a conditional form (43).[42]

(42) Standard reasoning problem:

$P_1, \ldots, P_n$

Does C follow from the above premises?

(43) Conditional format:

a. If P_1, then if $\ldots$, then if P_n, then C.

Is the above sentence true?

b. Whenever P_1 and $\ldots$ and P_n, C.

Is the above sentence true?

The conditional schemata in (43) are of course much more syntactically complex than the reasoning problem format in (42). This introduces important concerns about parsing difficulties and ambiguity. For example, naively translating the illusory inference from disjunction stimuli used by Walsh and Johnson-Laird (2004) into conditional format results in ambiguous sentences that will almost certainly be extremely difficult, if not impossible, for subjects to parse in an experiment:

[42]Other reasoning experiments ask "what, if anything, follows" from a sequence of premises, instead of presenting a putative conclusion C and asking whether C follows. This paradigm can also be converted into conditional form, by giving subjects an incomplete conditional sentence and asking them to complete the *then*-clause in a way that makes the entire sentence true.

(44) If either Jane is kneeling by the fire and she is looking at the window or otherwise Mark is standing at the window and he is peering into the garden, then, if Jane is kneeling by the fire, she must be looking at the window.

Happily, this problem can be significantly mitigated if we use syntactically simpler variants of this illusory inference. First, I point out that the illusory inference pattern can be recast with universal quantifiers doing the job of conjunction.

Consider (45), given in standard reasoning problem format.

(45) P_1: Every boy or every girl is coming to the party.
P_2: John is coming to the party.
Concl.: Bill is coming to the party.

I find this a very compelling inference, and I suspect the reader will agree. It is, however, fallacious: in a model where every girl and John come to the party, and no one else does, the premises are satisfied but the conclusion falsified.

Notice that, in the first premise of (45), the universal quantifiers are equivalent to arbitrarily large conjunctions, assuming that the domain of boys and girls is finite. In a model where b_1 and b_2 are the boys and g_1 and g_2 are the girls, the formulas in (46a) and (46b) are equivalent, and therefore equally apt representations of the literal meaning of P_1 in (45).

(46) a. $((\forall x \in boy')\ party'(x)) \vee ((\forall x \in girl')\ party'(x))$
b. $(party'(b_1) \wedge party'(b_2)) \vee (party'(g_1) \wedge party'(g_2))$

The formula in (46b) looks exactly like P_1 of the illusory inference from disjunction, and similarly for the straightforward representations of P_2 and the proposed conclusion of (45). It is therefore reasonable to call (45) a *quantified variant* of the illusory inference from disjunction.

Moreover, with one modification and one assumption, the theory of implicature I adopt here predicts that (45) should indeed be considered valid by reasoners, since the predicted strengthened meaning for P_1 in (45) will classically validate the inference, in a manner entirely parallel to the propositional illusory inference from disjunction. The required modification is that we consider,

rather than only alternatives *strictly stronger* than the literal meaning, alternatives that are simply *not weaker* than the literal meaning, as proposed by Spector (2007).[43] With this expanded set of alternatives, we will get (47b) as an alternative to (47a).

(47) a. Source of the alternatives:
Every boy or every girl is coming to the party.

b. Substituting existential for universal quantifiers and *and* for *or*:
Some boy and some girl are coming to the party.

Strengthening the alternative in (47b) into a secondary implicature, we get the formula in (48).

(48) $((\neg\exists x \in boy')\ party'(x)) \vee ((\neg\exists x \in girl')\ party'(x))$

From the conjunction of the secondary implicature in (48) with the literal meaning of P_1 in (45), we get, by distributivity of conjunction over disjunction,

$$\begin{aligned}(((\forall x \in boy')\ party'(x)) \wedge (((\neg\exists x \in boy')\ party'(x)) \vee ((\neg\exists x \in girl')\ party'(x))))\\ \vee (((\forall x \in girl')\ party'(x)) \wedge (((\neg\exists x \in boy')\ party'(x)) \vee ((\neg\exists x \in girl')\ party'(x)))) \ .\end{aligned}$$

Now the required assumption: we must assume that universal quantifiers have existential import, that is, their restrictors are never empty.[44] Consider the first disjunct above. If boys exist, then that disjunct is equivalent to

$$((\forall x \in boy')\ party'(x)) \wedge ((\neg\exists x \in girl')\ party'(x)) \ ,$$

[43] See footnote 37 on page 70, where I show that the results for the propositional illusory inference from disjunction are preserved under this modification of the theory. For a concise review of the independent arguments in favor of this modification, see also Schlenker (2012).

[44] It is also possible to shift this assumption from interpretation to the domain of reasoning. Without existential import, we get (something that entails) the slightly weaker strengthening in (i).

(i) If boys and girls exist, then every boy and no girl is coming to the party, or else every girl and no boy is.

If reasoners assume that boys and girls exist on grounds entirely independent from the interpretation of the linguistic signal, as they plausibly might, then the "fallacy" can be derived in a classical logic from this interpretation of P_1.

and similarly for the predicate $girl'$ and the second disjunct. This means that the conjunction of (48) with P_1 in (45) is a formula equivalent to (49a). In (49b) I give an English sentence with the interpretation in (49a), to help parse that formula.

(49) a. $(((\forall x \in boy')\ party'(x)) \wedge ((\neg\exists x \in girl')\ party'(x))) \vee (((\forall x \in girl')\ party'(x)) \wedge ((\neg\exists x \in boy')\ party'(x)))$

b. Every boy and no girl or every girl and no boy is coming to the party.

The reader can easily verify that, with the interpretation in (49a) for P_1 of (45), the proposed inference follows classically, by the same reasoning applied to the original illusory inference from disjunction.

Granting for the purpose of this discussion that (45) is a compelling fallacy, it provides a promising way to test the predictions of the interpretation-based account in this chapter. While (44) was ambiguous and very difficult to parse, (50a) and (50b), the conditional formulations of the reasoning problem in (45), are perfectly tractable.

(50) a. If every boy or every girl is coming to the party, then, if John is coming to the party, Bill will also come.

b. Whenever every boy or every girl is coming to the party, and John is coming to the party, Bill also comes.

Recall the prediction of the interpretation-based theory. Since, in (50), the crucial premise of (45) is in the antecedent of a conditional (*if*-clause or *whenever*-clause), the strengthening required to validate the inference is either not predicted to arise or it is predicted to arise less often. Consequently, there should be a significant decrease in acceptability between the standard reasoning problem format in (45) and either of the conditional formats in (50). This provides a general way to falsify any implicature-based account of compelling fallacies.

3.3.4 Conclusion

This chapter has so far defined a program for the study of failures of reasoning that roots compelling fallacies in interpretive processes, rather than in the general-purpose reasoning mechanisms themselves. I have shown that this program can be applied to a class of sophisticated reasoning data from the psychological literature, thus far ignored by the field of formal pragmatics, yielding a natural account that uses only independently motivated interpretive mechanisms. Finally, I showed how in principle we can test the predictions of interpretation-based theories in this spirit, helping to demarcate the line between reasoning and interpretation.

This program and the result in the foregoing sections are of significance to the psychological study of reasoning. Since most scholars of human reasoning do not have a background in linguistics and most linguists do not work on reasoning, extant theories of reasoning tend not to take advantage of the sophisticated theories of meaning that semanticists have developed over the past forty years. Consequently, the difference between general-purpose reasoning and interpretive processes is not well understood. The working hypothesis of this chapter, that human reasoning is entirely classical, is almost certainly false, at least in this strong formulation. However, most psychologists would agree that understanding how human reasoning differs from normative logic is an important step toward understanding human reasoning. Clearly, we can only trust our accounts of this intermediate step if we can also trust our understanding of the line between reasoning and interpretation. Without that, the scientist himself might be falling prey to illusions of human irrationality.

3.4 On the cardinalities of sets of scalar alternatives

In this section, I address an issue with modern theories of scalar implicature not discussed in the literature so far. Insofar as a theory of implicature requires a rich enough set of formal alternatives, as most modern theories of pragmatics in linguistics do, a problem of tractability arises. As it turns out, the two most precise accounts of the alternative-generating mechanism produce sets of formal alternatives whose cardinalities increase exponentially as a function of the number of atomic

propositions in the source.[45]

3.4.1 Introduction

Since the work of Paul Grice (1967, published in 1989), scalar implicatures as exemplified in (51) have been taken to be the result of processes of pragmatic inference. In a nutshell: starting from what the speaker of (51a) *literally said*, the hearer will consider what stronger statements the speaker *chose not to utter*, and conclude that the speaker believes all such statements to be false, as long as their falsehood is compatible with what she literally said.

(51) a. I saw John or Mary.

b. SCALAR IMPLICATURE: The speaker did not see both John and Mary.

I show that there is one technical component, universally employed in formal theories of scalar implicature, that has at least puzzling consequences under the assumption that these theories ought to be psychologically tenable. The issue can be summarized as follows. Every modern theory of scalar implicature makes crucial use of a set of scalar alternatives, sentences that are (in a precise sense) related to the sentence uttered by a speaker. These alternatives are the ones that a hearer will take into consideration when thinking of what the speaker could have said but chose not to. I point out in this section that the cardinalities of these sets increase at very fast rates, and moreover that, even for sentences with a relatively small number of coordinated clauses, the cardinalities of alternative-sets are very large numbers. If the theories of alternatives considered here are making claims about psychological processes, then these claims are very difficult to square with what we independently know about reasoning with alternatives. On the other hand, if these theories are to be taken only as mathematical idealizations or theories of pragmatic competence, then it becomes necessary to investigate what psychologically tenable heuristics might implement this competence. It does not follow from the facts I report here that modern formal approaches to scalar implicature are doomed. Instead, I aim to point out a collection of puzzling and previously unnoticed facts about formal pragmatics.

[45] The work presented in this section is a minimally modified version of a paper currently under revision.

3.4.2 Alternative-sets and their cardinalities

I take the work of Chierchia (2004), Sauerland (2004), Spector (2007), and Fox (2007) to be representative examples of contemporary formal approaches to scalar implicature. All of these theories assume or propose some mechanism that generates formal alternatives as a function of a source and possibly a context. I restrict my attention to explicit proposals about the set of formal alternatives, for obvious reasons. There are two approaches in the literature that fit this criterion.

3.4.2.1 Positive propositions

While making no commitments about the inner workings of the alternative-generating procedure, Spector's (2007) theory makes use of the set of positive propositions that can be constructed from a source S. The notion is simple and elegant, and is perhaps most naturally seen as a semantic concept: given a source S with a set A of atomic propositions, the set $P^+(A)$ of positive propositions based on A is the closure of A under conjunction and disjunction. Thus, for a simple source $S = a \vee b$, we get the alternative set

$$P^+(\{a,b\}) = \{a, b, a \wedge b, a \vee b\} \ .$$

The set of positive propositions based on a set of atoms A corresponds to the set of non-constant monotonic Boolean functions of $|A|$-many variables. Monotonic Boolean functions on n variables are precisely those that can be defined with the Boolean meet and join. The "non-constant" proviso simply excludes the *verum* and *falsum* functions from the list.

The number of monotonic Boolean functions on n variables $M(n)$ was defined by Dedekind in 1897 (see for example Kleitman, 1969). Accordingly, the set of all $M(n)$ for $n \in \mathbb{N}$ is known as the set of *Dedekind numbers*. The sequence grows very rapidly. As of mid 2013, no closed-form expression for $M(n)$ is known, and exact values for $M(n)$ are only known for $n \leq 8$. The set of positive propositions with n atoms inherits the mathematical unwieldiness of Dedekind numbers, since, for A a set of atoms with $|A| = n$, $|P^+(A)| = M(n) - 2$.[46]

[46] We subtract 2 because positive propositions are monotonic Boolean functions excluding *verum* and *falsum*.

3.4.2.2 The syntactic substitution approach

Katzir (2007) and Fox and Katzir (2011) propose a different mechanism for generating formal alternatives based on a notion of structural complexity. As explained by Katzir (2007), there are contexts where, in order to derive the observed scalar implicatures, a non-monotonic expression must be compared with a formal alternative that is structurally simpler than it without being its scale-mate. For example:

(52) a. I doubt that exactly three semanticists will sit in the audience.

b. I doubt that three semanticists will sit in the audience.

An utterance of (52a) suggests that the speaker does not find it unlikely that at least three semanticists will be in attendance. This means that (52b) must be a formal alternative to (52a). But the quantifier in (52b) is not a scale-mate of the one in (52a). The relevant relation between the two sentences seems to be one of complexity: (52b) is less complex than (52a). In the interest of space, I must refer the reader to the cited articles for a discussion of the merits of this approach.

Katzir's (2007) theory is given in a fully explicit way. First we need to define a substitution source. This will contain all elements that can legally substitute elements of the speaker's utterance when generating alternatives. The rationale is that anything in the substitution source is no more complex than the original utterance.

(53) Substitution source:

Let σ be a parse tree. The substitution source for σ is the union of the lexicon of the language with the set of all subtrees of σ.

Next, we define a relation of structural complexity, defined for syntactic structures. Finally, we define the set of structural alternatives in terms of structural complexity.

(54) Structural complexity:

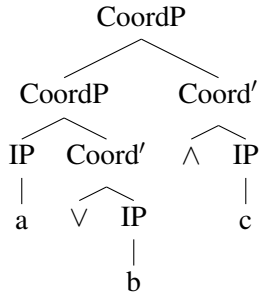
Let σ, σ' be parse trees. If we can transform σ into σ' by a finite series of deletions and replacements of constituents in σ with constituents of the same category taken from the substitution source of σ, we say that σ' is at most as complex as σ, in symbols $\sigma' \lesssim \sigma$.

(55) Structural alternatives:

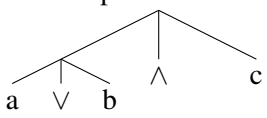
Let σ be a parse tree. The set of structural alternatives for σ is defined as $A(\sigma) := \{\sigma' : \sigma' \lesssim \sigma\}$.

In order to compare this theory directly with Spector's (2007) positive propositions, we need to make several simplifying assumptions. First, I assume that the parse trees that constitute the input to the procedure defined in (55) only have sentence level nodes, call them IPs, and coordination phrases, as in (56a). Given this very restrictive assumption, we can simplify these structures as in (56b). I further assume that the only elements available in the substitution source 53 are subtrees of the original utterance and binary connectives. Connectives have a very restricted distribution, occurring always and only between any two IP nodes. It will be useful to consider the lexical skeleton of a parse tree, that is the simplified parse-tree stripped of binary connectives, as in (56c). Finally, I assume that all parse-trees are left branching.

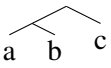
(56) a. Parse tree for $(a \vee b) \wedge c$



b. Simplification



c. Lexical skeleton



With these assumptions, we can find a simple closed-form expression that takes a natural number for the unique IPs in a tree and returns the number of alternatives predicted by Katzir's (2007) algorithm so constrained. The expression is the following, where n is the number of unique sentence-level nodes (IPs) in the source structure.

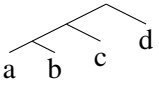
(3.1) $$\sum_{k<n} (2n-1)^{k+1} 2^k - k$$

I explain how we get to (3.1) shortly. First, it is important to remark that, whatever function (3.1) is, it constitutes a rather cautious lower bound on the number of alternatives generated by the uncon-

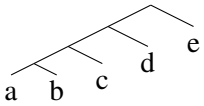
strained procedure. This is because the constraints used here exclude trees with both left and right branching, which allow for more complex structures than these simple left-branching trees. As we shall see, the results are already quite interesting, even for this lower bound.

What is the function in (3.1) doing? The expression $(2n-1)$ gives us the number of subconstituents of the lexical skeleton of a parse-tree with n distinct propositional atoms as terminal nodes, given the assumptions above.[47] Note that a lexical skeleton (so constrained) will have as many levels of embedding as there are propositional atoms. The sum in (3.1) is meant to range over these levels of embedding. The reader is encouraged to consider the lexical skeletons in (57) and convince him or herself that these mathematical expressions are correct given our assumptions.

(57) a. Lexical skeleton for four propositional atoms



b. for five atoms



The sum in function (3.1) starts at the level of zero embedding. How many single-node trees can we create given the substitution source? This will be simply the number of subconstituents of the original tree. For $k = 0$, that is what (3.1) returns. How many two-node trees can we create? It will be the number of subconstituents squared, times the number of connectives in our substitution source, for remember that we are considering lexical skeletons, so that a two-node tree is omitting the connective between its two nodes. The expression in (3.1), for $k = 1$, returns $(2n-1)^2 \cdot 2 - 1$. Why "minus one?" Because at embedding-level 1 we generate a tree identical to one already generated for embedding-level 0, namely the original tree. Similar considerations apply at further levels of embedding, as summarized in (58).

(58) In general, at embedding-level k, we will find

a. $k+1$ positions for subconstituents of the source — whence $(2n-1)^{k+1}$

b. k positions for connectives — whence 2^k

c. k trees that we have already counted in an earlier step — whence subtraction of k.

[47] Here is one way to see that this is the case. Each atom is a constituent on its own, giving us n. Consider now the tree from the top. Each non-terminal constituent contains one terminal constituent and one non-terminal constituent, except for the bottom non-terminal constituent in the tree, which is formed by two terminal nodes. This gives us $n-1$.

	2	3	4	n
1. Propositions	16	256	65,536	$2^{(2^n)}$
2. Positive propositions	4	18	166	Dedekind numbers: $M(n)-2$
3. Katzir (2007)	20	552	20,679	$\sum_{k<n}(2n-1)^{k+1}2^k-k$

Table 3: Number of alternatives by procedure, for a source with 2, 3, 4, and n atoms.

Finally, why add up the possible trees that can be created from the substitution source for each level of embedding in the original tree? The idea is to replicate the effects of an operation of node deletion, as used in Katzir's (2007) theory. This way, we need only consider instances of substitution, which greatly simplifies the formulation of the algorithm.[48] Notice that this was only possible given our simplifying assumptions about structure.

3.4.3 Discussion

Table 3 summarizes the findings in the preceding sections. It includes the total number of propositions made out of n atoms merely to provide an anchor that will be familiar to the reader.

The first interesting observation about Table 3 is that, for both algorithms, the size of the output set of alternatives increases very fast as a function of the number of atoms in the input. This is visible even for small numbers in Table 3 in the case of Katzir's (2007) theory. Indeed, in this theory the number of predicted alternatives grows (at least) at an exponential rate.[49] Although it is not obvious from simply looking at Table 3, the positive-propositions approach fares no better. A coarse asymptotic approximation places the lower bound on Dedekind numbers at $C(n,n/2)$, where C is the binomial coefficient. This lower bound itself grows at an exponential rate.[50] These facts immediately suggest that the question of the complexity of the problem of generating formal alternatives is of interest. While this dissertation does not contain a proper complexity result, I return to this issue at the end of this section. First, I discuss the absolute output values for small

[48] Another way to replicate the effects of deletion would be to put the empty tree, call it ε, in the substitution source. The problem with this strategy is that it makes it harder to eliminate structurally identical alternatives, which ought not to be a part of the final alternative-set. For example, a simple tree with just two atoms a b would generate the alternative a in two distinct ways: ε a and a ε. The strategy adopted here allows for an accurate closed-form expression that is simpler to formulate and explain.

