The Erotetic Theory of Reasoning

Bridges Between Formal Semantics and the Psychology of Deductive Inference

Philipp Koralus* and Salvador Mascarenhas†,‡

September 30, 2013

1. Preliminaries

1.1. Introduction

The capacity to reason and draw inferences is central to all advanced human endeavors. Pushed to its limits, this capacity makes possible quantum physics and formal logic, as well as systematic theories of this very capacity. Equally remarkably, we are subject to systematic failures of reasoning in all domains, ranging from reasoning about what is and what might be the case to reasoning about probabilities, often in ways that are strikingly predictable (Tversky and Kahneman, 1983; Evans, 2002; Byrne, 2005; Johnson-Laird, 2008, among many others).

We introduce in this paper 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. The core of our 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 paper. The erotetic theory of reasoning based on this principle combines two classes of ideas from the philosophy of language and linguistics. The first one is the idea that we interpret premises as strong answers to questions, represented in the form of mental models that we bring to the task of interpretation (Koralus, 2012). The second class of ideas concerns enriched notions of

*Oxford University, pkoralus@alumni.princeton.edu, http://www.koralus.net

†New York University, smasc@nyu.edu, http://files.nyu.edu/sdm330/public/

‡We would like to thank Hannes Leitgeb, Kit Fine, Philippe Schlenker, Anna Szabolcsi, Chris Barker, Felipe Romero, Dylan Bumford, Alistair Isaac, Franz Berto, and Philip Johnson-Laird for helpful comments and discussions.

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 we shall return. The foregoing example is just one of many intricate data points on reasoning that need to be captured.

As noted, we are not only interested in systematically capturing divergences in naïve 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 paper, we prove that this reasoning strategy exists. In a certain specific sense we 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.

1.2. The scope of this paper

In this paper, we focus 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 we have 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 naïve reasoning diverges from what we are entitled to conclude and how it is possible for us to come to reason correctly.

For our purposes, solving the problem of success means that we have to show that there exists a reasoning strategy using our naïve reasoning capacities that is classically sound, and, under performance idealizations, classically complete. We want 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 we consider to be a particularly interesting set of data on propositional reasoning. The erotetic theory of reasoning we develop below captures a significant catalog of empirically documented patterns of naïve reasoning that diverge from the norms of what we are entitled to conclude, listed in Table 1 (on page 4). In addition to data from existing experimental literature, we present a novel illusory inference that, to our knowledge, no extant theory of reasoning captures.[1]

1.3. 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 paper.

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).

We also have important points of agreement with those defending rational analysis approaches to reasoning. A

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

Connective	Result	Reference	Examples in text	Supplementary examples
Not	Few list all alternatives compatible with negated conjunction	Khemlani et al. (2012)	exx. 15 & 16, p. 26	
Not	Most can list what corresponds to negated disjunction	Khemlani et al. (2012)		ex. 24, p. 1
Not	Easy to list a case that falsifies conditional	Oaksford and Stenning (1992)		ex. 25, p. 1
—	"Explosion" highly counterintuitive [(*P and not P*)→*Q*]	Harman (1986)	ex. 20, p. 31	
Or	Disjunctive syllogism is harder than disjunctive modus ponens	Rips (1994)		ex. 26, p. 1
Or	Disjunctive syllogism easier if categorical premise comes first	García-Madruga et al. (2001)	ex. 12 & 13, p. 24	
Or	Illusory inferences from disjunction and categorical premise	Walsh and Johnson-Laird (2004)	ex. 9, p. 23	
Or	Control problems with disjunction and categorical premise	Walsh and Johnson-Laird (2004)		ex. 27, p. 2
Or	Strategic supposition makes intractable problems easier	Johnson-Laird (2008)	exx. 19 & 10, p. 31	
Or	Disjunction introduction [*P*→(*P or Q*)] counterintuitive	Braine et al. (1984)	ex. 18, p. 28	
Or/If	Fallacies with conditionals and disjunction	Johnson-Laird (2008)		ex. 28, p. 2
Or/If	Illusory inferences with conditional embedded in disjunction	Johnson-Laird and Savary (1999)		ex. 29, p. 3
Or/If	Control problems with conditionals embedded in disjunction	Johnson-Laird and Savary (1999)		ex. 30, p. 3
If	Modus ponens is extremely easy	Braine and Rumain (1983)	ex. 21, p. 33	
If	Modus ponens easier than modus tollens	Evans et al. (1993)	ex. 21, p. 33	
If	Affirming consequent more rapid than denying antecedent	Barrouillet et al. (2000)		ex. 32, p. 5
If	Order effects on modus tollens	Girotto et al. (1997)		ex. 33, p. 5
If	Illusions of consistency with sets of biconditional statements	Johnson-Laird et al. (2004)		ex. 34, p. 5
If	Control problems for biconditional consistency judgments	Johnson-Laird et al. (2004)		ex. 35, p. 6
If	Illusory inference from disjunction to conditional	*This paper*	ex. 22, p. 35	

Table 1: Some core data on naïve reasoning captured by the erotetic theory (*Examples under "Supplementary examples" can be found at* `http://www.koralus.net/ETRFurtherExamples.pdf`)

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.[2] The aim we propose for our naïve 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.

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. 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 3 we introduce semi-formally the central components of the theory, to be fully formalized in section 4. For the present section, we concentrate on the principles whereby successive premises are combined, and

[2]The apparent tension between the approach in this paper 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 paper, 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.

explain intuitively what makes the account of reasoning in this paper 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. This is a more general and precise version of the proposal that we interpret statements relative to a question, represented in the form of a mental model, that a hearer or reasoner seeks to answer (Koralus, 2012). 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 naïve 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 3 and 4 we show precisely how the two parts of the erotetic principle are implemented in our proposed formal system.

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).[3]

(3) P_1: Either Jane is kneeling by the fire and she is looking at the window or else 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 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.

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.

[3] 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 (Mascarenhas, 2013):

(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.

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

In face of the fallacies that we are subject to in naïve 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 naïvely endorsed, but our reasoning abilities are not irretrievably lost to these factors. When pressed, even naïve 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 naïve reasoning. This may encourage some to think that it suffices to have a model of naïve 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 naïve reasoning is not in the first instance to produce classically correct reasoning. However, it is possible to use the resources of naïve 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.

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.

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) 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 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 naïve 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.[4] This much has been established in the reasoning literature. For example, the disjunctive syllogism in (5) becomes significantly more acceptable to naïve 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

[4] 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 meanings.

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 that represents every possibility with respect to some propositional atom. As discussed in section 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.

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 4 and 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. 19).

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. 21, since the former requires an optional operation that filters out contradictory alternatives.[5]

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. 9 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 naïve 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. 20, and ex. 18.

Naïve 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

[5]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.

grasped even through naïve reasoning. For example, modus ponens is predicted to be extremely easy (ex. 21) 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.

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*.[6] 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 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.

[6] 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 7).

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.[7] The crucial notion however is that of the verification relation and what it says about how sentences with connectives are exactly verified. In this paper, we abstract away from model-theoretic considerations 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 4.1, but limitations of space prevent us from giving an explicit model theory for exact verification semantics.[8]

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 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 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.

[7] 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.

[8] 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.

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 (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 paper 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 paper gain welcome independent motivation.

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 5.

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.[9] 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 $\mathcal{N}$ of mental model nuclei is a non-empty set containing the smallest representational units of the system for propositional reasoning, standing for atomic propositions:

$$\mathcal{N} = \{p, q, \ldots\} \qquad \dashv$$

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

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

$$\mathcal{N}^+ = \{p, \neg p, \neg\neg p, \neg\neg\neg p, \ldots, q, \neg q, \neg\neg q, \neg\neg\neg q, \ldots\} \qquad \dashv$$

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 '$\sqcup$'.

[9] 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.

Definition 3 (Mental model molecules). Every mental model nucleus in $\mathcal{N}^+$ 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 3.3, we say that the *alternatives* in a mental model are all of the mental 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) $(p \sqcup q) \sqcap (p \sqcup r) = p$ (ii) $(p \sqcup q \sqcup r) \sqcap (p \sqcup q \sqcup s) = p \sqcup q$

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) $p \sqsubseteq p \sqcup q$ (ii) $p \sqcup q \sqsubseteq p \sqcup q$

To sum up, we propose the following algebraic structure.

Definition 4 (Mental model molecular structures). A structure $\mathfrak{M} = \langle \mathcal{M}, \sqcup, \sqcap, 0 \rangle$ is a mental model molecular structure iff all of the following conditions hold: (1) $\mathcal{N}^+ \subseteq \mathcal{M}$. (2) $\langle \mathcal{M}, \sqcup, \sqcap \rangle$ is a lattice, that is, the operations $\sqcup$ and $\sqcap$ obey the laws of commutativity, associativity, and absorption. (3) For any α in $\mathcal{M}$, $\alpha \sqcup 0 = \alpha$, 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 $\langle \mathcal{M}, \sqsubseteq \rangle$, 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 $\mathcal{M}$ be mental model nuclei standing for atomic propositions; the set of those nuclei is merely included in $\mathcal{M}$. 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 $\mathfrak{M} = \langle \mathcal{M}, \sqcup, \sqcap, 0 \rangle$,

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

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 $\langle \mathcal{M}, \sqcup, \sqcap, 0 \rangle$, the set $\mathbb{M}$ of mental models is the smallest set containing every subset of $\mathcal{M}$ 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. ⊣

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 4 and reintroduce it in section 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 $\langle \Gamma, B \rangle$.⊣

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

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*.