[49] The closed-form expression above is a sum of exponential functions. It is also easy to show that it is in $O(n^n)$.

[50] The binomial coefficient $C(n,n/2)$ is in fact in $\Theta(2^n n^{-0.5})$.

numbers, as can be glanced from Table 3, as well as the differences between the two theories of formal alternatives.

3.4.3.1 On the values for small numbers

The values seen in Table 3 for small numbers are already rather large. This too is less obvious for the positive-propositions approach, which seems manageable even for $n = 4$ when compared to the syntactic approach. However, at $n = 7$, Dedekind numbers overtake the syntactic alternatives, and both are astronomical.[51] I submit that these absolute values are interesting in their own right, irrespective of considerations of relative growth.

It is a matter of debate in the literature on scalar implicature whether what are known as secondary implicatures are generated by a global process of pragmatic reasoning (Sauerland, 2004) or by a narrowly grammatical process, involving unpronounced exhaustification operators (Chierchia, 2004; Fox, 2007). However, *primary* implicatures, given their weaker epistemic nature, are usually taken to be the result of proper pragmatic reasoning. For the case of grammatically generated scalar implicatures, these kinds of large numbers of alternatives are arguably of little significance, for similarly large numbers of alternatives already figure in principle in current theories of questions or of alternative-sensitive items like "only." But this move is harder to argue for, or altogether untenable, in the case of scalar implicatures generated by pragmatic reasoning rather than grammatical mechanisms. If pragmatic reasoning is a certain kind of *reasoning*, then we expect the absolute numbers of alternatives predicted to be taken into consideration to fit into what we independently know about reasoning with sets of alternative propositions.

The psychological literature on reasoning offers theories of reasoning with and about alternative propositions. In particular, mental model theory, as proposed by Johnson-Laird (1983) and collaborators, is an account of the human faculty of reasoning whose hallmark is precisely a theory of how humans entertain and manipulate alternative propositions, called alternative mental models. Now, the role of alternative mental models in reasoning is not the same as that of formal alternatives in

[51] To wit, the seventh Dedekind number (subtracting 2 as explained above) is greater than two trillion ($2,414,682,040,996$). The syntactic approach generates "only" around four billion alternatives ($4,176,541,270$) for the same atoms.

pragmatic reasoning. Alternative mental models represent ways that the world could be like, according to the information being attended to by a reasoner, while formal alternatives represent utterances that a speaker could have made but chose not to. However, the kinds of manipulations of alternative mental models proposed by mental model theory are not unlike the manipulations required by theories of implicature in formal pragmatics. For example, mental model theory provides mechanisms for checking the consequences (entailments) of mental models and for excluding particular mental models from the set under consideration.

The connection I am drawing between sets of alternatives in scalar implicatures and alternative mental models is by no means perfect. But, as far as I can see, there is nothing in the reasoning literature that looks *as much* like the alternatives in formal pragmatics as the alternative mental models of mental model theory. It is therefore at least interesting to ask whether the formal alternatives in theories of scalar implicature fit into the mental models picture of reasoning with and about alternative mental models.

The short answer is no. In mental model theory, important correlations are established between the number of alternative mental models under consideration and humans' performance in reasoning tasks. Every study within this paradigm indicates that humans can efficiently reason with/about rather small numbers of alternative mental models at any given point in time, with five to seven concurrent mental models being the limit.[52] These results cannot be squared with the astronomical numbers of formal alternatives predicted by the theories considered in this chapter.

This observation is of some significance. If pragmatic reasoning makes use of alternatives as numerous as predicted by current theories, then it bears no relation to other known kinds of reasoning with and about alternatives. Consequently, either pragmatic reasoning is insensitive to these astronomical numbers of alternatives, or our theories of formal alternatives are merely elements of theories of pragmatic competence, offering no principled insight into the way in which the mind computes and manipulates alternatives in pragmatic processes. Notice that both possibilities pose interesting puzzles for linguistics and psychology that deserve further investigation.

[52]For a recent discussion of some of these results, see Johnson-Laird (2008). For correlations between the number of mental models under consideration and reaction times for very small numbers (one vs. two models), see for example Walsh and Johnson-Laird (2004).

1. If we take these theories of formal alternatives to be making claims about mental processes, we must conclude that pragmatic reasoning looks nothing like general-purpose reasoning. This is because we do not see in pragmatic reasoning the same kinds of constraints on numbers of alternatives that we see in general-purpose reasoning.

2. If instead we do *not* consider these theories of formal alternatives to provide an insight into actual mental processes, then there is an important gap in our understanding of pragmatic reasoning. Namely, which clever heuristics does the mind use, in the process of computing scalar implicatures, that allow it to only consider a manageable number of formal alternatives?

It is important to note that 2. may have an answer. Despite the fact that current theories of formal alternatives generate astronomical numbers of alternatives, I am familiar with no results that prove (or even argue) that these full sets of alternatives are *required* to account for the observed scalar implicatures. The syntactic approach for one will generally produce non-negligible amounts of syntactically distinct but equivalent alternatives. For example, a source like $a \vee b \vee c$ will generate the alternatives a, $a \wedge a$, and $a \wedge a \wedge a$. Clearly, an account of scalar implicature ought to be able to make do with only one of these. The positive propositions approach, due to its semantic nature, does not include equivalent-but-distinct alternatives, but it still produces alternatives that are entirely idle in the computation of scalar implicatures.[53] In sum, these theories of formal alternatives may well be adequate and efficacious in the context of a theory of pragmatic competence, but there is no a priori reason to think that dramatically more frugal heuristics cannot be discovered that also do the job.

3.4.3.2 Comparing the two alternative-generating procedures

Katzir's (2007) substitution approach generates more alternatives than there are propositions for very small numbers. This is due to the syntactic nature of the algorithm. The procedure generates many collections of equivalent formulas such as a and $a \wedge a$, even in the case of only two atomic

[53] Recall how earlier in this chapter, I showed that the syntactic approach of Katzir (2007) can generate all of the required implicatures for a sentence of the shape $(a \wedge b) \vee c$. But the syntactic approach does not generate one of the alternatives predicted by the positive propositions approach, showing that that theory also overgenerates. I discuss this result in more detail in the next section.

propositions. This prompts the question of how many *equivalence classes* of propositions are generated by this syntactic approach.

For $n = 2$, the equivalence classes of Katzir's (2007) alternatives, given the assumptions made explicit above, are the same as positive propositions. However, for $n = 3$, we get 17 equivalence classes of formal alternatives, versus 18 for Spector's (2007) positive propositions approach. This suggests that Katzir's theory, after suitable modifications that prevent equivalent alternatives from being added to the list, might be more economical than the positive propositions approach.

It is useful to see a concrete example. I focus on secondary implicatures, since as far as I can see the pragmatic reasoning / grammar distinction is immaterial in this respect. Sentences as in (59a), first discussed in the pragmatics literature by Spector (2007), have the implicature in (59b).

(59) a. Either John and Mary or Bill will come to the party. $(a \wedge b) \vee c$

b. Neither John nor Mary will come, or else Bill won't come. $(\neg a \wedge \neg b) \vee \neg c$

The alternative present in the positive propositions approach but not in Katzir's (2007) theory is $(a \wedge b) \vee (a \wedge c) \vee (b \wedge c)$; in words "at least two of a, b, and c." First, I point out that this alternative is not required to derive the observed implicature in (59b). In fact, as (59b) suggests, we actually need only one alternative to derive the implicature, namely $(a \vee b) \wedge c$. Second, this alternative does not give rise to an observed secondary implicature. The implicature would be that only one of a, b, and c is true, which is manifestly not an implicature of (59a).[54]

These observations allow for a weak but interesting conclusion: Katzir's (2007) theory, when restricted to equivalence classes and under the simplifying assumptions made above, contains fewer unnecessary propositions than the positive propositions approach, for the case of $n = 3$. A general result must be left to future work, and will in fact not be easy to achieve. For the simplifying assumptions made in the foregoing sections about syntactic structure make a crucial difference for *which*, though not significantly *how many*, alternatives are generated in Katzir's (2007) theory.

[54] Nor is it predicted to be a an implicature of (59a) under the theories of scalar implicature I consider here. For Sauerland (2004), it is not innocently excludable, while for Spector (2007) (59a) is not an optimal answer in an information state that entails the negation of $(a \wedge b) \vee (a \wedge c) \vee (b \wedge c)$. I refer the reader to both papers for the details.

3.4.3.3 Implications for the complexity of pragmatic theories

The numbers and functions in Table 3 do *not* directly allow us to find complexity measures for the alternative-generating algorithms considered here, or in general for the problem of generating formal alternatives. Instead, they relate a measure of structural complexity of the input (the number of atomic formulas that occur in the input) with the cardinality of the output set (the set of alternatives generated). This is not a direct measure of the complexity of the alternative-generating algorithm itself. For this reason, it was enough to take high-level mathematical descriptions of the two alternative-generating algorithms discussed here, as given by both Spector (2007) and Katzir (2007). However, we must distinguish the complexity of the problem of generating alternatives, to which the work in this chapter does not provide an answer, from the complexity of the problem of calculating scalar-implicatures. In fact, the numbers on Table 3 will be crucial when investigating the *effective* complexity of calculating scalar implicatures, irrespective of actual implementations of the alternative-generating algorithms.

Any algorithmic account of scalar implicature must define a function that takes as one of its input arguments a set of formal alternatives. The scalar-implicature algorithm will then have to scan this list of formal alternatives, checking each of them for certain properties. Say the scalar-implicature algorithm itself takes merely linear time on all of its input parameters taken together, including the set of formal alternatives. Even in this case, the rate of growth of the search space will matter to the effective time the scalar-implicature algorithm will take to calculate implicatures. In other words, the rate at which sets of formal alternatives grow as a function of the source of alternatives will have a direct impact on the complexity of the broader scalar-implicature algorithm, no matter how efficient its inner computations are. In this sense, the results in this chapter offer a first glance at the effective complexity of formal theories of pragmatics that make use of these alternative-generating algorithms.

Taken as measures of search-space growth, the results in Table 3 predict that calculating implicatures is computationally intractable. As observed above, both the syntactic algorithm of Katzir (2007) and the positive propositions approach grow at least at an exponential rate. Consequently, either calculating scalar implicatures is in fact an intractable problem in the general case, or the mind

is using different alternative-generating procedures that produce more manageable sets of formal alternatives.

Is it impossible that the human mind is using alternative-generating procedures that take exponential time? Unfortunately, it is very difficult to find straightforward predictions based on the assumption that the mind is working with exponentially growing alternatives as seen above. One tempting route would be to look for an exponential increase in processing time of scalar implicatures as a function of the number of atoms in the input. The difficulty with this strategy is that the mere fact that the output of alternative-generating procedures increases exponentially does not tell us *where* we would be able to see the effects of this exponential increase. All we predict is that *there are* values for the input where the processing load would increase exponentially. In other words, the human mind could be such an efficient computer that we would only see the effects of these exponentially growing sets for large numbers of atoms in the input. These numbers might well be large enough to make it impossible in practice to find processing-time results, given entirely independent restrictions on the length of sentences that humans are capable of parsing.

To summarize, looking at the rate of growth of the outputs of alternative-generating procedures gives us a preliminary insight into the effective complexity of theories of scalar implicature that use them. However, rate of growth does not offer a direct way to check for the psychological plausibility of these theories.

3.4.4 Conclusion

Modern theories of scalar implicature make crucial use of sets of alternatives. In this section I showed how the two most precisely defined alternative-generating procedures in the literature (a) generate very large sets even for the case of small inputs and (b) generate sets whose size increases (at least) exponentially as a function of the input. These facts, while by no means damning to current work in formal pragmatics, highlight a puzzling state of affairs that deserves investigation. If these theories of alternative-generation should have psychological import, then one is at a loss trying to integrate them within the broader existing research on cognitive mechanisms that deal with reasoning and with alternatives. If on the other hand they are to be taken only as theories of

pragmatic competence, then we must ask what cognitive mechanisms implement this description of competence; what heuristics does the mind use that allow it to consider only a manageable subset of these large collections of alternatives, while deriving the observed implicatures.

Chapter 4

Reasoning about probabilities — the conjunction fallacy

4.1 Introduction

The conjunction fallacy is one of the most well-known puzzles about human reasoning. Consider the following setup and question.

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Which is more probable?
>
> 1. Linda is a bank teller.
> 2. Linda is a bank teller and is active in the feminist movement.

About 85% of subjects rank the conjunctive option 2. above the simpler option 1. This is a fallacy according to standard probability theory, for the probability of a conjunction of independent propositions cannot be higher than that of one of its conjuncts. Since Tversky and Kahneman reported these results in a seminal (1983) paper, the conjunction fallacy has generally been taken

to show that human reasoners, including a majority of highly educated ones,[55] make choices that directly violate simple theorems of probability theory.

This chapter is divided in two parts. In section 4.2 I give an account of the conjunction fallacy that explains it with recourse to two linguistic phenomena and a psychological one. First, I argue that option 1. in the problem given above is interpreted roughly as "Linda is a bank teller and she may or may not be active in the feminist movement." I derive this interpretation from the predicted primary implicatures of the original sentence.[56] Second, I show that, in Kratzer's (1991) theory of modality in terms of comparative probabilities, "Linda is active in the feminist movement" is a better possibility than "Linda might not be active in the feminist movement," under the assumption that "Linda is active in the feminist movement" is a better possibility than "Linda is not active in the feminist movement." Finally, I argue that subjects likely make this assumption on independent grounds, as it is an instance of what Tversky and Kahneman (1974) call base-rate neglect. Base-rate neglect is the phenomenon of human reasoning whereby people fail to take into account the prior probabilities of events, overestimating the usefulness of particularly salient chunks of information when calculating priors. In this specific case, subjects take the probability that Linda is active in the feminist movement to be higher than 50%, while most likely only a rather small percentage of women are in fact active in the feminist movement.

Section 4.3 contains the second, more tentative part of this chapter. In it I take a broader look at the representativeness heuristic, Kahneman and Tversky's original explanation of the conjunction fallacy, and by and large still the dominant theory of the phenomenon. I raise novel questions about how a representativeness account is to be applied to complex sentences, an essential but rarely discussed component of the conjunction fallacy explanation in that theory. I then explore the connection between representativeness judgments and likelihood judgments, searching for cases where the two differ.

[55] Tversky and Kahneman (1983) tested undergraduates with no background in probability or statistics, first year graduate students, and doctoral students with considerable backgrounds in statistics and decision theory. They found no effect of statistical sophistication in their indirect conjunction fallacy experiments. For the transparent case, where subjects are given exactly the two options I give above, more statistically sophisticated subjects did perform appreciably better than more naive ones: 36% committed the conjunction fallacy.

[56] See the discussion in section 3.2.2, p. 61, for a review of the kind of pragmatic theory I assume.

I conclude this introduction with a brief review of essential background on the conjunction fallacy, representativeness, and interpretation-based accounts related to mine.

Background on the relevant experiments Tversky and Kahneman (1983) demonstrated the conjunction fallacy in three main classes of experimental situations. The version of the Linda problem just given is a *direct-transparent*, or simply transparent, test. In it, subjects are shown both and only the conjunctive option and its least intuitively likely conjunct. By contrast, in the *direct-subtle* test, subjects are given both options together with a number of fillers. Finally, in *indirect* tests, one group of subjects is given the conjunctive option and a number of fillers, while another group sees the conjuncts together with fillers. In the order I just gave them, the three tests gradually dilute the comparative element of the task, to the point where, in indirect tests, there seems to be no opportunity for comparison between the conjunction and its conjuncts (but see the discussion on page 98 and footnote 59 for an argument that there might be). Rates of commission of the conjunction fallacy are comparable across all three classes of experiments.

Representativeness The authors explain the conjunction fallacy in terms of their representativeness heuristic. Representativeness is "an assessment of the degree of correspondence between a sample and a population, an instance and a category, and act and an actor" (Tversky and Kahneman, 1983). In the cases I look at in this chapter, representativeness effectively reduces to prototypicality. Although the notion has been criticized for being too vague and suspiciously all-explaining (Gigerenzer, 1996), it has more recently been made precise in terms of prototype theory (Kahneman and Frederick, 2002) and exemplar theory (Nilson et al., 2008).

The representativeness explanation of the conjunction fallacy goes as follows. Humans substitute for a difficult question about probabilities an easier question about representativeness. Instead of assessing the likelihood of certain statements in the Linda experiment, subjects are answering the question "how typical is Linda, given this description of her personality, as an example of this category of people?" The conjunctive option ranks higher on representativeness than the conjunct "Linda is a bank teller;" this is the case on the formalized versions of representativeness cited above, but it can be also be seen by directly asking subjects questions about typicality. Thus, the represen-

tativeness heuristic predicts the fallacious ranking.

Extant interpretation-based accounts As noted in Chapter 1, a majority of research conducted on reasoning tends not to consider systematically issues of interpretation. The conjunction fallacy is a notable exception to this tendency. As Kahneman (2011) has recently observed, scholars of reasoning have criticized virtually every linguistic aspect of the experimental design, from the meaning of "likely" to that of "and." For example, Gigerenzer (1991) argues that the conjunction fallacy is an artifact of the particular experimental design used by Tversky and Kahneman (1983), based on the fact that, if one asks questions about the frequencies of the same properties over a population of say 100 individuals as opposed to a question about probabilities, the conjunction fallacy virtually disappears. The literature on the conjunction fallacy is truly vast, and no other classical piece of data on reasoning has sparked as much controversy.

The work I present in section 4.2 shares important properties with one route to explaining the conjunction fallacy, pursued by Dulany and Hilton (1991) and, as far as I can tell, not revisited until now. Because this is the only piece of work on the conjunction fallacy that shares part of the spirit of mine, it is pertinent and useful to survey briefly its strengths and weaknesses.

In this article, Dulany and Hilton explore the idea that epistemic implicatures that assert ignorance can absolve from the conjunction fallacy at least those subjects that are reliably computing those implicatures. Applying Grice's work to the conjunction fallacy, they propose that some subjects may be interpreting the "bank teller" option as committing to ignorance about whether Linda is active in the feminist movement or not. They submit that it is in fact rational for reasoners to assign a low probability to the ignorance component of the non-conjunctive option, and they conduct a number of experiments attempting to discern when an experimental subject is interpreting the non-conjunctive option in this way.

This idea is crucial to my proposal in section 4.2, summarized earlier in this introduction, and the authors deserve all the credit for stating and exploring it for the first time. There are however important gaps in their account. Some of these gaps are due entirely to the state of the art of pragmatics at the time of writing (the paper was published in 1991). For example, their assumptions

about when and how the relevant implicature will arise are not quite correct, and as is natural more recent developments in formal pragmatics allow me to be much more precise about these questions. But there are also substantive gaps in their informal account which mine resolves. First, their article and their account are about the transparent cases, and the account does not extend obviously to the direct-subtle and indirect conjunction fallacy tests. Second, they do not offer an explanation of *why* it may in fact be rational to reject the relevant statements of ignorance. Third, perhaps because they do not explore this issue of rationality, they do not identify the independent biases that the conjunction fallacy can be reduced to, once one factors out the effects of interpretation.