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.[10]

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

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

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

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 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.[11]

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

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

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

Where $\Gamma' = \Gamma \times \{\delta \in \Delta : (\neg\exists\beta \in B)\ (\forall b \in \{\delta\} \times \beta)\ \textsc{Contr}(b)\}$. The function $B^*(\Gamma)$ is defined as follows (recall that $\mathfrak{M}$ is the molecular structure underlying $\mathfrak{D}$, as per Definition 4).[12]

[11] 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 $\bigsqcap$ with $\bigsqcup$ 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 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.

[12] 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.

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

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 $B^*(\Gamma)$ in the definitions that follow. ⊣

Example 6. 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:

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

Notice how the discourse is updated with $a \sqcup b$, while the background set is updated with two categorical mental models {a} and {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 7. Suppose we have a background $B = \{\{\neg a\}\}$ establishing that what is represented by a is not the case:

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

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

Example 8. $\langle\{a \sqcup b, c\},\emptyset\rangle[\{d\}]^C = \langle\{a \sqcup b, c\} \times \{d\},\{\{d\}\}\rangle = \langle\{a \sqcup b \sqcup d, c \sqcup d\},\{\{d\}\}\rangle$

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.

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

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

$$\begin{aligned} \|p\|^{\mathfrak{D}} &= \{p\}, \text{ for atoms } p \\ \|\varphi \textit{ or } \psi\|^{\mathfrak{D}} &= \|\varphi\|^{\mathfrak{D}} \cup \|\psi\|^{\mathfrak{D}} \\ \|\varphi \textit{ and } \psi\|^{\mathfrak{D}} &= \|\varphi\|^{\mathfrak{D}} \times \|\psi\|^{\mathfrak{D}} \\ \|\textit{not } \varphi\|^{\mathfrak{D}} &= \textsc{Neg}(\|\varphi\|^{\mathfrak{D}}) \end{aligned}$$ ⊣

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 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.[13,14]

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 that 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,

$$\textsc{Neg}(\Gamma) = \textsc{Neg}(\{\alpha_0, \ldots, \alpha_n\}) = \{\neg a_{00}, \ldots, \neg a_{0m_0}\} \times \cdots \times \{\neg a_{n0}, \ldots, \neg a_{nm_n}\} \,. \qquad \dashv$$

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

Example 9. (*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:

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

$$[\{k\}]^{\textsc{Up}} = \langle\{k \sqcup l\},\{\{k\}\}\rangle$$

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.

[13] 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.

[14] 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.

(9) 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. $\textsc{Neg}(\{a \sqcup b \sqcup c\}) = \{\neg a, \neg b, \neg c\}$

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,

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

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 10. We can now complete the last step of Example 9 with an application of MR, providing the conclusion mental model {l}.

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

Example 11. Beside 'pulling out a conjunct' from a categorical mental model, MR also allows us to 'reduce' alternatives.

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

Disjunctive syllogism, though valid, is in fact endorsed less often than the fallacy in Example 9 (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 12. (*Disjunctive syllogism with categorical premise before disjunctive premise*)

P_1	John won't come to the party.	{¬j}
P_2	John will come to the party, or else Mary will.	{j, m}
C	Mary will come to the party.	{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 13. (*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.	{¬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 9, 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:

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

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, then the erotetic theory needs a way to eliminate contradictory alternatives from mental models. Although naïve 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).

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

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

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

$$\textsc{dne}(\alpha) = \bigsqcup \{\textsc{dne}(a) : a \in \mathit{Atoms}(\mathfrak{M}) \ \&\ a \sqsubseteq \alpha\} \qquad \dashv$$

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 14. $\langle \{a \sqcup \neg\neg\neg a\}, \emptyset\rangle[\ ]^{\mathrm{F}}[\ ]^{\mathrm{F}} = \langle \{a \sqcup \neg a\}, \{\{a\}, \{\neg a\}\}\rangle[\ ]^{\mathrm{F}} = \langle \emptyset, \{\{a\}, \{\neg a\}\}\rangle$

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 15. (*Negation of conjunction*)

P It is not the case that A and B and C. NEG{a ⊔ b ⊔ c}

In cases like this, evaluating the negation of a conjunction yields as many alternatives as there are atomic propositions in the conjunction: NEG({a ⊔ b ⊔ c}) = {¬a, ¬b, ¬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*.[15] 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.