4.2 The conjunction fallacy as reasoning about ignorance

4.2.1 Secondary implicatures: an insufficient interpretation-based account

Consider the following very attractive explanation of the conjunction fallacy in terms of interpretive processes. I focus on the transparent task. Subjects are comparing the two options: they are being given a choice between B and $B \wedge F$, and so the two alternatives are surely salient during the process of pragmatic enrichment. Given this salient comparison and choice situation, it is expected that they should interpret the two options exclusively, where exclusivity does not lead to contradiction. Thus, the non-conjunctive option (B) is successfully strengthened as in (60a), while the conjunctive option is not strengthened, since (60b) is a contradiction.[57]

(60) a. $(B \wedge \neg(B \wedge F)) \Leftrightarrow (B \wedge \neg F)$

b. $(B \wedge F \wedge \neg B) \Leftrightarrow \bot$

Granting that subjects consider the probability of F to be higher than that of $\neg F$ and that B and F are independent propositions, the observed ranking is no longer fallacious. This is due to the theorem

[57] A variation on this explanation is to consider that subjects might take the experimenter to be asserting $B \vee (B \wedge F)$. This disjunction is a violation of what is known as Hurford's constraint, a generalization about the felicity of disjunctions where one of the disjuncts entails the other. As Chierchia et al. (2012) point out, in such cases it is obligatory to interpret the weaker disjunct exhaustively with respect to the stronger disjunct, as in (60a), while the stronger disjunct is not strengthened.

in (61).[58]

(61) An elementary theorem of probability theory:

If $\varphi \cap \theta = \psi \cap \theta = \emptyset$*, then* $P(\varphi) > P(\psi) \Rightarrow P(\varphi \wedge \theta) > P(\psi \wedge \theta)$.

Tversky and Kahneman (1983) consider a version of this explanation and argue against it. Recall the *indirect* version of their experiment, where they observed that the conjunction fallacy is prevalent even if, instead of asking subjects to choose one of the two crucial options as most likely, one asks them to rank the options (more) independently on a nine point scale of likelihood. They argue that, because the conjunctive option is not being pitted against its constituent conjuncts, there is no reason to expect subjects to get an exclusive interpretation for the simpler conjuncts. Whether this follows or not depends on the specifics of their indirect experiment. According to their (1983) paper, the indirect task did not compare the conjunctive option with its individual conjuncts. Instead, one group ranked a number of options that included the conjunctive option but neither of its conjuncts, while another group ranked a list of options that included "its separate constituents" (*op. cit.*, p. 298) but not the conjunctive option. It is unclear from the cited paper whether this second group was further divided into two subgroups, each of which saw only one of the conjuncts, or whether the second group always saw *both* individual conjuncts. The question is crucial to assess whether the indirect test provides an argument against the implicature analysis just sketched above. As Katzir (2007) and Fox and Katzir (2011) point out, relevant formal alternatives can come from the larger context in which an utterance was made, incorporating content that may even be entirely absent from the target utterance itself. Thus, all that is needed to get pragmatic strengthening from B to $B \wedge \neg F$ is for "active in the feminist movement" (F) to be a salient proposition in the context. If the second indirect-task group saw both B and F in the list that was to be ranked, then it is quite possible that they interpreted each option as being exclusive of the other. This would derive the observed ranking without need for an account that posits non-normative heuristics for judging the

[58] Notice that both or either of these assumptions may well be unwarranted, so that this interpretation-based approach to the conjunction fallacy does *not* altogether explain away the fallacy. Instead, it reduces the conjunction fallacy to independently observed biases that can be more transparently and precisely stated.

probabilities of conjunctions.[59]

In another experiment, Tversky and Kahneman (1983) substituted (62a) for the original "Linda is a bank teller" and ran the direct "which is more probable" target question on (62a) vs. (62b). The conjunction fallacy was still observed, despite a drop in acceptance rates from 85% in the original experiment to 57% in this followup.[60]

(62) a. Linda is a bank teller whether or not she is active in the feminist movement.

b. Linda is a bank teller and is active in the feminist movement.

Because (62a) is explicitly noncommittal about whether Linda is active in the feminist movement, it is unlikely if not impossible that it will be strengthened pragmatically to "not active in the feminist movement." This argument is valid, for (62a) is no longer entailed by (62b), Hurford's constraint is no longer being violated, and therefore the strengthening discussed above is not obligatory. On any view of pragmatics I am familiar with, if a pragmatic process is not obligatory and its computation results in a contradiction, competent speakers are very unlikely to execute it.[61]

[59]This strikes me as the interpretation most consonant with the letter of the cited article. But even if the experimental design was such that different people in the "separate conjuncts" condition only saw one of the conjuncts, something tantamount to the relevant strengthening could still have happened. This condition in the experiment involved a number of filler items beyond the target conjunct(s). For the Linda problem, one of the fillers was "Linda is a member of the League of Women Voters." Subjects who saw the crucial "Linda is a bank teller" conjunct definitely saw this filler as well, irrespective of whether they saw the other conjunct "Linda is active in the feminist movement." By the same reasoning given above, these subjects likely interpreted "Linda is a bank teller" in an exclusive way with respect to "Linda is a member of the League of Women Voters." Finally, if "Linda is a member of the League of Women Voters" is intuitively ranked above *the proposition* "(that) Linda is a bank teller," its exclusion from the interpretation of *the sentence* "Linda is a bank teller" would have a similar effect to exclusion of "Linda is active in the feminist movement," bringing the ranking of the single conjunct below that of the conjunction, while proving no fallacious reasoning in what concerns the conjunction.

[60]Tversky and Kahneman (1983) report but do not discuss this improvement in performance. Interestingly, they report identical (57%) rates of conjunction fallacy in betting scenarios ("If you could win $10 by betting on an event, which of the following would you choose to bet on?"), where it is reasonable to think that experimental subjects suspend the calculation of implicatures. The rationale for this suspension of pragmatics is that betting scenarios are not necessarily situations with normal communicative cooperativeness. That is, betters know, not just that the stakes are high, but also that it is crucial to be very cautious and therefore *literal* with respect to the interpretation of the options. Bookies do not typically reward betters who merely implied the observed outcome.

[61]This is not quite the argument that Tversky and Kahneman (1983) give. In their view, what is interesting about (62) is that "the extension of 'Linda is a bank teller whether or not she is active in the feminist movement' clearly includes the extension of 'Linda is a bank teller and is active in the feminist movement"' (*op. cit.*, page 299). Presumably, the authors find the inclusion obvious because they consider "whether or not she is active in the feminist movement" to be a tautology. But this is unwarranted. First, notice the simple fact that, to take that clause to be a tautology in the relevant sense is to accept that it could block the scalar implicature while *meaning*, as it were, nothing (or rather, something trivial). Second, and more importantly, appending this locution to (62b) as in (i) (footnote continued on the next page) results in a sharp contradiction, which is unexpected if it were a tautology.

While the experimental results with the stimuli in (62) show that the conjunction fallacy can be observed without the interpretation in (60a), they do so by introducing a new ingredient into the mix, namely the proposition expressed by the second conjunct of (62a). Happily, the literature offers a better controlled argument making the same point.

Tentori et al. (2004) used a simple and effective experimental paradigm to show that the conjunction fallacy occurs even without the pragmatically enriched interpretation in (60a). They asked subjects to choose the most probable statement from sets of three statements. The following is a representative example of their stimuli.[62]

> The Scandinavian peninsula is the European area with the greatest percentage of people with blond hair and blue eyes. This is the case even though (as in Italy) every possible combination of hair and eye color occurs. Suppose we choose at random an individual from the Scandinavian population.
>
> Which do you think is the most probable? (Check your choice.)
>
> 1. The individual has blond hair.
> 2. The individual has blond hair and blue eyes.
> 3. The individual has blond hair and does not have blue eyes.

They found that around 60% of subjects ranked option 2. as the most likely, committing the conjunction fallacy. Importantly, it is highly implausible that any of these subjects interpreted option 1. to mean "the individual has blond hair and not blue eyes," for this would make option 1. equivalent to option 3. The potential strengthening of 1. would therefore attribute uncooperativeness to the

(i) Linda is a bank teller and is active in the feminist movement, whether or not she is active in the feminist movement. ($\Leftrightarrow \bot$)

Notice that it does matter whether the intuition of contradictoriness in (i) is a pragmatic or purely semantic intuition. There may well be good reasons to think that strings like "whether or not she is active in the feminist movement" are in fact tautologies as far as their literal content is concerned, in which case our intuition about (i) might be of the Moore-paradoxical kind (see for example Hintikka, 1962). Be that as it may, I know of no good reason to believe that subjects in these experiments are accessing only the literal content of these locutions.

[62] The experiment was run on students at the University of Padua. The translation I give here is the same as presented by Tentori et al. (2004).

experimenter. That is, to strengthen 1. above is to assume that the experimenter, who is presumably responsible for uttering the three options, offered two equivalent options for subjects to choose from.

The upshot of these experiments is the following. Pragmatic strengthening of the non-conjunctive option in the conjunction fallacy does *not* account for all observed instances of the conjunction fallacy. In other words, scalar implicatures do not immediately explain the conjunction fallacy away. This is not to say that a strengthening of this sort is not responsible *in part* for the conjunction fallacy, in the cases where the strengthening is plausible. Indeed, it is at least an interesting fact that, where we can be confident that the strengthening is blocked, we get conjunction fallacy rates around or under 60%, as opposed to 85% where the strengthening is possible and likely to occur. In the next few sections, I show how all of these data points can be accounted for by recourse to a different kind of pragmatic process.

4.2.2 Primary implicatures

A crucial intuition behind my account can be summarized as in (63).

(63) When the context suggests strongly that the question *whether* φ is resolved, speakers/reasoners reject statements that commit to ignorance about φ.

I discuss the rationale for (63) later in section 4.2.4. First, it is important to see how (63) explains the conjunction fallacy in the cases considered in this paper.

The first ingredient involves primary implicatures. In modern formal pragmatics (see for example Sauerland, 2004), primary implicatures, also known as *epistemic* implicatures, are inferences of the sort "the speaker of S is not in a position to assert φ," or, in typical cases, "the speaker of S is ignorant about *whether* φ," for S an utterance of a sentence and φ a relevant alternative to that sentence.[63,64] For example, if a speaker utters (64a), a statement with (at least) the relevant alterna-

[63] The question of what precisely constitutes a "relevant alternative" in this sense is a subject of debate in the literature. The assumptions made in this chapter about the relevant alternatives are entirely uncontroversial on anyone's account, so I refrain from espousing a particular theory of formal alternatives for these purposes. My assumptions in this chapter are also perfectly compatible with the fully spelled-out theory of implicature from Chapter 3.

[64] It is unclear when exactly the formally generated primary implicature of the form "speaker is not in a position to assert" is interpreted in the stronger form "speaker does not know whether." The issue will not make a significant difference for the account in this chapter. Se also footnote 66, p. 103.

tives in (64b), she will generally be committed to the primary implicatures in (64c), derived from the formal alternatives in (64b).

(64) a. John or Mary is coming to the party.

b. {*John is coming to the party*, *Mary is coming to the party*}

c. The speaker does not know whether John is coming to the party, and the speaker does not know whether Mary is coming to the party.

In (64), the relevant alternatives with respect to which we calculate ignorance implicatures are given explicitly in the original utterance (64a). This is an interesting property of natural language disjunctions, but alternatives can also come from the broader context.

(65) *Mary got a terrific grade in her math exam. She also got a comparable grade in her chemistry exam.*

a. Mary told John she had aced her math exam.

b. Mary didn't tell John she had aced her chemistry exam.

c. The speaker of a. does not know whether Mary told John that she had aced her chemistry exam.

The contextual information (in *italics*) given in (65) makes Mary's math exam *and* her chemistry exam salient, but the utterance in (65a) speaks only of her math exam. In this situation, it is natural for a hearer of (65a) to derive the *secondary* implicature in (65b). This shows that the alternative "Mary told John that she had aced her chemistry exam" must be available to pragmatic computations. This can only be the case if, when computing the implicatures of (65a), hearers have access to the phrase "chemistry exam," which in turn can only happen if hearers have access to salient alternatives from a context larger than simply the sentence whose implicatures are being calculated.[65]

It follows that this contextually supplied alternative will also be available to pragmatic computations of primary implicatures, responsible for deriving (65c). Now, whenever (65b) is derived as an implicature of (65a), the ignorance implicature in (65c) will be blocked, for the speaker is in fact

[65]This point is related to that made in section 4.2.1, starting on p. 98, regarding access to contextual alternatives in Katzir's (2007) theory of formal alternatives. See also footnote 32 in Chapter 3, p. 63.

assumed to have an opinion about the salient question of whether Mary told John that she aced her chemistry exam.[66]

But if the strong secondary implicature in (65b) is somehow blocked, the weaker primary implicature in (65c) will surface. This process, I argue, is also operative in the conjunction fallacy cases discussed above, where we can be confident that the strong "not active in the feminist movement" (respectively, "does not have blue eyes") implicature is being blocked. Consequently, I predict that, whenever the relevant secondary implicature is blocked, hearers will still calculate a relevant, though weaker, primary implicature, provided that the crucial alternative is present in the context of the target sentence, under a reasonably broad view of context. Specifically, for the Linda problem and the Scandinavian individual problem discussed above, we get primary implicatures that can be roughly paraphrased as follows.

(66) a. Linda is a bank teller.

STRENGTHENING: Linda is a bank teller and one does not have a position on whether she is a feminist or not.

b. The individual has blond hair.

STRENGTHENING: The individual has blond hair and one does not have a position on whether he/she has blue eyes or not.

[66] I substantially simplify the issue of how primary implicatures interact with secondary implicatures, in order to present the spirit of my account in a theory-neutral way. Technically, within the class of theories of implicature espoused in Chapter 3, the primary implicature for (65a) is as in (i).

(i) The speaker of (65a) does not believe that Mary told John that she had aced her chemistry exam.

This is weaker than what I give in (65c), which entails that the speaker does not believe that Mary did *not* tell John either. The primary implicature in (i) is in fact compatible with the speaker believing that Mary did not tell John, so that the primary (ignorance) implicature does not need to be "blocked" in any meaningful way to accommodate the secondary implicature. But now we must ask what happens when the secondary implicature is blocked. Do we get the stronger ignorance inference in (65c), or are we left with the weaker (i)? As far as I know, the literature does not address this specific question. My introspective judgment is that we do get the stronger ignorance inference, whence the (simplified) paraphrase in (65c) and elsewhere in this section. Happily, this question will not matter. I turn to the original Linda problem: all that is required in my formal account is that the option "Linda is a bank teller" entail that (the speaker believes that) she *may not* be active in the feminist movement. In the absence of the secondary implicature (that she is not active in the feminist movement), the required epistemic proposition is surely entailed by the technically predicted primary implicature "the speaker does not believe that Linda is active in the feminist movement."

Primary implicatures are about epistemic states, they comment on the speaker's ignorance about certain facts. With that in mind, it would seem that (67) gives more accurate paraphrases of the predicted strengthened meaning of the sentences in question.

(67) a. Linda is a bank teller.

STRENGTHENING: Linda is a bank teller and the speaker does not have a position on whether she is a feminist or not.

b. The individual has blond hair.

STRENGTHENING: The individual has blond hair and the speaker does not have a position on whether he/she has blue eyes or not.

But who is "the speaker?" There are in principle two immediate possibilities. The holder of this epistemic attitude might be some speaker distinct from the experimental subject, whose existence is accommodated by the subject, perhaps even identified with the experimenter herself.[67] But it is also plausible that subjects consider the sentences in question not as assertions by some loosely identified speaker, but rather as potential assertions of their own. In other words, experimental subjects in these tasks may consider the act of picking one of the sentences as most likely over its alternatives to be tantamount to uttering that sentence. They may be asking themselves "what am I committing to by choosing this sentence, thereby uttering it?" In this case, the relevant "speaker" in (67) is in fact the experimental subject, rather than a second or a third party.

This question is an interesting one, but it is only of relative importance to my account which of these views is correct. Thus, I opt for the more neutral paraphrase "one does not have a position" as in (66), simply to draw attention away from the issue of who the relevant "speaker" is. I return to the issue briefly in the next section.

4.2.3 The account in a theory-neutral formulation

My account of the conjunction fallacy in cases where the relevant secondary implicatures are blocked builds on the above facts about primary implicatures. It goes as follows, using the setup

[67] In Chapter 3, section 3.2, I assume that secondary implicatures figure into reasoning in this way. Nothing in that account hinges on this assumption, and the choice was made only to allow me to spell out the account fully.

from the Linda problem for concreteness.

First, reasoners erroneously assign a probability to Linda's being active in the feminist movement that is rather high, above 50% in any event. I provide some independent reasons to believe that this is the case in the next section. Second, reasoners calculate primary implicatures committing "the speaker" to ignorance about relevant contextual propositions. For the Linda problem, and abstracting away from the particular way in which the relevant secondary implicatures get blocked, this means that the two options given in the transparent test are interpreted as follows.

(68) a. Linda is a bank teller, and one does not have a position on whether she is active in the feminist movement.

b. Linda is a bank teller and she is active in the feminist movement.

What is the normative ranking of the propositions in (68) with respect to their probability, granting that the probability that Linda is active in the feminist movement is higher than 50%? This will be given by the conjunction rule in (69).

(69) $P(\varphi \wedge \psi) = P(\varphi) \cdot P(\psi)$

Notice that the two options given in (68) are of the form $\varphi \wedge \psi$, respectively $\varphi \wedge \theta$. Notice also the simple arithmetic fact that, for $a \neq 0$, we will have $a \cdot b > a \cdot c$ just in case $b > c$. Consequently, and given that both options in (68) include as a conjunct "Linda is a bank teller," which presumably is not assigned probability zero, we can factor out this first conjunct when comparing the probabilities of the two conjunctions. All that matters, then, is what probabilities reasoners assign to the second conjuncts of each option in (68):

(70) a. One does not have a position on whether Linda is active in the feminist movement.

b. Linda is active in the feminist movement.

By assumption, reasoners believe that Linda is likely active in the feminist movement (70b). But then they cannot consistently assign a higher probability to (70a), which asserts ignorance about a proposition they in fact consider rather likely. The only internally consistent ranking of the propo-

sitions in (70) is therefore (70b) $>$ (70a). It follows that the internally consistent ranking of (68) is (68b) $>$ (68a). Insofar as (68a) is (roughly) the way in which reasoners interpret the non-conjunctive option they saw ("Linda is a bank teller"), as I argued in the preceding section, the consistent ranking reasoners ought to choose assigns a higher probability to the conjunctive option than to the non-conjunctive option.

Crucially, this is not a violation of the conjunction rule in (69). For the non-conjunctive option "Linda is a bank teller" only *appears* to be properly entailed by the conjunctive option. Instead, the two options are incompatible with each other. On this view, reasoners are not violating the norm on combining the individual probabilities assigned to events, so that the inferential behavior we observe should not properly be described as a conjunction fallacy. There may be other fallacies reasoners are committing when solving the Linda problem, and in fact I will argue shortly that they are making a mistake when they assign a high probability to the proposition that Linda is active in the feminist movement. But those mistakes will be distinct from, and entirely independent of, conjunction errors.

It is useful at this point to distil the account into its essential components.

(71) For a transparent Linda-type problem whose options are A and $A \wedge B$,

a. reasoners calculate primary (ignorance) implicatures for the A option, committing "the speaker" to ignorance about B;

b. reasoners consider that $P(B) > .5$;

c. option A pays a hefty price in the currency of probabilities because it asserts ignorance about a proposition reasoners do not believe one is ignorant about.

Component (71a) was argued for in detail in the preceding section, and (71b) will me motivated in the next section. (71c) however remains somewhat underspecified. Depending on who is "the speaker" who is said to be ignorant, (71c) can have one of two distinct rationales.

If "the speaker" is the subject of experiment himself, (71c) can be understood as entirely rational behavior, as spelled out above. The subject believes that Linda is indeed active in the feminist movement, so that his belief that *he* is not opinionated about that proposition has to be extremely

low. But if "the speaker" is the experimenter, a little more needs to be said, as self-consistency becomes at least a less immediate rationale for (71c).

One possibility is that subjects take the experimenter to be drawing the same conclusions they are. I return to the Linda example again, for concreteness. The experimenter has access to the same evidence as I, the subject, so she must have also concluded that Linda is likely to be active in the feminist movement. From there, we can reason in a way similar to the self-consistency route explored above. But there is another class of possibilities. Perhaps subjects do not assume that the experimenter reasoned in the same way as they; subjects might even recognize that they have no idea what the experimenter's position on Linda's activity in the feminist movement is. Even in that case, there is a plausible rationale for (71c). Subjects might take the experimenter to be in a privileged epistemic position, given that she is responsible for concocting / reporting the story they just read. Though subjects may not know what the experimenter thinks about Linda's activity in the feminist movement, they would seem justified to believe that the experimenter either knows the facts or at least has a position on those facts. In this case, option *A* pays a price for contradicting an independent assumption of experimental subjects, namely that whoever administers a psychological experiment is in possession of all the relevant knowledge. I do not settle this question in the present work, and must leave these possibilities open for future research.

4.2.4 Base-rate neglect

In the superb (1974) paper on judgment under uncertainty, Tversky and Kahneman list a rich inventory of heuristics and biases operative in humans' reasoning about probabilities. The representativeness heuristic was first articulated in this article. One of the manifestations of representativeness that Tversky and Kahneman discuss is what they called "insensitivity to prior probability of outcomes," or base-rate neglect. The essence of this bias is that reasoners effectively ignore the prior probabilities of outcomes when assessing the probabilities of those outcomes.

In an experiment reported in that article, subjects were shown personality descriptions of several individuals and were tasked with assessing, for each description, the probability that it belonged to an engineer and the probability that it belonged to a lawyer. Subjects were split into two condi-

tions. In one, they were told that the group of individuals from which the descriptions came was comprised of 100 people, of which 70 were engineers and 30 were lawyers. In the other condition, they were told the group had 30 engineers and 70 lawyers. Thus, the prior probabilities that any given description belonged say to an engineer were very different in the two conditions. The normative standard of probability theory demands that, for the same description, a higher probability be assigned to the probability that it belongs to an engineer in condition one (70% engineers in the group) than in condition two (30% engineers in the group). Instead, Tversky and Kahneman found no significant difference between the probabilities assigned by subjects in each condition. Subjects' answers appeared to be ignoring the base-rate frequencies of the two professions altogether, as was expected if they were substituting a question about representativeness for a question about (normative) probabilities.[68]

I propose that base-rate neglect of this kind is operative in the conjunction fallacy experiments I discussed above. I turn to Linda again for concreteness. The base-rate frequency of women active in the feminist movement is almost certainly rather low.[69] Let us say two out of every ten Ivy-league-educated women (for recall that Linda went to Harvard) are active in the feminist movement. No matter how well the description given of Linda matches that of someone active in the feminist movement, the low prior probability should have an important impact on the probability assessment about Linda's involvement in the feminist movement. The account given above assumes instead that subjects are assigning a rather high probability to the relevant propositions. Application of the representativeness heuristic in the form of base-rate neglect makes this a perfectly warranted

[68]Tversky and Kahneman (1974) showed also that humans are perfectly capable of using prior probabilities in a normative way, it simply has to be the case that there is no other information available to them. The authors gave an almost identical task to a different group of subjects, except that these subjects did not see personality sketches, and were instead asked to judge the probability that a randomly selected individual from those two groups of 100 people would be an engineer. Here, the answers were the correct .7 for condition one and .3 for condition two. The representativeness heuristic has a very strong draw, but in the absence of the required property descriptions with respect to which to assess representativeness, humans suddenly show an ability to apply basic principles of probability theory correctly. Unfortunately though, this effect consonant with the norm disappears once descriptions are introduced, even if they are entirely *irrelevant* to the question at hand. Given the opportunity, it seems like humans will always prefer to use representativeness judgments, quite possibly as a way to maximize the relevance of the information given in the experiment: "if I was given a description of an individual *beside* some numerical probabilities, surely I am meant to focus on that description as I answer this question."

[69]According to a HuffPost/YouGov poll from 2013, only 23% of American women consider themselves feminists (`http://big.assets.huffingtonpost.com/tabs_gender_0411122013.pdf`). I submit that probably an even smaller percentage would say they are active in the feminist movement.

assumption. If, as Tversky and Kahneman (1974) argue, humans substitute questions about representativeness for questions about probabilities whenever possible, then we expect these base-rates not to figure into their assessment of the likelihoods of the relevant atomic propositions in the Linda experiment. A low base-rate frequency of individuals with the Q property will affect the normative judgment of the probability that a has the Q property, but it will not affect a judgment about how representative of property Q individual a is.

This is not to say that we need to accept the representativeness heuristic wholesale, as given by Tversky and Kahneman (1974). Representativeness offers an appealing explanation of *why* base-rate frequencies are so often ignored in favor of information from individual descriptions. But whatever one thinks of that explanation, *that* people tend to ignore base-rate frequencies is a very well established fact about human reasoning. That fact, whatever its cause, is all that is needed to motivate this particular component of my account.

4.2.5 The conjunction fallacy in Kratzer's (1991) theory of modality

In this section I give a formal account of how ignorance figures into reasoning in the conjunction fallacy, in the context of a well-known theory of modality from linguistic semantics.

4.2.5.1 Motivations for the formal account

My explanation of the conjunction fallacy in terms of reasoning about ignorance was stated informally above. I believe that this is a useful move given that the spirit of the account is of value irrespective of specific implementations. Conversely, there is also value in spelling out fully the more novel aspects of the account. In particular, the issue of how exactly ignorance figures into reasoning was left largely open. I argued that the two options in the transparent Linda problem are interpreted roughly as follows.

(68) a. Linda is a bank teller, and one does not have a position on whether she is active in the feminist movement.

b. Linda is a bank teller and she is active in the feminist movement.

The probabilities of each of the second conjunctions in (68) are what is crucial. But how exactly do we assess the probability of propositions of the form "one does not have a position on . . . "? One reasonable way to find an answer to that question is to look at semantic theories that assign precise truth-conditions to epistemic and probability talk.

Kratzer (1991) offers a theory of modals like "might" and "must," as well as locutions such as "it is likely that." Kratzer's proposal does so while maintaining a semantic ontology that is in principle compatible with humans' poor performance in reasoning tasks involving probabilities. The theory is built not on measure functions that map propositions onto a bounded continuum of probability values, but rather on a notion of *comparative possibility*. The idea is that the contents of sentences of English do not receive real-valued probabilities, instead they are assigned qualitative rankings based on underlying qualitative rankings on possible worlds (in one version of the theory). This system is strictly less expressive than probability theory, but it allows one to give empirically adequate accounts of almost the entirety of probability talk.[70] Interestingly, it turns out to derive the observed ranking of options in Linda-style problems, given an analysis of their interpretation in the spirit of the one I give above, and making a few reasonable assumptions.

The first assumption, perhaps better described as a simplification, is that the content of the crucial primary implicature can be sufficiently faithfully paraphrased as (72), to be compared with (71a).[71]

(72) Linda is a bank teller, and she may or may not be a feminist.

[70] Recent work, in particular by Yalcin (2010) and Lassiter (2011), has shown that Kratzer's original (1991) theory validates certain normatively and intuitively invalid inference patterns involving probabilities and disjunction. This has motivated some scholars to abandon qualitative theories such as Kratzer's in favor of the full expressive power of probability theory. But the debate is still ongoing; in a talk presented at Harvard in November 2012, Paul Portner showed that it is possible to block the undesired inference patterns in Kratzer's theory by assigning a bigger role to context in determining the ordering source, allowing it to vary between premises in an argument. Moreover, even if Kratzer's (1991) theory turns out to have deeper and unsolvable issues, there exist other theories in logical space that preserve the attractive qualitative character of Kratzer's approach but validate all of the right inference patterns (Holliday and Icard, 2013).

[71] Strictly speaking, all that is needed for the account to work is the weaker (i). In keeping with the stronger paraphrases I used in preceding sections I will prove the result for the stronger (72).

(i) Linda is a bank teller, and she may not be a feminist.

This assumption is required because Kratzer does not give a semantics for locutions like "one does not have a position on." As far as I can see, (72) offers what is intuitively a close-enough paraphrase. I now give the technical elements from Kratzer's theory necessary to derive the desired result. My discussion of these technical aspects includes only what is absolutely required, and I do not provide any of the motivations for these technical components. I refer the reader to Kratzer's body of work for those details.

4.2.5.2 Technical background and the result

This theory of modality relativizes modal statements to an *ordering source* and a *modal base*. The ordering source is a partial ordering on possible worlds, the nature of which will depend on the kind of modality. For the case of epistemic modality, the ordering source makes sure that more "normal," less "surprising" worlds are ranked higher than more bizarre worlds with respect to a world of evaluation. This epistemic ordering is related to the notion of similarity introduced by Stalnaker (1968, 1975).

Definition 22 (Ordering source). I write $u \geq_{g(w)} v$ iff u is at least as close as v is to the ideal provided by the ordering source g at the world of evaluation w. Since the ordering source will remain constant, I omit the subscript and write simply $u \geq v$.

The modal base consists of a body of modally relevant propositions, and it can in principle vary between worlds. For epistemic modality, modal bases are bodies of information, or information states, representing what is believed at each given world.[72]

Definition 23 (Modal base). The function $f : W \rightarrow \wp(W)$ takes a world and returns an information state, modeled as a proposition. It represents what is believed at a certain world.

We can now define the central notion of comparative possibility in the theory.

[72] Believed by whom? As in the informal exposition of the account in section 4.2.3, I gloss over the issue of whose information state is at stake, the experimental subject or some other party.

Definition 24 (Better possibilities). A proposition φ is a better possibility than ψ at a world w relative to a modal base f (I omit the ordering source for simplicity), in symbols $\varphi \succ_w \psi$, iff $(\exists u \in f(w) \cap (\varphi - \psi))\ (\forall v \in f(w) \cap (\psi - \varphi))\ u > v$.

In words, φ is better than ψ at w if there is an accessible world u that is a φ and not ψ world that is better than any accessible world that is a ψ and not φ world. To apply this account of modality to the sentences we are interested in, we must now give a definition for the modal "may/might," which I represent as $\Diamond$.

It is a matter of some debate in the literature on this class of semantics whether one can safely assume that any set of worlds will always contain $>$-maximal worlds. A world is $>$-maximal with respect to a set of worlds F if there is no world that outranks it in F.

Definition 25 (Maxima). A world w is a maximum relative to a set of worlds F, in symbols $\textsc{Max}_F(w)$, iff $(\neg\exists u \in F)\ u > w$.

This *limit assumption* is endorsed, among others, by Stalnaker (1981), and it simplifies both definitions and proofs significantly.[73] Armed with it, we can define the required modal in a reasonably simple fashion.

Definition 26 (Might). A statement of the form "might φ," in symbols $\Diamond\varphi$, is true at a world w relative to a modal base f, and under the limit assumption, iff $(\exists u \in f(w))\ \textsc{Max}_{f(w)}(u)\ \&\ u \in \varphi$.

The full list of technical assumptions required to derive the result in this section is the following. All of these assumptions are rather common in the literature, but none is trivial.

1. Limit assumption: for any set of worlds, there will always be one or more maxima.
2. Correctness of the modal base: $w \in f(w)$.
3. Monotonicity of the modal base: $u \in f(v) \implies f(u) \subseteq f(v)$.

First, I note that correctness and monotonicity have obvious counterparts in standard modal logic, respectively reflexivity and transitivity. Second, it is interesting to reflect on what these assumptions commit us to regarding how humans reason about epistemic possibilities.

[73] The limit assumption has an important dissenter in Lewis (1981), who shows that for certain sentences involving quantification over infinite domains the limit assumption is incoherent. The debate is very much outside the scope of this dissertation.

Correctness essentially means that reasoners assume that, while they may be ignorant about some facts, they are not mistaken. On the other hand, monotonicity says that, while reasoners can imagine their information states being different in different situations (worlds), they assume that they can only gain information, never lose information. These two informal glosses are related, expressing a certain form of epistemic optimism (or hybris).

These observations are of interest because this formal account should not simply be a way to derive a formal result about the conjunction fallacy. Ideally, it would offer some insight into what is behind the phenomenon. I suggest, though I cannot explore the issue further in this dissertation, that the correspondence between the assumptions and epistemic optimism I just drew is the insight offered by this technical account. In other words, this account does not explain the fallacy away, instead it reduces it to (A) a particular interpretation of the predicate "be likely" in terms of comparative possibilities, and (B) the two related cognitive biases I dub epistemic optimism.

I conclude this section with the theorem required to derive the conjunction fallacy. Together with the interpretive considerations on primary implicatures that I argued for in section 4.2.2, this theorem predicts the inferential behavior observed in the conjunction fallacy literature.

Theorem 1. *If $\varphi \succ_w \neg\varphi$, then $\varphi \succ_w \Diamond\varphi \wedge \Diamond\neg\varphi$. In words, if φ is a better possibility than $\neg\varphi$ at w, then φ is also a better possibility than "maybe φ, maybe not φ."*

Proof. By assumption and Definition 24 we have a world, call it a, such that $a \in \varphi$, $a \in f(w)$, and $(\forall v \in f(w))\ v \notin \varphi \implies a > v$. We must now show that there is some world in $\varphi - (\Diamond\varphi \wedge \Diamond\neg\varphi)$ that is better than any world in $(\Diamond\varphi \wedge \Diamond\neg\varphi) - \varphi$. Since the second set contains only $\neg\varphi$ worlds, it suffices to show that $a \in \varphi - (\Diamond\varphi \wedge \Diamond\neg\varphi)$. Since $a \in \varphi$, we get that $a \in \varphi - (\Diamond\varphi \wedge \Diamond\neg\varphi)$ just in case $a \notin \Diamond\varphi \wedge \Diamond\neg\varphi$. Assume that $a \in \Diamond\varphi \wedge \Diamond\neg\varphi$, to derive a contradiction. By Definition 26, this entails that $(\exists u \in f(a))\ \textsc{Max}_{f(a)}(u)$ & $u \notin \varphi$. Call this element b; we then have $\textsc{Max}_{f(a)}(b)$ and $b \notin \varphi$. By monotonicity of the modal base, $b \in f(w)$. Since $b \notin \varphi$ and $\varphi \succ_w \neg\varphi$, we get $a > b$. By correctness of the modal base, $a \in f(a)$, so $\neg\textsc{Max}_{f(a)}(b)$. This is a contradiction. □

4.3 Assessing the role of representativeness in the conjunction fallacy

Tversky and Kahneman (1983) argue that, when asked about probabilities, humans substitute for the actual question they are asked a similar question about *representativeness* rather than about likelihood. I conclude this chapter with a preliminary study on the conjunction fallacy that attempts to tease likelihood and representativeness apart, so as to compare their contributions to the conjunction fallacy.

4.3.1 Representativeness and judgments of likelihood

As discussed in connection with base-rate neglect (section 4.2.4), the strategies that humans tend to deploy to assess the likelihoods of statements do not conform to the norms of probability theory. Tversky and Kahneman (1983) offer an explanation for this fact: instead of asking themselves a question about probabilities, humans ask questions about representativeness. For example, in the case of the Linda problem, humans ask themselves question (73b) instead of the targeted (73a).

(73) a. How likely is it that Linda is a bank teller?

b. How typical an example of a bank teller is Linda?

Representativeness substitutions as in (73) allow the authors to give an elegant account of several important classes of facts. The conjunction fallacy, as discussed in the preceding sections, is a good example, but it is not the only one.

In a relatively recent article, Kahneman and Frederick (2002) revisit the representativeness heuristic, responding to some of the criticism leveled against the original heuristics and biases approach (Tversky and Kahneman, 1974) and giving a more explicit account of what is meant by representativeness in terms of prototype theory. The authors point out that, while the transparent Linda problem may well be amenable to analyses alternative to representativeness, three classes of closely related facts were left unexplained by these alternative accounts at the time of writing of that paper (2002).

1. Subtle comparisons — conjunction fallacy cases where subjects are given a number of fillers

along with the crucial A and $A\&B$ options,

2. Between-subjects comparisons — where no subject sees *both* of the crucial options A and $A\&B$,

3. The virtually perfect correspondence between judgments of likelihood and judgments of representativeness for particular statements, irrespective of conjunction effects.[74]

Points 1. and 2. were addressed earlier in this chapter, where I offered a direct account of 1. and argued that 2. either reduces to 1. or needs more study, depending on particulars of the experiments carried out by Tversky and Kahneman (1983).[75] The discussion in this dissertation so far does not apply to point 3. above, which is left entirely open. Indeed, if humans' judgments of probability always track their judgments of representativeness, then the representativeness heuristic seems to be on the right track, no matter how many attractive alternative explanations there may be for certain subclasses of data, like the conjunction fallacy.

The question then arises: can we tease the two notions apart? That is, can we find sentences that rank high on likelihood but low on representativeness? If so, we have found potential ways of falsifying the representativeness heuristic in a strong way. The most immediate strategy for finding such cases is to explore the asymmetry between the predicates "representative" and "likely." To wit, "likely" takes a clausal complement, while "representative" takes an individual and a property. For example, it is hard to concoct bona fide representativeness questions to correspond to likelihood questions as in (74). When asked about (74a) or (74b), what representativeness question do we substitute for the likelihood questions we see?

(74) a. How likely is it that it is raining?

b. How likely is it that the Goldbach conjecture is true?

This seems like a difficult question, but it is perhaps not impossible. One could argue that (74a) contains an event or situation argument and that that argument has a topical role in the sentence,

[74] For the case of the Linda data, this is the fact that the mean *likelihood* rankings of (non-conjunctive) descriptions of Linda are virtually identical to the mean *representativeness* rankings of those descriptions. For example, different groups of subjects were asked to give an answer to (73a) and (73b). The rankings turned out to be identical.

[75] See the beginning of section 4.2.1 (p. 97), as well as footnote 59 (p. 99).

so that the representativeness question is "To what extent is the current situation a typical example of a raining situation?" Let us grant that this analysis is reasonable. Can we find other ways of making representativeness assessments impossible? I propose that conjunction-fallacy problems offer a promising alternative route.

One interesting property that the data from Tversky and Kahneman (1983) have is that every item subjects had to rank was about the same individual. In fact the individual was the syntactic subject of every sentence subjects were asked about. For the Linda story, the items were as in (75).

(75) a. Linda is a teacher in elementary school.

b. Linda works in a bookstore and takes Yoga classes.

c. Linda is active in the feminist movement.

d. Linda is a psychiatric social worker.

e. Linda is a member of the League of Women Voters.

f. Linda is a bank teller.

g. Linda is an insurance salesperson.

h. Linda is a bank teller and is active in the feminist movement.

What happens if we introduce an item into the fray that is *not* about Linda? Unfortunately, none of the articles cited in this dissertation's bibliography reports an experiment with this property. That is, every conjunction fallacy (or more generally representativeness) study I have reviewed uses collections of sentences that are clearly about the same topical individual, usually the sentential subject. Now, simply adding a sentence about "Mary" to the list in (75) does not make for a particularly interesting experiment. A cost might be incurred for throwing off subjects by giving them a sentence so different from the others, but presumably subjects would take whatever information they have about Mary and ask themselves a simple representativeness question about Mary. That assessment would produce a ranking, allowing them to place the new sentence in an appropriate position relative to the other sentences. The representativeness heuristic would account for this immediately.

Consider now what happens when we conjoin two sentences about different individuals. For concreteness, compare the standard (76a) with (76b).

(76) a. Linda is a bank teller and is active in the feminist movement.

b. Linda is a bank teller and Mary is a school teacher.

It is clear what the representativeness analysis of (76a) is, for the representativeness question is simply "to what extent is Linda a typical example of someone who is a bank teller and active in the feminist movement." But what is the representativeness question for (76b)? One might think that the question will involve asking how good an example of a bank teller Linda is, how good an example of a school teacher Mary is, and by some arithmetic process combining the two degrees of representativeness. First, notice that this would require the representativeness heuristic to be sensitive to the structure of the sentences under consideration, a rather substantive move that has not been suggested in the literature. Regardless, what would be the result of this operation of combination? It ought to be a degree of representativeness for the *conjunction* (76b), but one is at a loss trying to decide what the category and individual required for a well-formed statement about representativeness would be.

Again, the literature as far as I can tell does not provide an answer to this question.[76] It seems that, whatever the degree of representativeness one assigns to (76b), one will have to make a choice about the relevant individual and the relevant category. I do not see a straightforward way to generalize extant formalizations of representativeness so as to make a coherent prediction about the ranking of (76b).[77]

[76]This is true not only of the early, informal expositions of the representativeness heuristic (Tversky and Kahneman, 1974, 1983), but also of Kahneman and Frederick's (2002) formal account in terms of prototype theory. Alternative formalizations in terms of exemplar theory (see for example Nilson et al., 2008) also do not offer an answer.

[77]This is not to say that it is impossible to rescue the representativeness account, especially if one formalizes the heuristic itself from scratch, using standard formal-language theory tools.

One strategy would be to make the higher-level heuristic more complex and sensitive to the linguistic structure of its argument. The original representativeness account already required, though this was never made explicit, that some amount of linguistic structure be accessed. For example: instead of assessing the likelihood of a sentence S, one assesses how representative the subject of S is of the predicate of S. This much already prompts difficult questions, the most salient of which being whether subjects and predicates are the relevant categories, as opposed to say topics and comments. But the challenge from composite representativeness assessments I outline above requires more access to structure. One possibility would be to define a function L from sentences of English into degrees of likelihood inductively, so that in the base case L is representativeness. For concreteness, $L(\varphi \wedge \psi) = L(\varphi) \circ L(\psi)$, for some arithmetic operation $\circ$, but for monoclausal structures (the base case) $L(P(x)) = R(P)(x)$.

Again, I have no reason to think that this question cannot be answered correctly and elegantly. Instead, I am pointing out that the question is open, and that one promising route to an answer seems to involve a significant rethinking of how the representativeness heuristic itself works, requiring the use of formal linguistic methods.

In the next section, I present a pilot experiment I conducted in an attempt to separate representativeness from probability. The experiment builds on the challenge to representativeness accounts that is introduced by conjunctions with entirely distinct topical individuals. The results of this pilot are very tentative, but I believe suggestive that this line of research holds promise. If the design of this study is on the right track, some development of it may allow us to answer more ambitious questions about the role of representativeness in judgments of likelihood.

4.3.2 Pilot design

I selected and slightly adapted four conjunction-fallacy setups from the literature, all included in Appendix C. Each story was clearly about some individual a. I composed three statements following the guidelines in (77).

(77) A A statement about d but not especially representative of d,

B A statement *not* about d and independently likely.

C The conjunction of A and B.

For concreteness, the following was one of my sets of stimuli.

> In what follows, assume that BNC is a major US bank.
>
> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

(78) Target sentences:

A Linda is a bank teller.

B BNC engaged in risky trading in the mid 2000's.

C Linda is a bank teller, and BNC engaged in risky trading in the mid 2000's.

Notice that, while (78B) is not representative of the topical individual (Linda), it is reasonable to think that subjects will rank it high on a likelihood scale, assuming rather minimal familiarity with recent events in the American economy.

For each statement in each setup, I asked a question about likelihood and a question about representativeness.

(79) Questions:

a. To what degree does Linda represent a typical example of the following statement?

b. How likely do you think the following statement is?

In total there were 24 conditions: 4 setups with 3 sentences each, 2 target questions about each sentence. Each subject was assigned either to a representativeness task or a likelihood task, so that no individual had to switch gears between the two question-types exemplified in (79). Each subject saw all four setups in a random order and answered only one randomly selected question about each of the four setups. Thus, each subject provided four data points. I collected minimal demographic information about each subject, so as to control for possible effects of education.

Subjects rated the sentences they saw on a scale of 0 to 5, and they were given the option to answer "cannot answer / not applicable" if they wanted to. I gave subjects this option due to the unavoidable awkwardness resulting from asking subjects about representativeness on a target that wasn't about the topic individual. For example, the representativeness tasks for sentence (78B) looked like the following.

> In what follows, assume that BNC is a major US bank.
>
> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> **To what degree does Linda represent a typical example of the following statement?**
>
> *BNC engaged in risky trading in the mid 2000's.*

The rationale for offering a "cannot answer / not applicable" choice was that in principle it was possible that some subjects simply couldn't make sense of the question, while others could. I wanted to be able to discern between these two possible reactions to questions of this sort.

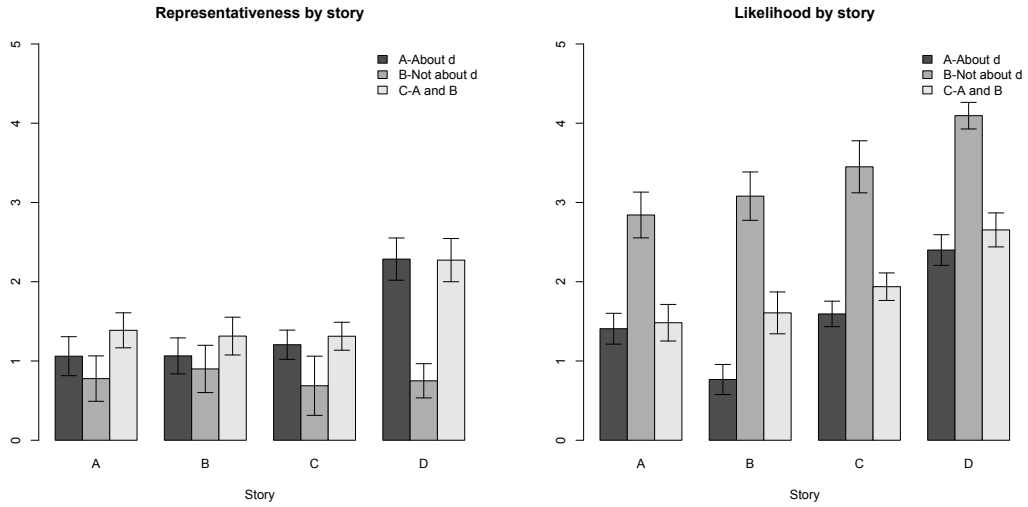


Figure 2: Representativeness and likelihood by conjunction-fallacy setup, with 95% confidence intervals. Stories A–D are as given in Appendix C.

The experiment was run on Amazon's Mechanical Turk (MTurk) crowdsourcing platform. Two-hundred subjects based in the US whose first language was English took the experiment. No special effort was made to recruit participants, nor was the experiment advertised beyond the standard ad that every MTurk task gets. Subjects were given seven minutes to complete the experiment and received a USD $0.20 compensation. They took on average just under three minutes to complete the experiment. No subject took more than five minutes.

4.3.3 Results and discussion

The data from one subject were unusable, as this individual did not answer any of the questions.[78] Of the remaining 199, all answered every question. Somewhat surprisingly, very few subjects responded to any questions with "cannot answer / not applicable," and indeed the distribution of this answer was rather uniformly scattered through all conditions. I chose to remove these responses from the analysis. Figure 2 shows the results by item (that is by setup). Some interesting tendencies are apparent from Figure 2.

[78]The implementation of the experimental design was such that subjects had to select each answer actively, including the "cannot answer / not applicable" option. We can thus be sure that the unresponsive subject either clicked through the experiment without selecting any options, or experienced some mysterious technical problem.

First, B sentences rank significantly higher on likelihood than on representativeness. This was as intended, for recall that B sentences are not about the topical individual, but were designed to be considered independently likely. Second, in the likelihood task, there appears to be a tendency for C sentences to rank higher than A sentences, a conjunction effect. This tendency seems overall less pronounced in the representativeness task.

A by-subjects analysis of the results proved difficult to make due to the specific design of this experiment. The items (the four setups) were quite different from each other and not enough data was collected from each subject. Consequently, it was difficult to determine the best way to collapse the items and get meaningful by-subject data. The most straightforward strategy is to take averages for those targets for which more than one item was seen, but this yielded only one significant result: representativeness ratings were very significantly lower than likelihood ratings for B sentences. Conjunction fallacy effects did not come out significant, nor did other differences between the targets.[79]

It is difficult to assess how seriously one should take this fact, given that this method of aggregating by-subject data involves very disparate comparisons between subjects and targets. For some pairs of subject-target one is looking at exactly one data point, for others one is looking at an average of two or three, for others there are no data points. The possible combinations were numerous enough, and the subject pool small enough, that further aggregating the by-subject data by say the number and type of data-points being averaged over scattered the data significantly.[80]

A less sophisticated but at least as reasonable by-subject analysis showed a new interesting result. Taking only the first item and question each subject saw (approximately eight data points per condition), I found that representativeness and likelihood assessments were not different for A targets and very significantly different for B targets, as before. But I also found that these ratings were significantly different for C targets ($p < .04$).[81] This is a promising result.

[79]The by-item analysis also only showed a significant result for B sentences. This too should be taken with a grain of salt, since the sample was so small (four items).

[80]In future research, I will address this difficulty by collecting more data from each subject and in a more systematic way, to make sure that I get comparable numbers of data points to average over for each subject. Along with more items to ask questions about, this will allow for a straightforward enough way to organize the data by subject.

[81]I found the same effect in a somewhat hybrid analysis, where I looked at each of the four data-points provided by each subject as independent responses. This gave me 796 data points. Not only were representativeness and likelihood

It is not surprising that representativeness and likelihood differ for B targets, since these were precisely the ones that are entirely unrelated to the topical individual. Moreover, the representativeness question about these targets was very awkwardly phrased, as discussed above, so that one would imagine the representativeness assessment could be brought down to some extent simply by virtue of that awkwardness. More interesting is the fact that C sentences also ranked higher on likelihood than on representativeness. Here, the representativeness question was far less awkward, if at all, so that the divergence between likelihood and representativeness rankings cannot be immediately explained away on a representativeness heuristic view.

This demonstrates the potential of the paradigm. The representativeness heuristic predicts that asking humans about likelihood will amount to asking them about representativeness, but this pilot study found that in conjunctions of statements with different topical individuals the representativeness heuristic as presented in the literature makes inadequate predictions, since representativeness and likelihood no longer go hand in hand.

significantly different for C targets, I also found a significant conjunction fallacy effect ($p < .05$) on the likelihood scale, but not on the representativeness scale. That is, C targets ranked significantly higher than A targets on likelihood, but not on representativeness. The non-standard nature of this method of aggregating data should give us pause, but not too much. All of the other methods of looking at this data by subject showed this very tendency. That is, C sentences were always ranked higher than A sentences on likelihood, while this tendency was appreciably less pronounced on the representativeness scale. Insofar as this analysis is indicative of a real phenomenon, a further challenge is posed for the representativeness heuristic. These are conjunction fallacy effects in radical indirect form where representativeness assessments predict no effect.

Appendix A

Getting classical reasoning on the erotetic theory

A.1 Preliminaries

The central results of this appendix are soundness and completeness of a special case of the erotetic theory of reasoning for classical propositional semantics. Specifically, I show that the particular class of erotetic reasoning strategies that use Inquire on every propositional atom mentioned in discourse is sound and complete for classical propositional semantics. To accomplish this, I must first give a precise definition of derivations within the (unqualified) erotetic theory of reasoning. In what follows, I abbreviate Erotetic Theory of Reasoning as ETR. Also in the interest of readability, I omit outer square brackets when talking about operations on mental model discourses, absent an input mental model discourse. That is, instead of $[\Delta]^{\mathrm{Up}}$, I write Δ^{Up}, to refer to Update with the argument Δ.

Definition 27 (Derivations in ETR). A derivation $\mathscr{D}$ is a non-empty sequence of operations on mental model discourses, such that $\langle\{0\},\emptyset,i\rangle\mathscr{D} = \langle\Gamma,B,i\rangle$, for i not a suppositional index. We call Γ the *conclusion* of $\mathscr{D}$, and the smallest set containing every Δ such that Δ^{Up} occurs somewhere in $\mathscr{D}$ the set of *hypotheses* of $\mathscr{D}$.

Armed with this notion of derivation, we can now define derivability in the usual way.

Definition 28 (Derivability in ETR). For Γ a mental model and Σ a set of mental models, we say that Γ is derivable from Σ, in symbols $\Sigma \vdash_{\text{ETR}} \Gamma$, iff there is a derivation $\mathscr{D}$ with all of its hypotheses in Σ and with conclusion Γ. When no confusion arises, we omit the subscript 'ETR' and write simply $\Sigma \vdash \Gamma$.

For example, it is true that $\{\{p \sqcup q, \neg p\}, \{p\}\} \vdash \{q\}$. Proof: let $\mathscr{D} = \langle \{p \sqcup q, \neg p\}^{\text{Up}}, \{p\}^{\text{Up}}, q^{\text{MR}} \rangle$. Notice also that $\{\{p \sqcup q, \neg p\}, \{q\}\} \vdash \{p\}$, for consider $\mathscr{D} = \langle \{p \sqcup q, \neg p\}^{\text{Up}}, \{q\}^{\text{Up}}, p^{\text{MR}} \rangle$. Definition 28 thus tracks derivability in the system in the most general way, not simply classical derivability. We need to define a more constrained notion of derivability which considers only that subset of the ETR derivations in Definition 27 that is sound for classical propositional semantics. I do this in the next section.

Before moving on to soundness, I define and discuss an indispensable translation from the language of mental models to the language of propositional formulas. Because the two languages are different, soundness and completeness for classical logic must be stated via a translation.

Definition 29 (Translating mental models into propositional formulas). Let a molecular structure $\mathfrak{M}$ be given. We first define @, a function from mental molecules to propositional formulas as follows, for $a \in \mathit{Atoms}(\mathfrak{M})$ and α, β molecules of arbitrary complexity.

$$0^{@} = \top$$

$$a^{@} = a$$

$$(\alpha \sqcup \beta)^{@} = \alpha^{@} \wedge \beta^{@}$$

The translation $*$ from mental models based on $\mathfrak{M}$ into propositional formulas is defined thus:

$$\emptyset^{*} = \bot$$

$$\{\alpha\}^{*} = \alpha^{@}$$

$$(\Gamma \cup \Delta)^{*} = \Gamma^{*} \vee \Delta^{*}$$

For convenience when stating the relevant theorems, we generalize $*$ to apply to sets of mental models. For $\Sigma = \bigcup_{i \leq n}\{\Gamma_i\}$, let $\Sigma^* = \bigcup_{i \leq n}\{\Gamma_i^*\}$.

The precise results shown here make crucial use of this translation. Specifically, I will show that, whenever there is *a certain well defined kind* of ETR derivation of some mental model Γ from set of premises Σ, then the translation Γ^* will classically follow from the translation of the premises Σ^* (soundness).

Conversely, I will also show that, whenever the translation Γ^* of some mental model Γ classically follows from Σ^*, then there is an ETR derivation (again, within a particular well defined class) of Γ from Σ. For the special case in which we have an empty set of premises (i.e. $\Sigma = \emptyset$), this means that a certain well-defined set of ETR theorems coincides with the set of classical tautologies. This result almost constitutes classical completeness, but one last step is required: every propositional formula φ has a classically equivalent formula $\varphi^\vee$ such that there is some mental model Γ with $\Gamma^* = \varphi^\vee$. This is true because every formula in the class of formulas in *disjunctive normal form* is the translation of some mental model.[82] Given the disjunctive normal form theorem for classical propositional logic, every formula is equivalent to a formula in disjunctive normal form. Therefore, while apparently less expressive, the language of mental models contains all that is needed to cover classical reasoning completely.

I now give detailed proof-sketches of soundness and completeness.

A.2 Soundness

As explained above, we need to characterize the class of ETR derivations that guarantee classical validity. The required definition of classical ETR derivations is the following.[83]

[82] I simplify matters in this discussion. Strictly speaking, every DNF formula *up to certain equivalences* (e.g. idempotence) is the translation of some mental model. For example, in a propositional language, the formulas p and $p \vee p$, while equivalent, are distinct objects in the language. In the mental model language defined in this dissertation however the objects $\{p\}$ and $\{p, p\}$ are properly identical and therefore indistinguishable. Thus, the formula $p \vee p$ is *not* the translation of any mental model. However, there is a formula equivalent to it, namely p, that is the translation of a mental model, namely $\{p\}$.

[83] It is important to remark that, while every derivation that falls under Definition 30 will be classically sound, not all classically sound derivations will be classical derivations according to Definition 30. This is because we wanted to give a definition of classical derivations that guaranteed soundness but that was simple enough, for the purposes

Definition 30 (Classical derivations). A non-empty sequence of operations on mental models $\mathscr{D}^C$ is a classical ETR derivation just in case (1) it is a derivation in the sense of Definition 27 and (2) every occurrence of Δ^{Up} and Δ^{Sup} in $\mathscr{D}^C$ is immediately preceded by a sequence of p^{Inq}, for each atom p that occurs somewhere in Δ. We will write $\Sigma \left|\!\!\frac{}{\text{CETR}}\right. \Gamma$ iff there is a classical derivation $\mathscr{D}^C$ with conclusion Γ and all its hypotheses in Σ.

What characterizes classical ETR derivations is the careful application of Inquire to each propositional atom occuring in a model Δ *right before* the update (or supposition-update) with Δ. It is useful to get an intuitive grasp on why this is the central ingredient for achieving classical reasoning with the erotetic system given here. The more logically experienced reader can probably skip the next few paragraphs and move to the statement of Lemma 2 on the next page.

Mental models in the erotetic theory of reasoning are in a sense typically underspecified. This was discussed in detail in Chapter 2, section 2.2.3.3: mental models on the erotetic theory of reasoning represent only the exact verifiers for complex propositions. Interestingly, this property of mental models does not *per se* give us the (welcome and intended) non-classical inference patterns of ETR. Rather, it is the way in which Q-update treats these mental models as *questions* and extremely strong *answers* to those questions that gives rise to non-classical reasoning patterns that are valid in ETR.[84] Consider as an example the fallacy of Affirming the Consequent (AC), which as discussed above is derivable in ETR. When the premise $\{q\}$ is interpreted in the context of $\{p \sqcup q, \neg p\}$, it is Q-update that is responsible for the output $\{p \sqcup q\}$: the molecule q has a non-empty intersection with the molecule $p \sqcup q$, and an empty intersection with molecule $\neg p$, so only the former molecule survives;

of expository clarity as well as mathematical elegance. Consequently, alternative definitions exist that still guarantee classical soundness and that more exactly characterize the class of such derivations. For present purposes, this is of little importance: I want to show that *there is* a subset of derivations guaranteed to be classically sound and that implement Part II of the erotetic principle.

[84] ETR captures certain other non-classical properties of naive reasoning in ways unrelated to Q-update, but these are no obstacle to soundness or completeness. For example, recall that in ETR the order in which premises are updated will in principle make a difference. This dynamic feature of the system was explored in the account of the order effect on modus tollens and disjunctive syllogism inferences: when the categorical premise is processed first, the inferences go through more easily. Crucially, this dynamic property only has an effect on the complexity of the derivation. It is not that in ETR modus tollens cannot be derived if the categorical premise is processed at the end. Rather, deriving a modus tollens inference with this order of premises is more complex, as it involves an application of Filter. This contrasts with modus tollens with the categorical premise processed first, where there is no need for Filter. Since the dynamic properties of the system only have an effect on the complexity of derivations, never on whether something is derivable or not, they will not present any difficulties in proving soundness and completeness.

following this update with molecular reduction on p gives us AC. Imagine now that we somehow blocked the effects of Q-update. Then the update with $\{q\}$ in the context of $\{p \sqcup q, \neg p\}$ would amount to the C-update, returning a simple conjunction of mental models (I disregard the effects of the background set B and the index i in this explanation), namely the model $\{p \sqcup q, \neg p \sqcup q\}$. From this resulting model there is no application of molecular reduction that could possibly return $\{p\}$, so we no longer validate AC. This example generalizes: if we somehow block the Q-portion of Update, reducing its effects to those of C-update, we get classical reasoning.

A possible way to get classical reasoning would thus be to define a special Update operation that used only C-update and never Q-update. However, this strategy would amount to introducing a novel rule of update exclusively dedicated to classical reasoning, which would undermine the strength of claims of generality and systematicity of the erotetic theory of reasoning: if it is possible to define a new rule of update that behaves classically, and if that new rule of update is actually strictly simpler than the non-classical one, it ought to be extremely easy to learn. We thus reject this possibility, and maintain that the rule of mental model update is exactly as defined earlier in this chapter, in terms of Q-update and C-update.

Happily, there is another strategy: we can block the effects of Q-update *without* changing the definition of Update, simply by making sure that every new update is interpreted against a mental model context that has been systematically extended with respect to the propositional atoms that occur in the new update. The relevant form of extension in preparation for a new update amounts to asking all atomic yes-no questions on atoms in the new update. Intuitively, this means that a reasoner is paying full attention to every single possibility compatible with the information given: this kind of attention comes at a big cost of (potential) exponential blowup of alternatives, but it has as its central benefit the guarantee of classical results. I illustrate this strategy by showing how it blocks AC.

Suppose a reasoner has just heard $\{q\}$ in the context of $\{p \sqcup q, \neg p\}$. She now chooses to Inquire on q in the same context, before the update with $\{q\}$. The expanded context is then $\{p \sqcup q, \neg p \sqcup q, \neg p \sqcup \neg q\}$ (for recall that by definition Inquire on q amounts to updating with $\{q, \neg q\}$ followed by Filter). What will be the effect of updating with $\{q\}$, the second premise? Recall that Q-update

will eliminate every element of the context that has an empty intersection with $\{q\}$. But in the new expanded context, each molecule contains either q or $\neg q$, and therefore the only molecules with empty intersections with $\{q\}$ will be those molecules in the context that contain $\neg q$. These molecules we would have wanted to eliminate anyway after the update with $\{q\}$, since they would include both $\neg q$ and q and thus be contradictions. The result of updating the expanded context with $\{q\}$ is thus $\{p \sqcup q, \neg p \sqcup q\}$. Clearly, there is no way of getting $\{p\}$ from this model, and thus we have blocked AC. As before, this example generalizes, and the effects of Q-update in a context that is expanded with respect to the relevant atoms are completely harmless for the purposes of classically sound reasoning. Thus:

Lemma 2. *Let a mental model discourse $\langle \Gamma, B, i \rangle$ and a mental model Δ be given, and let $\langle \Gamma', B', i \rangle$ be the result of inquiring in $\langle \Gamma, B, i \rangle$ on each propositional atom that occurs somewhere in Δ. Then $\langle \Gamma', B', i \rangle [\Delta]^{\mathrm{Up}} [\,]^{\mathrm{F}} = \langle \Gamma', B', i \rangle [\Delta]^{C} [\,]^{\mathrm{F}}$.*

Proof sketch. Let $\langle \Gamma', B', i \rangle [\Delta]^{Q} = \langle \Gamma'', B', i \rangle$. By definition of Q-update, $\Gamma'' = \Gamma' - E$, where $E = \{\gamma \in \Gamma' : (\bigsqcap \Delta) \sqcap \gamma = 0\}$. Notice that $\bigsqcap \Delta$ is either 0 or some other mental model molecule. If $\bigsqcap \Delta = 0$, then $E = \Gamma'$, $\Gamma'' = \emptyset$, and the lemma follows immediately from the definition of Update. For recall that when Q-update returns $\emptyset$ Update is by definition C-update.

Suppose then that $\bigsqcap \Delta \neq 0$, and call it δ to reduce clutter in our formulas. Each $\gamma \in \Gamma'$ will be in E just in case $\gamma \sqcap \delta = 0$. For any $\gamma \in \Gamma'$, it will have an empty intersection with δ just in case for all $p \sqsubseteq \delta$, $p \not\sqsubseteq \gamma$. Choose any such $p \sqsubseteq \delta$. Because $\gamma \in \Gamma'$ and Γ' is the result of inquiring on each atom that occurs somewhere in Δ, it must be that $\neg p \sqsubseteq \gamma$, or that $p = \neg q$ and $q \sqsubseteq \gamma$. Either way, $\gamma \sqcup \delta$ contains an atom and its negation and is therefore contradictory. Consequently, $\gamma \in E$ just in case $\gamma \in \Gamma'$ and $\{\gamma\} \times \Delta$ contains only contradictions. Clearly, these are γs whose counterparts would not be present in $\langle \Gamma', B', i \rangle [\Delta]^{C} [\,]^{\mathrm{F}}$ anyway, given the definition of C-update as $\times$ (together with manipulations of B) and the definition of Filter, so nothing valuable is lost in $\Gamma' - E$ and the lemma follows. □

With this intermediate result, I can now give a sketch of soundness via the $*$ translation.

Theorem 3 (Soundness via translation). $\Sigma \mid\!\!\!\overline{}_{\mathrm{CETR}} \Gamma \implies \Sigma^{*} \mid\!\!\!\overline{}_{\mathrm{CPL}} \Gamma^{*}$.

Proof sketch. I prove by induction on the length of classical ETR derivations. The base case is trivial, since $\{0\}$ is translated as the tautology.

The inductive step uses the hypothesis that a derivation of length n is classically sound, to show for each operation that, if that operation occurs at step $n+1$, classical validity will be preserved. Inquire and Filter are trivial: Inquire amounts to excluded middle, and Filter to $\varphi \vee \bot \leftrightarrow \varphi$ and double negation elimination for literals. The effects of Molecular Reduction amount to weakening of disjuncts, which is classically valid. For Δ^{Up}, we must use Lemma 2 and show that C-update preserves classical validity, as follows.

C-update is the mental model conjunction of the old Γ with a certain subset of the new Δ, namely the set containing only those $\delta \in \Delta$ that do not contradict any of the mental models in the background of established facts B. Notice that B contains only (though not necessarily all) models that could be classically gotten from Σ. This means that, if a $\delta \in \Delta$ contradicts some model in B, then there is some $\Gamma in \Sigma$ such that all of its counterparts in $\Gamma \times \{\delta\}$ would be contradictions. Therefore, it is safe to exclude from Δ all such δ, and the background set B is well-behaved with respect to classical soundness. Finally, I observe that mental model conjunction is analogous to conjunction introduction followed by some finite number of applications of the law of distributivity, to get to a disjunctive normal form.

Showing the validity of the suppositional operations Δ^{Sup} and Dep requires a few intermediate steps. First, one must show that any application of Δ^{Sup} will be followed (not necessarily immediately) by an application of Dep. This falls out of our definition of derivations, which makes sure there are no uncanceled suppositions. Now, Γ in a discourse with supposition (that is $\langle \Gamma, B, \langle B', i, S \rangle \rangle$) will not necessarily classically follow from the hypotheses in Σ. What we do have is that Γ in a suppositional discourse classically follows from the supposed mental model S, stored in the index slot of the suppositional discourse, together with the hypotheses in Σ. In symbols, $\Sigma^* \cup \{S^*\} \models_{\text{CPL}} \Gamma^*$. Given (the semantic correlate of) the deduction principle in classical logic, we also have $\Sigma^* \models_{\text{CPL}} S^* \rightarrow \Gamma^*$. But $S^* \rightarrow \Gamma^*$ is classically equivalent to $\Gamma^* \vee (\text{NEG}(S))^*$. This last formula is simply the $*$-translation of the result of the operation Dep, so Dep preserves classical validity. Given that Δ^{Sup} must always be followed by Dep later in the derivation, suppositional

discourses introduce no complications for soundness. □

A.3 Completeness

That ETR derivations are complete for classical semantics may not seem particularly surprising. The erotetic theory of reasoning captures inference patterns that are too strong for the standards of classical logic, so the surprising result would be if some classical inferences were lost in the process. Nevertheless, I outline in this section the proof of classical completeness for the classical erotetic theory of reasoning. The proof strategy used here will be familiar to readers acquainted with standard completeness proofs for classical propositional logic.

Since I use the traditional proof-technique of maximally consistent sets, I start by recasting consistency for the classical erotetic theory in the most natural way.

Definition 31 (CETR-consistency). A set of mental models Σ is CETR-consistent just in case $\Sigma \not\vdash_{\text{CETR}} \emptyset$. In what follows, we omit the prefix CETR and refer to this property simply as "consistency."

A crucial fact for completeness is that double negation elimination holds, both in the classical version of the erotetic theory and in the unqualified version.

Theorem 4. $\textsc{Neg}(\textsc{Neg}(\Gamma)) \vdash_{\text{CETR}} \Gamma$.

Interestingly, the proof of Theorem 4 is quite complex. Contrary to what one might have thought, it follows by no means *immediately* from the definition of Filter and its double negation elimination effects. This is because Filter eliminates double negations from literals, and therefore some work needs to be done to show that the operation NEG, which applies to mental models, introduces no surprises. The proof of Theorem 4 is rendered quite complex by the fact that it not true in general that $\textsc{Neg}(\textsc{Neg}(\Gamma)) = \Gamma$, which in turn is due to the fact that each application of NEG may involve more than one application of (the mental model correlate of) distribution of conjunction over disjunction. Consequently, $\textsc{Neg}(\textsc{Neg}(\Gamma))$ will often contain redundant material absent from the original Γ, and is therefore often distinct from it.

The first lemma required for completeness notes that, if a mental model Γ cannot be derived from some set of models Σ, then Σ can be put together with the negation of Γ, preserving consistency.

Lemma 5. $\Sigma \not\vdash_{\text{CETR}} \Gamma \implies \Sigma \cup \{\text{NEG}(\Gamma)\}$ *is consistent.*

Proof sketch. The equivalent claim $\Sigma \cup \{\text{NEG}(\Gamma)\}$ is inconsistent $\implies \Sigma \vdash_{\text{CETR}} \Gamma$ follows with a simple suppositional derivation in ETR, together with Theorem 4. □

The required model existence lemma is given below, via the translation $*$. It states that we can always find a classical model for the translations of CETR-consistent sets of mental models. Recall that classical models are valuation functions v from propositional atoms into truth values, 0 for false and 1 for true. Classical models are extended to full valuations $[\![\cdot]\!]_v$ that are defined for formulas of arbitrary complexity. Full valuations assign truth values to complex formulas as a function of the valuation v and the (standard) definitions of the propositional connectives. The existence of maximally consistent sets of mental models, as well as the useful properties of maximally consistent sets, is shown in the usual way. Because there are no ETR-specific elements in these proofs, I omit them.

Lemma 6 (Model existence). *If Σ is consistent, then there is a classical valuation v such that $[\![\Delta^*]\!]_v = 1$ for each $\Delta \in \Sigma$.*

Proof sketch. Find a maximally consistent extension Σ' of Σ, put $v(p) = 1$ iff $\{p\} \in \Sigma'$, for atomic (in the classical propositional sense) p, and extend v to a full valuation $[\![\cdot]\!]_v$. Next, show by induction on Δ that $[\![\Delta^*]\!]_v = 1$ iff $\Delta \in \Sigma'$. The atomic case follows by definition.

Suppose $\Delta = \{\alpha \sqcup \beta\}$. By the $*$-translation, $[\![\{\alpha \sqcup \beta\}]\!]_v = 1$ iff $[\![\alpha]\!]_v = [\![\beta]\!]_v = 1$, so the claim follows by the induction hypothesis together with maximal consistency of Σ'. The converse direction uses the fact that $\{\alpha \sqcup \beta\} \vdash_{\text{CETR}} \{\alpha\}, \{\beta\}$ (a simple molecular reduction). If $\{\alpha \sqcup \beta\} \in \Sigma'$ then by the properties of maximally consistent sets $\{\alpha\} \in \Sigma'$ and $\{\beta\} \in \Sigma'$. The result then follows from the induction hypothesis.

Suppose that $\Delta = \Delta' \cup \Delta''$. $[\![(\Delta' \cup \Delta'')^*]\!]_v = 1$ iff $[\![\Delta'^*]\!]_v = 1$ or $[\![\Delta''^*]\!]_v = 1$. By induction hypothesis $\Delta' \in \Sigma'$ or $\Delta'' \in \Sigma$. By maximal consistency of Σ' we get $\Delta' \cup \Delta'' \in \Sigma'$. Conversely, if $\Delta' \cup \Delta'' \in \Sigma'$

then by maximal consistency of Σ' we get $\Delta' \in \Sigma'$ or $\Delta'' \in \Sigma'$. The result then follows from the induction hypothesis.

Finally, since $\Sigma \subseteq \Sigma'$, we have $[\![\Delta^*]\!]_v = 1$ for each $\Delta \in \Sigma$. □

Theorem 7 (Completeness via translation). $\Sigma^* \mid\!\!=\!\!=_{\text{CPL}} \Gamma^* \implies \Sigma \mid\!\!\sim_{\text{CETR}} \Gamma$.

Proof sketch. We prove the contrapositive $\Sigma \not\mid\!\!\sim_{\text{CETR}} \Gamma \implies \Sigma^* \not\mid\!\!=\!\!=_{\text{CPL}} \Gamma^*$. Assume $\Sigma \not\mid\!\!\sim_{\text{CETR}} \Gamma$. By Lemma 5 $\Sigma \cup \{\text{NEG}(\Gamma)\}$ is consistent. By Lemma 6 there must be a classical valuation such that $[\![\Delta^*]\!] = 1$ for each $\Delta \in \Sigma$ and $[\![(\text{NEG}(\Gamma))^*]\!] = 1$, which entails that $[\![\Gamma^*]\!] = 0$. This is equivalent to $\Sigma^* \not\mid\!\!=\!\!=_{\text{CPL}} \Gamma^*$. □

Appendix B

Supplementary examples for the erotetic theory of reasoning

B.1 Negation

Example 24. (*Negation of inclusive disjunction is easier than negation of conjunction*). Khemlani et al. (2012) report that when participants had to state what was possible given denials of conjunctions and disjunctions, the negated conjunctions yielded 18% correct responses whereas the negated disjunctions yielded 89% correct responses. Likewise, when they had to formulate denials of conjunctions and disjunctions, they made correct denials for 0% of conjunctions but for 67% of inclusive disjunctions. Denying a conjunction require a large number of alternatives, generating the full range of alternatives requires extensive creative use of inquire. This is not the case for denying a disjunction.

It is not the case that A or B or C $\quad\quad$ $\textsc{Neg}(\{a,b,c\})$

$$\textsc{Neg}(\{a,b,c\}) = \{\neg a\} \times \{\neg b\} \times \{\neg c\} = \{\neg a \sqcup \neg b \sqcup \neg c\}$$

Notice that inquiring on any of the disjuncts does not yield any new alternatives. In other words, the default interpretation already yields the full range of alternatives (exactly one).

Example 25. (*Negation of the conditional*). If asked to falsify the conditional people manage to produce the *P and not Q* case (Oaksford & Stenning 1992). On the erotetic theory of reasoning, only a very small amount of creativity is predicted to be required.

What do you need to make "if p then q" false? $\textsc{Neg}(\{p \sqcup q, \neg p\})$

$$\begin{aligned}\textsc{Neg}(\{p \sqcup q, \neg p\}) &= \{\neg p, \neg q\} \times \{\neg\neg p\} \\ &= \{\neg p \sqcup \neg\neg p, \neg q \sqcup \neg\neg p\}\end{aligned}$$

Updating an empty mental model discourse with this result and filtering contradictions, we easily obtain *P and not Q*:

$$\langle\{\neg p \sqcup \neg\neg p, \neg q \sqcup \neg\neg p\}, \emptyset, i\rangle[\,]^{\mathrm{F}} = \langle\{p \sqcup \neg q\}, \{\{p\}, \{\neg q\}\}, i\rangle$$

B.2 Disjunction

Example 26. (*Disjunctive syllogism is harder than disjunctive modus ponens*). An interesting observation is that inferring the truth of one conjunct from the falsity of another, although manageable, is in fact harder than inferring the truth of the consequent of a conditional with a disjunctive antecedent from the truth of one of the disjuncts. This is a surprising fact, since the premises in the latter case are more complex (Rips 1994). On the erotetic theory, disjunctive syllogism requires an application of the contradiction filter, while disjunctive modus ponens is immediate.

Disjunctive syllogism

P_1	John will come to the party, or else Mary will.	$\{j, m\}$
P_2	John won't come to the party.	$\{\neg j\}$
C	Mary will come to the party.	$\{m\}$

$$\langle\{0\},\emptyset,i\rangle[\{j,m\}]^{\mathrm{Up}} = \langle\{\{j,m\},\emptyset,i\rangle$$
$$[\{\neg j\}]^{\mathrm{Up}} = \langle\{\{j \sqcup \neg j, m \sqcup \neg j\},\{\{\neg j\}\},i\rangle$$
$$[\,]^{\mathrm{F}} = \langle\{\{m \sqcup \neg j\},\{\{\neg j\},\{m\}\},i\rangle$$
$$[m]^{\mathrm{MR}} = \langle\{\{m\},\{\{\neg j\},\{m\}\},i\rangle$$

Disjunctive modus ponens

P_1	If there is an ace or a king then there is a queen.	$\{a \sqcup q, k \sqcup q, \neg a \sqcup \neg k\}$
P_2	There is an ace.	$\{a\}$
C	There is a queen.	$\{q\}$

$$\langle\{0\},\emptyset,i\rangle[\{a \sqcup q, k \sqcup q, \neg a \sqcup \neg k\}]^{\mathrm{Up}} = \langle\{a \sqcup q, k \sqcup q, \neg a \sqcup \neg k\},\emptyset,i\rangle$$
$$[\{a\}]^{\mathrm{Up}} = \langle\{a \sqcup q\},\{\{a\},\{q\}\},i\rangle$$
$$[q]^{\mathrm{MR}} = \langle\{q\},\{\{a\},\{q\}\},i\rangle$$

Example 27. (*Illusory inference from disjunction and categorical premise*). Walsh & Johnson-Laird (2004) report that 90% of participants make the fallacious illusory inference below, while 78% draw the correct conclusion in the control problem. Thus, making the illusory inference is in fact slightly easier than making the correct control inference. On the erotetic theory, this is captured by the fact that the control problem requires an additional application of filter.

Illusory inference

P_1	Either Jane is kneeling by the fire and she is looking at the TV or else Mark is standing at the window and he is peering into the garden.	$\{k \sqcup l, s \sqcup p\}$
P_2	Jane is kneeling by the fire.	$\{k\}$
C	Jane is looking at the TV.	$\{l\}$

$$\langle\{0\},\emptyset,i\rangle[\{k\sqcup l, s\sqcup p\}]^{\mathrm{Up}} = \langle\{k\sqcup l, s\sqcup p\},\emptyset,i\rangle$$
$$[\{k\}]^{\mathrm{Up}} = \langle\{k\sqcup l\},\{\{k\},\{l\}\},i\rangle$$
$$[l]^{\mathrm{MR}} = \langle\{l\},\{\{k\},\{l\}\},i\rangle$$

Control

P_1 Either Jane is kneeling by the fire and she is looking at the TV or else Mark is standing at the window and he is peering into the garden. $\{k\sqcup l, s\sqcup p\}$

P_2 Jane is not kneeling by the fire. $\{\neg k\}$

C Mark is standing by the window. $\{s\}$

$$\langle\{0\},\emptyset,i\rangle[\{k\sqcup l, s\sqcup p\}]^{\mathrm{Up}} = \langle\{k\sqcup l, s\sqcup p\},\emptyset,i\rangle$$
$$[\{\neg k\}]^{\mathrm{Up}} = \langle\{k\sqcup l\sqcup\neg k, s\sqcup p\sqcup\neg k\},\{\neg k\},i\rangle$$
$$[\,]^{\mathrm{F}} = \langle\{s\sqcup p\sqcup\neg k\},\{\{\neg k\},\{s\},\{p\}\},i\rangle$$
$$[s]^{\mathrm{MR}} = \langle\{s\},\{\{\neg k\},\{s\},\{p\}\},i\rangle$$

B.3 Conditionals

Example 28. (*Fallacies with conditionals and disjunction*). Conditionals together with disjunctions can yield fallacious inferences (Johnson-Laird 2008).

P_1 If Pat is here then Viv is here. $\{p\sqcup v, \neg p\}$

P_2 Mo is here or else Pat is here, but not both. $\{m\sqcup\neg p, p\sqcup\neg m\}$

C Pat and Viv are here, or else Mo is here. $\{p\sqcup v, m\}$

$$\langle\{0\},\emptyset,i\rangle[\{p\sqcup v,\neg p\}]^{\mathrm{Up}} = \langle\{p\sqcup v,\neg p\},\emptyset,i\rangle$$
$$[\{m\sqcup\neg p,p\sqcup\neg m\}]^{\mathrm{Up}} = \langle\{p\sqcup v\sqcup m\sqcup\neg p,p\sqcup v\sqcup\neg m,\neg p\sqcup m,\neg p\sqcup p\sqcup\neg m\},\emptyset,i\rangle$$
$$[\,]^{\mathrm{F}} = \langle\{p\sqcup v\sqcup\neg m,\neg p\sqcup m\},\emptyset,i\rangle$$
$$[p\sqcup v]^{\mathrm{MR}} = \langle\{p\sqcup v,\neg p\sqcup m\},\emptyset,i\rangle$$
$$[m]^{\mathrm{MR}} = \langle\{p\sqcup v,m\},\emptyset,i\rangle$$

Example 29. (*Illusory inference with conditional embedded in disjunction*). Illusory inferences can also be obtained with a conditional in a disjunction (Johnson-Laird & Savary 1999).

Illusion 0

P_1 If there is a king in the hand then there is an ace in the hand, or else if there is a queen in the hand then there is an ace in the hand. $\{k\sqcup a,\neg k,q\sqcup a,\neg q\}$

P_2 There is a king in the hand. $\{k\}$

C There is an ace in the hand. $\{a\}$

$$\langle\{0\},\emptyset,i\rangle[\{k\sqcup a,\neg k,q\sqcup a,\neg q\}]^{\mathrm{Up}} = \langle\{k\sqcup a,\neg k,q\sqcup a,\neg q\},\emptyset,i\rangle$$
$$[\{k\}]^{\mathrm{Up}} = \langle\{k\sqcup a\},\{\{k\},\{a\}\},i\rangle$$
$$[a]^{\mathrm{MR}} = \langle\{a\},\{\{k\},\{a\}\},i\rangle$$

Illusion 1

P_1 If there is a king in the hand then there is an ace in the hand, or else if there isn't a king in the hand then there is an ace in the hand. $\{k\sqcup a,\neg k,\neg k\sqcup a,\neg\neg k\}$

P_2 There is a king in the hand. $\{k\}$

C There is an ace in the hand $\{a\}$

$$\langle\{0\},\emptyset,i\rangle[\{k\sqcup a,\neg k,\neg k\sqcup a,\neg\neg k\}]^{\mathrm{Up}} = \langle\{k\sqcup a,\neg k,\neg k\sqcup a,\neg\neg k\},\emptyset,i\rangle$$
$$[\{k\}]^{\mathrm{Up}} = \langle\{k\sqcup a\},\{\{k\},\{a\}\},i\rangle$$
$$[a]^{\mathrm{MR}} = \langle\{a\},\{\{k\},\{a\}\},i\rangle$$

(Pseudo-)Illusion 2

The inference in this problem is not an illusion in the same sense in which the previous problems are, which are fallacies in classical logic. If we take *or* as inclusive, the conclusion follows. If we take *or* as exclusive, the premises are contradictory, in which case the conclusion classically follows as well.

P_1 If there is a king in the hand then there is an ace, or else there is an ace in the hand. $\{k\sqcup a,\neg k,a\}$

P_2 There is a king in the hand. $\{k\}$

C There is an ace in the hand. $\{a\}$

$$\langle\{0\},\emptyset,i\rangle[\{k\sqcup\neg a,a\sqcup\neg k\}]^{\mathrm{Up}} = \langle\{k\sqcup a,\neg k,a\},\emptyset,i\rangle$$
$$[\{k\}]^{\mathrm{Up}} = \langle\{k\sqcup a\},\{\{k\},\{a\}\},i\rangle$$
$$[a]^{\mathrm{MR}} = \langle\{a\},\{\{k\},\{a\}\},i\rangle$$

Example 30. (*Control problems are predicted to yield correct inferences*).

Control 1

P_1 There is a king in the hand and there is not an ace in the hand, or else there is an ace in the hand and there is not a king in the hand. $\{k\sqcup\neg a,a\sqcup\neg k\}$

P_2 There is a king in the hand. $\{k\}$

C There isn't an ace in the hand. $\{\neg a\}$

$$\langle\{0\},\emptyset,i\rangle[\{k\sqcup\neg a,a\sqcup\neg k\}]^{\mathrm{Up}}=\langle\{k\sqcup\neg a,a\sqcup\neg k\},\emptyset,i\rangle$$
$$[\{k\}]^{\mathrm{Up}}=\langle\{k\sqcup\neg a\},\{\{k\},\{\neg a\}\},i\rangle$$
$$[\neg a]^{\mathrm{MR}}=\langle\{\neg a\},\{\{k\},\{\neg a\}\},i\rangle$$

Control 2

P_1 If there is a king in the hand then there is an ace in the hand, or else there is not a king in the hand. $\{k\sqcup a,\neg k\}$

P_2 There is a king in the hand. $\{k\}$

C There is an ace in the hand. $\{a\}$

$$\langle\{0\},\emptyset,i\rangle[\{k\sqcup a,\neg k\}]^{\mathrm{Up}}=\langle\{k\sqcup a,\neg k\},\emptyset,i\rangle$$
$$[\{k\}]^{\mathrm{Up}}=\langle\{k\sqcup a\},\{\{k\},\{a\}\},i\rangle$$
$$[a]^{\mathrm{MR}}=\langle\{a\},\{\{k\},\{a\}\},i\rangle$$

Example 31. (*Illusory inference from disjunction, novel experiment*). As predicted by Q-update and our account of supposition, subjects accept as valid fallacious conditional conclusions in setups similar to Example 27. The following data were collected from 20 participants using an online survey. Numbers next to example titles below indicate proportions of "Yes" responses to the question of whether the conclusion follows from the premise.

Illusion 1 — 20/20

P_1 There is a jack and a king or there is a queen and an ace.

C If there is a jack, then there is a king.

$$\langle\{0\},\emptyset,i\rangle[\{j\sqcup k,q\sqcup a\}]^{\mathsf{Up}} = \langle\{j\sqcup k,q\sqcup a\},\emptyset,i\rangle$$
$$[\{j\}]^{\mathsf{Sup}} = \langle\{j\sqcup k\},\{\{j\},\{k\}\},\langle\emptyset,i,\{\{j\}\}\rangle\rangle$$
$$[k]^{\mathrm{MR}} = \langle\{k\},\{\{j\},\{k\}\},\langle\emptyset,i,\{\{j\}\}\rangle\rangle$$
$$[\,]^{\mathsf{Dep}} = \langle\{j\sqcup k,\neg j\},\emptyset,i\rangle$$

Illusion 2 — 19/20

P_1 There is a queen or there is a king and an ace.

C If there is a king then there is an ace.

$$\langle\{0\},\emptyset,i\rangle[\{q,k\sqcup a\}]^{\mathsf{Up}} = \langle\{q,k\sqcup a\},\emptyset,i\rangle$$
$$[\{k\}]^{\mathsf{Sup}} = \langle\{k\sqcup a\},\{\{k\},\{a\}\},\langle\emptyset,i,\{\{k\}\}\rangle\rangle$$
$$[a]^{\mathrm{MR}} = \langle\{a\},\{\{k\},\{a\}\},\langle\emptyset,i,\{\{k\}\}\rangle\rangle$$
$$[\,]^{\mathsf{Dep}} = \langle\{k\sqcup a,\neg k\},\emptyset,i\rangle$$

Control 1, invalid — 0/20

P_1 There is a jack or there is a queen or an ace.

C If there is a jack, then there is a queen.

Supposing "there is a jack" upon updating with the first premise will not lead to a conclusion of "there is a queen." This fallacy is not derivable using default procedures.

Control 2, invalid — 4/20

P_1 There is an ace or a king or there is a jack or a queen.

C If there is an ace, then there is a king or a jack or a queen.

Supposing "there is an ace" will not reduce the alternatives in the discourse to "there is a king or jack or a queen."

Control 3, valid — 15/20

P_1 There is a queen or there is a king or a jack.

C If there is no queen, then there is a king or a jack.

An instance of disjunctive syllogism, see Example 26.

Example 32. (*Affirming the consequent and denying the antecedent*). Ordinary reasoners are somewhat prone to the fallacies of affirming the consequent (AC) and denying the antecedent (DA). However, AC is endorsed more rapidly than DA (Barrouillet et al. 2000). AC can be derived by default updating and molecular reduction. By contrast, unless we consider very roundabout derivations, DA can only be derived if we suppose that reasoners take an extra interpretive processing step of generating an implicature from "if P then Q" to "P if and only if Q" (Grice 1989). This might explain why the reaction time involved in endorsing DA has been found to be longer than that involved in AC.

Affirming the consequent

P_1 If the card is long then the number is even. $\{l \sqcup e, \neg l\}$

P_2 The number is even. $\{e\}$

C The card is long. $\{l\}$

$$\langle \{0\}, \emptyset, i \rangle [\{l \sqcup e, \neg l\}]^{\mathrm{Up}} = \langle \{l \sqcup e, \neg l\}, \emptyset, i \rangle$$

$$[\{e\}]^{\mathrm{Up}} = \langle \{l \sqcup e\}, \{\{l\}, \{e\}\}, i \rangle$$

$$[l]^{\mathrm{MR}} = \langle \{l\}, \{\{l\}, \{e\}\}, i \rangle$$

Denying the antecedent

P_1 If the card is long then the number is even [with implicature to biconditional]. $\{l \sqcup e, \neg l \sqcup \neg e\}$

P_2 The card is not long. $\{\neg l\}$

C The card is not even. $\{\neg e\}$

$$\langle\{0\},\emptyset,i\rangle[\{l \sqcup e, \neg l \sqcup \neg e\}]^{\mathrm{Up}} = \langle\{l \sqcup e, \neg l \sqcup \neg e\},\emptyset,i\rangle$$
$$[\{\neg l\}]^{\mathrm{Up}} = \langle\{\neg l \sqcup \neg e\},\{\{\neg l\},\{\neg e\}\},i\rangle$$
$$[\neg e]^{\mathrm{MR}} = \langle\{\neg e\},\{\{\neg l\},\{\neg e\}\},i\rangle$$

Example 33. (*Order effect on modus tollens*). Modus tollens becomes easier if the negative premise is encountered before the conditional (Girotto et al. 1997).

P_1 If John comes to the party, Mary will come to the party. $\{j \sqcup m, \neg j\}$

P_2 Mary won't come to the party. $\{\neg m\}$

C John won't come to the party. $\{\neg j\}$

$$\langle\{0\},\emptyset,i\rangle[\{j \sqcup m, \neg j\}]^{\mathrm{Up}} = \langle\{j \sqcup m, \neg j\},\emptyset,i\rangle$$
$$[\{\neg m\}]^{\mathrm{Up}} = \langle\{j \sqcup m \sqcup \neg m, \neg j \sqcup \neg m\},\{\{\neg m\}\},i\rangle$$
$$[\,]^{\mathrm{F}} = \langle\{\neg j \sqcup \neg m\},\{\{\neg m\},\{\neg j\}\},i\rangle$$
$$[\neg j]^{\mathrm{MR}} = \langle\{\neg j\},\{\{\neg m\},\{\neg j\}\},i\rangle$$

P_1 Mary won't come to the party. $\{\neg m\}$

P_2 If John comes to the party, Mary will come to the party. $\{j \sqcup m, \neg j\}$

C John won't come to the party. $\{\neg j\}$

$$\langle\{0\},\emptyset,i\rangle[\{\neg m\}]^{\mathrm{Up}} = \langle\{\neg m\},\{\{\neg m\}\},i\rangle$$
$$[\{j \sqcup m, \neg j\}]^{\mathrm{Up}} = \langle\{\neg j \sqcup \neg m\},\{\{\neg m\},\{\neg j\}\},i\rangle$$
$$[\neg j]^{\mathrm{MR}} = \langle\{\neg j\},\{\{\neg m\},\{\neg j\}\},i\rangle$$

Example 34. (*Illusory consistency judgments with sets of biconditional statements*). Biconditional statements can yield mistaken judgments about consistency in certain cases (Johnson-Laird et al. 2004). The broad explanatory strategy on any mental model theory

would be to say that people are disposed to evaluate a set of statements as consistent if they can easily find a mental model in which all the statements hold. We will assume the following: (1) *P if and only if Q* is represented in mental models of the form $\{p \sqcup q, \neg p \sqcup \neg q\}$ (which is helpfully equivalent to the mental model resulting from $\|P \text{ if } Q \text{ and } Q \text{ if } P\|$ followed by the filter operation). (2) Reasoners are disposed to try to find a mental model in which a series of biconditional statement holds by interpreting the first biconditional, making a supposition, and continuing to update with the rest of the biconditionals. If the supposition yields the empty set, reasoners either conclude that the biconditionals are inconsistent or try again with a different supposition. Otherwise, they conclude that the biconditionals are consistent.

97% of responses correctly identified the following statements as consistent.

Consistent case where straightforward updating suffices

P_1	The chair is saleable if and only if it is elegant.	$\{s \sqcup e, \neg s \sqcup \neg e\}$
P_2	The chair is elegant if and only if it is stable.	$\{e \sqcup t, \neg e \sqcup \neg t\}$
P_3	The chair is saleable or it is stable, or both.	$\{s, t, s \sqcup t\}$

$$\langle\{0\},\emptyset,i\rangle[\{s \sqcup e, \neg s \sqcup \neg e\}]^{\text{Up}} = \langle\{s \sqcup e, \neg s \sqcup \neg e\},\emptyset,i\rangle$$
$$[\{s\}]^{\text{Sup}} = \langle\{s \sqcup e\},\{\{s\},\{e\}\},\langle\emptyset,i,\{s\}\rangle\rangle$$
$$[\{e \sqcup t, \neg e \sqcup \neg t\}]^{\text{Up}} = \langle\{s \sqcup e \sqcup t\},\{\{s\},\{e\},\{t\}\},\langle\emptyset,i,\{s\}\rangle\rangle$$
$$[\{s, t, s \sqcup t\}]^{\text{Up}} = \langle\{s \sqcup e \sqcup t\},\{\{s\},\{e\},\{t\}\},\langle\emptyset,i,\{s\}\rangle\rangle$$

Straightforward updating with a supposition does not yield an empty set, so the consistency judgment is easy to reach.

Example 35. (*Control problem for biconditional consistency judgments*). In contrast to the previous example, only 39% judged the set of statements below to be consistent. Here,

straightforward updating results in the empty model.

P_1	The chair is unsaleable if and only if it is inelegant.	$\{\neg s \sqcup \neg e, \neg\neg s \sqcup \neg\neg e\}$
P_2	The chair is inelegant if and only if it is unstable.	$\{\neg e \sqcup \neg t, \neg\neg e \sqcup \neg\neg t\}$
P_3	The chair is saleable or it is stable, or both.	$\{s, t, s \sqcup t\}$

$$\begin{aligned}
\langle\{0\},\emptyset,i\rangle[\{\neg s \sqcup \neg e, \neg\neg s \sqcup \neg\neg e\}]^{\mathrm{Up}} &= \langle\{\neg s \sqcup \neg e, \neg\neg s \sqcup \neg\neg e\},\emptyset,i\rangle \\
[\{\neg s\}]^{\mathrm{Sup}} &= \langle\{\neg s \sqcup \neg e\},\{\{\neg s\},\{\neg e\}\},\langle\emptyset,i,\{\neg s\}\rangle\rangle \\
[\{\{\neg e \sqcup \neg t, \neg\neg e \sqcup \neg\neg t\}\}]^{\mathrm{Up}} &= \langle\{\neg s \sqcup \neg e \sqcup \neg t\},\{\{\neg s\},\{\neg e\},\{\neg t\}\},\langle\emptyset,i,\{\neg s\}\rangle\rangle \\
[\{s, t, s \sqcup t\}]^{\mathrm{Up}} &= \langle\emptyset,\{\{\neg s\},\{\neg e\},\{\neg t\}\},\langle\emptyset,i,\{\neg s\}\rangle\rangle
\end{aligned}$$

Thus, the same reasoning strategy that yields a consistency judgment in the previous problem yields an inconsistency judgment in this case. What about the 39% who did obtain the consistency judgment on this problem? By choosing a "less natural" supposition and using contradiction filter, it is possible to find a mental model integrating the premises of this reasoning problem as well.

$$\begin{aligned}
\langle\{0\},\emptyset,i\rangle[\{\neg s \sqcup \neg e, \neg\neg s \sqcup \neg\neg e\}]^{\mathrm{Up}} &= \langle\{\neg s \sqcup \neg e, \neg\neg s \sqcup \neg\neg e\},\emptyset,i\rangle \\
[\{\neg\neg s\}]^{\mathrm{Sup}} &= \langle\{\neg\neg s \sqcup \neg\neg e\},\{\{\neg\neg s\},\{\neg\neg e\}\},\langle\emptyset,i,\{\neg\neg s\}\rangle\rangle \\
[\{\{\neg e \sqcup \neg t, \neg\neg e \sqcup \neg\neg t\}\}]^{\mathrm{Up}} &= \langle\{\neg\neg s \sqcup \neg\neg e \sqcup \neg\neg t\}, \\
&\qquad \{\{\neg\neg s\},\{\neg\neg e\},\{\neg\neg t\}\},\langle\emptyset,i,\{\neg\neg s\}\rangle\rangle \\
[\,]^{\mathrm{F}} &= \langle\{s \sqcup e \sqcup t\}, \\
&\qquad \{\{\neg\neg s\},\{s\},\{\neg\neg e\},\{e\},\{\neg\neg t\},\{t\}\},\langle\emptyset,i,\{\neg\neg s\}\rangle\rangle \\
[\{s, t, s \sqcup t\}]^{\mathrm{Up}} &= \langle\{s \sqcup e \sqcup t\}, \\
&\qquad \{\{\neg\neg s\},\{s\},\{\neg\neg e\},\{e\},\{\neg\neg t\},\{t\}\},\langle\emptyset,i,\{\neg\neg s\}\rangle\rangle
\end{aligned}$$

Example 36. (*Conditional transitivity*). We take it that when reasoners try to see if a conditional conclusion follows from premises, they are primed by the linguistic meaning of "if ... then" to use the supposition operation, since that operation is recruited by the linguistic meaning of "if ... then".

P_1	If P then Q.	$\{p \sqcup q, \neg p\}$
P_2	If Q then R.	$\{q \sqcup r, \neg q\}$
C	If P then R.	$\{p \sqcup r, \neg p\}$

$$
\begin{aligned}
[\{p\}]^{\text{Sup}} &= \langle \{p\}, \{\{p\}\}, \langle \{p\}, \emptyset, i \rangle \rangle \\
[\{p \sqcup q, \neg p\}]^{\text{Up}} &= \langle \{p \sqcup q\}, \{\{p\}, \{q\}\}, \langle \{p\}, \emptyset, i \rangle \rangle \\
[\{q \sqcup r, \neg q\}]^{\text{Up}} &= \langle \{p \sqcup q \sqcup r\}, \{\{p\}, \{q\}, \{r\}\}, \langle \{p\}, \emptyset, i \rangle \rangle \\
[\{p \sqcup r\}]^{\text{MR}} &= \langle \{p \sqcup r\}, \{\{p\}, \{q\}, \{r\}\}, \langle \{p\}, \emptyset, i \rangle \rangle \\
[\,]^{\text{Dep}} &= \langle \{p \sqcup r, \neg p\}, \emptyset, i \rangle
\end{aligned}
$$

Appendix C

Experimental materials

C.1 Story A, Linda — based on Tversky and Kahneman (1983)

In what follows, assume that BNC is a major US bank.

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

(80) Question:

a. To what degree does Linda represent a typical example of the following statement?

b. How likely do you think the following statement is?

(81) Targets:

a. Linda is a bank teller.

b. BNC engaged in risky trading in the mid 2000's.

c. Linda is a bank teller, and BNC engaged in risky trading in the mid 2000's.

C.2 Story B, Bill — based on Tversky and Kahneman (1983)

In what follows, assume that "The Superlatives" is an amateur jazz band.

Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.

(82) Question:

a. To what degree does Bill represent a typical example of the following statement?

b. How likely do you think the following statement is?

(83) Targets:

a. Bill is an amateur jazz player.

b. “The Superlatives” includes a keyboard player.

c. Bill is an amateur jazz player, and ”The Superlatives” includes a keyboard player.

C.3 Story C, Scandinavian individual — based on Tentori et al. (2004)

The Scandinavian peninsula, including Sweden, Norway, and Finland, has a high percentage of people with blond hair and blue eyes. This is the case even though (as in the US) every possible combination of hair and eye color occurs. Suppose we choose at random an individual from the Scandinavian peninsula.

(84) Question:

a. To what degree does the individual represent a typical example of the following statement?

b. How likely do you think the following statement is?

(85) Targets:

a. The individual has dark hair.

b. Norway has universal healthcare.

c. The individual has dark hair, and Norway has universal healthcare.

C.4 Story D, Donna — based on Politzer and Noveck (1991)

In high school, Donna got excellent grades at math and science. She likes human contact, she enjoys helping others, and she is very determined.

(86) Question:

a. To what degree does Donna represent a typical example of the following statement?

b. How likely do you think the following statement is?

(87) Targets:

a. Donna is a theoretical physicist.

b. Most scientists at UCLA's physics department are male.

c. Donna is a theoretical physicist, and most scientists at UCLA's physics department are male.

References

Anderson, Craig A. (1982). Inoculation and counterexplanation: Debiasing techniques in the perseverance of social theories. *Social Cognition*, 1(2):126–139.

Arkes, Hal R., David Faust, Thomas J. Guilmette and Kathleen Hart (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, 73(2):305–307.

Armstrong, David M. (2004). *Truth and Truthmakers*. Cambridge: Cambridge University Press.

Baron, Jonathan (1993). Deduction as an example of thinking. *Behavioral and Brain Sciences*, 16(2):336–337.

Barrouillet, Pierre, Caroline Gauffroy and Jean-François Lecas (2008). Mental models and the suppositional account of conditionals. *Psychological Review*, 115(3):760–771.

Barrouillet, Pierre, Nelly Grosset and Jean-François Lecas (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition*, 75(3):237–266.

Barrouillet, Pierre and Jean-François Lecas (1998). How can mental models theory account for content effects in conditional reasoning? A developmental perspective. *Cognition*, 67(3):209–253.

Bauer, Malcolm I. and Philip N. Johnson-Laird (1993). How diagrams can improve reasoning. *Psychological Science*, 4(6):372–378.

Braine, Martin D. and David P. O'Brien (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review*, 98(2):182–203.

Braine, Martin D. S., Brian J. Reiser and Barbara Rumain (1984). Some empirical justification for a theory of natural propositional logic. In Gordon H. Bower, editor, *The Psychology of Learning and Motivation*, chapter 18, pages 317–371. New York: Academic Press.

Braine, Martin D. S. and Barbara Rumain (1983). Logical reasoning. In J. H. Flavell and E. M. Markman, editors, *Handbook of Child Psychology: vol 3, Cognitive Development*, pages 263–339. New York: Wiley.

Chierchia, Gennaro (2004). Scalar implicatures, polarity phenomena, and the syntax/semantics interface. In A. Belleti, editor, *Structures and Beyond*. Oxford: Oxford University Press.

Chierchia, Gennaro, Danny Fox and Benjamin Spector (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Paul Portner, Claudia Maienborn and Klaus von Heusinger, editors, *Semantics: An International Handbook of Natural Language Meaning*. Berlin: Mouton de Gruyter.

Ciardelli, Ivano A (2009). *Inquisitive semantics and intermediate logics*. Master's thesis, University of Amsterdam.

Davidson, Donald (1967). Truth and meaning. *Synthese*, 17(3):304–323.

Dulany, Don E. and Denis J. Hilton (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, 9(1):85–110.

Evans, Jonathan St BT, Stephen E Newstead and Ruth MJ Byrne (1993). *Human reasoning: The psychology of deduction*. Psychology Press.

Fine, Kit (2012). A difficulty for the possible world analysis of counterfactuals. *Synthese*.

von Fintel, Kai (1999). The presupposition of subjunctive conditionals. In Uli Sauerland and Orin Percus, editors, *The Interpretive Tract. MIT Working Papers in Linguistics*, volume 25, pages 29–44. Cambridge, MA: MITWPL.

Fox, Danny (2007). Free choice disjunction and the theory of scalar implicature. In Uli Sauerland and Penka Stateva, editors, *Presupposition and Implicature in Compositional Semantics*, pages 71–120. Pelgrave McMillan.

Fox, Danny and Roni Katzir (2011). On the characterization of alternatives. *Natural Language Semantics*, 19:87–107.

Fox, John F (1987). Truthmaker. *Australasian Journal of Philosophy*, 65(2):188–207.

van Fraassen, Bas (1969). Facts and tautological entailments. *The Journal of Philosophy*, 66(15):477–487.

García-Madruga, Juan A, S Moreno, N Carriedo, F Gutiérrez and PN Johnson-Laird (2001). Are conjunctive inferences easier than disjunctive inferences? a comparison of rules and models. *The Quarterly Journal of Experimental Psychology: Section A*, 54(2):613–632.

Gigerenzer, Gerd (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe and M. Hewstone, editors, *European Review of Social Psychology*. Chichester, England: Wiley.

Gigerenzer, Gerd (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky (1996). *Psychological Review*, 103(3):592–596.

Gillies, Anthony S (2009). On truth-conditions for if (but not quite only if). *Philosophical Review*, 118(3):325–349.

Girotto, Vittorio, Alberto Mazzocco and Alessandra Tasso (1997). The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*, 63:1–28.

Grice, Paul (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Groenendijk, Jeroen (2008). Inquisitive Semantics: Two possibilities for disjunction. ILLC Prepublications PP-2008-26, ILLC.

Groenendijk, Jeroen and Floris Roelofsen (2009). Inquisitive semantics and pragmatics. In *Presented at the Workshop on Language, Communication, and Rational Agency, Stanford, May 2009*.

Groenendijk, Jeroen and Floris Roelofsen (2010). Radical inquisitive semantics. In *Preliminary version, presented at the Colloquium of the Institute for Cognitive Science, University of Osnabrueck*.

Groenendijk, Jeroen and Martin Stokhof (1990). Dynamic Montague grammar. In L. Kálmán, editor, *Proceedings of the Second Symposium on Logic and Language*.

Groenendijk, Jeroen and Martin Stokhof (1991). Dynamic predicate logic. *Linguistics and Philosophy*, 14:39–100.

Haiman, John (1978). Conditionals are topics. *Language*, 54(3):564–589.

Hamblin, Charles L. (1958). Questions. *Australasian Journal of Philosophy*, 36(3):159–168.

Harman, Gilbert (1979). If and modus ponens. *Theory and Decision*, 11(1):41–53.

Harman, Gilbert (1986). *Change in view: Principles of reasoning*. MIT press Cambridge, MA.

Heim, Irene (1982). *The semantics of definite and indefinite noun phrases*. Ph.D. thesis, University of Massachusetts Amherst.

Hertwig, Ralph and Gerd Gigerenzer (1999). The conjunction fallacy revisited: how intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12:275–305.

Hintikka, Jaakko (1962). *Knowledge and Belief: an Introduction to the Logic of the two Notions*. Cornell University Press Ithaca.

Hoch, Stephen J (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4):719–731.

Hodges, Wilfrid (1993). The logical content of theories of deduction. *Behavioral and Brain Sciences*, 16(2):353–354.

Holliday, Wesley H. and Thomas F. Icard (2013). Logic, probability, and epistemic modality. Unpublished manuscript, UC Berkley and Stanford.

Horn, Laurence (1972). *On the semantic properties of the logical operators in English*. Ph.D. thesis, UCLA.

Horn, Laurence (2000). From *if* to *iff*: conditional perfection as pragmatic strengthening. *Journal of Pragmatics*, 32:289–326.

Jackendoff, Ray (1983). *Semantics and Cognition*. MIT Press: Cambridge, MA.

Johnson-Laird, Philip N (1970). The perception and memory of sentences. *New Horizons in Linguistic. Harmondsworth, Middlesex England: Penguin Book*, pages 261–270.

Johnson-Laird, Philip N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press.

Johnson-Laird, Philip N (2008). Mental models and deductive reasoning. In Lance Rips and J Adler, editors, *Reasoning: studies in human inference and its foundations*, pages 206–222. Cambridge: Cambridge University Press.

Johnson-Laird, Philip N and Fabien Savary (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71(3):191–229.

Johnson-Laird, Philip N and Rosemary Stevenson (1970). Memory for syntax. *Nature*, 227(412).

Johnson-Laird, Philip Nicholas and Ruth MJ Byrne (1991). *Deduction*. Erlbaum Hillsdale, NJ.

Johnson-Laird, PN, Paolo Legrenzi and Vittorio Girotto (2004). How we detect logical inconsistencies. *Current Directions in Psychological Science*, 13(2):41–45.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, Daniel and Shane Frederick (2002). Representativeness revisited: attribute substitution in intuitive judgment. In Thomas Gilovich, Dale Griffin and Daniel Kahneman, editors, *Heuris-*

tics and Biases: The Psychology of Intuitive Judgment, pages 49–81. Cambridge: Cambridge University Press.

Kamp, Hans (1981). A theory of truth and semantic representation. In Jeroen Groenendijk, Theo Janssen and Martin Stokhof, editors, *Formal methods in the study of language*, pages 277–322. Amsterdam: Matematisch Centrum.

Karttunen, Lauri (1976). Discourse referents. In James McCawley, editor, *Syntax and Semantics: Notes from the Linguistic Underground*, volume 7, pages 363–386. New York: Academic Press.

Katzir, Roni (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, 30:669–690.

Khemlani, Sangeet, Isabel Orenes and Philip N. Johnson-Laird (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24(5):541–559.

Kleitman, Daniel (1969). On Dedekind's problem: the number of monotone boolean functions. In *Proceedings of the American Mathematical Society*, volume 21, pages 677–682.

Koralus, Philipp (2012). The open instruction theory of attitude reports and the pragmatics of answers. *Philosopher's Imprint*, 12(14).

Koralus, Philipp and Salvador Mascarenhas (2013). The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, 27:312–365.

Koriat, Asher, Sarah Lichtenstein and Baruch Fischhoff (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):107–118.

Kratzer, Angelika (1991). Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*. Berlin: Walter de Gruyter.

Kratzer, Angelika and Junko Shimoyama (2002). Indeterminate pronouns: the view from Japanese. In *Third Tokyo Conference on Psycholinguistics*.

Lassiter, Daniel (2011). *Measurement and Modality: The Scalar Basis of Modal Semantics*. Ph.D. thesis, New York University.

Lewis, David (1981). Ordering semantics and premise semantics for conditionals. *Journal of Philosophical Logic*, 10:217–234.

Mackie, John Leslie (1973). *Truth, probability and paradox: Studies in philosophical logic*. Oxford University Press.

Marr, David (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company.

Mascarenhas, Salvador (2009). *Inquisitive Semantics and Logic*. Master's thesis, ILLC.

Mascarenhas, Salvador (2011). Licensing by modification: the case of positive polarity pronouns. In Ana A. Guevara, Anna Chernilovskaya and Rick Nouwen, editors, *Proceedings of* Sinn und Bedeutung *16*, pages 417–429.

Morris, Bradley J and Uri Hasson (2010). Multiple sources of competence underlying the comprehension of inconsistencies: A developmental investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2):277–287.

Nilson, Hakan, Peter Juslin and Henrik Olsson (2008). Exemplars in the mist: the cognitive substrate of the representativeness heuristic. *Scandinavian Journal of Psychology*, 49:201–212.

Oaksford, Michael and Nicholas Chater (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, Michael and Keith Stenning (1992). Reasoning with conditionals containing negated constituents. *Journal of experimental psychology. Learning, memory, and cognition*, 18(4):835–854.

Oberauer, Klaus (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, 53(3):238–283.

Politzer, Guy and Ira A. Noveck (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research*, 20(2):83–103.

Rips, Lance (1994). *The Psychology of Proof*. Cambridge, MA: MIT Press.

Sauerland, Uli (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27:367–391.

Schlenker, Philippe (2012). The semantics/pragmatics interface. In Maria Aloni and Paul Dekker, editors, to appear in *Handbook of Semantics*. Cambridge: Cambridge University Press.

Schroyens, Walter J, Walter Schaeken and Géry d'Ydewalle (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking & reasoning*, 7(2):121–172.

Spector, Benjamin (2007). Scalar implicatures: exhaustivity and Gricean reasoning. In Maria Aloni, Paul Dekker and Alastair Butler, editors, *Questions in Dynamic Semantics*. Elsevier.

Stalnaker, Robert (1968). A theory of conditionals. *Studies in Logical Theory, Americal Philosophical Quarterly*, 2.

Stalnaker, Robert (1975). Indicative conditionals. *Philosophia*, 5:269–286.

Stalnaker, Robert (1981). A defense of conditional excluded middle. In William L. Harper, Robert Stalnaker and Glenn Pearce, editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*, pages 87–104. D. Reidel Publishing Company, Dordrecht.

Starr, William (forthcoming). What *if*? *Philosopher's Imprint*.

Stenning, Keith and Michiel van Lambalgen (2008a). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Stenning, Keith and Michiel van Lambalgen (2008b). Interpretation, representation, and deductive reasoning. In J Adler and Lance Rips, editors, *Reasoning: studies in human inference and its foundations*, pages 223–249. Cambridge: Cambridge University Press.

Tenenbaum, Josua B, Thomas L Griffiths and Charles Kemp (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

Tentori, Katya, Nicolao Bonini and Daniel Osherson (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28:467–477.

Tversky, Amos and Daniel Kahneman (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185:1124–1131.

Tversky, Amos and Daniel Kahneman (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315.

Walsh, Clare and Philip N. Johnson-Laird (2004). Coreference and reasoning. *Memory and Cognition*, 32:96–106.

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

Yalcin, Seth (2010). Probability operators. *Philosophy Compass*, 5(11):916–937.