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 3.2, and is therefore costly.

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

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

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 16. (*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 15 as a starting point:

$$\langle\{\neg a, \neg b, \neg c\}, \emptyset\rangle[\{a\}]^{\text{Inq}} = \langle\{\neg a, \neg b \sqcup a, \neg b \sqcup \neg a, \neg c \sqcup a, \neg c \sqcup \neg a\}, \emptyset\rangle$$

$$[\{b\}]^{\text{Inq}} = \langle\{\neg a \sqcup b, \neg a \sqcup \neg b, \neg b \sqcup a, \neg c \sqcup a \sqcup b, \neg c \sqcup \neg a \sqcup b, \neg c \sqcup a \sqcup \neg b, \neg c \sqcup \neg a \sqcup \neg b\}, \emptyset\rangle$$

$$[\{c\}]^{\text{Inq}} = \langle\{\neg a \sqcup b \sqcup c, \neg a \sqcup b \sqcup \neg c, \neg a \sqcup \neg b \sqcup c, \neg a \sqcup \neg b \sqcup \neg c,$$

$$\neg b \sqcup a \sqcup c, \neg b \sqcup a \sqcup \neg c, \neg c \sqcup a \sqcup b\}, \emptyset\rangle$$

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 theory that makes questions about what is not already explicitly considered lead to better reasoning.[16] We prove the following in section 6.

[15] 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

[16] There is some empirical data supporting the notion that inquiring on what is not already explicitly represented can lead to improved reasoning.

(10) **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 17. Systematic inquiry blocks the disjunctive illusory inference discussed in Example 9. 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 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 18.

P_1	There is an ace.	$\{a\}$
C	There is an ace or a queen.	$\{a, q\}$

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.

$$\langle\{0\},\emptyset\rangle[\{a\}]^{\mathsf{Up}} = \langle\{a\},\{\{a\}\}\rangle$$

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

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

$$[a]^{\mathsf{MR}} = \langle\{q, a\},\{\{a\}\}\rangle$$

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.

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. As it turns out, the requisite operations for supposition supply a novel account of the interpretation from 'if p then q' into mental models.

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 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 our paper, has to be triggered for counterfactual suppositions.

Definition 18 (Suppose).

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

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.[17]

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

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

[17] 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": $\langle \Gamma, B', \langle B, i, \Sigma \rangle\rangle[\Delta]^{\text{SF}} = \langle \Gamma, B', \langle B, i, \Sigma \cup \{\Delta\}\rangle\rangle[\Delta]^{\text{Up}}$. The depose operation (Definition 19) would also have to adapted to handle multiple suppositions, thus: $\langle \Gamma, B', \langle B, i, \Sigma \rangle\rangle[\ ]^{\text{Dep}} = \langle \Gamma', B'', i \rangle = \langle \Gamma \cup \textsc{Neg}(S_0) \cup \cdots \cup \textsc{Neg}(S_n), B^*(\Gamma', \Gamma'), i \rangle$, for $\Sigma = \{S_0, \ldots, S_n\}$.

The idea here is that deposing the supposition characteristically yields two alternatives. In one 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 19. (*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.

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

With supposition, the erotetic theory of reasoning is now powerful enough to provide classical completeness (see section on metatheory). 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 naïve 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 20. 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, $\{p\}$ and $\{\neg p\}$) 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.

$$\langle \emptyset, \emptyset, \mathrm{i} \rangle [\{\neg \mathrm{q}\}]^{\mathrm{Sup}} = \langle \emptyset, \emptyset, \langle \emptyset, \mathrm{i}, \{\neg \mathrm{q}\} \rangle \rangle$$

$$[\ ]^{\mathrm{Dep}} = \langle \emptyset \cup \textsc{Neg}(\{\neg \mathrm{q}\}), \{\{\neg\neg \mathrm{q}\}\}, \mathrm{i} \rangle$$

$$= \langle \{\neg\neg \mathrm{q}\}, \{\{\neg\neg \mathrm{q}\}\}, \mathrm{i} \rangle$$

$$[\ ]^{\mathrm{F}} = \langle \{\mathrm{q}\}, \{\{\neg\neg \mathrm{q}, \mathrm{q}\}\}, \mathrm{i} \rangle$$

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. There are important differences between the accounts with empirical consequences, but a full discussion of these issues is beyond the scope of this paper (Koralus *in preparation*).

Definition 20 (Conditional as supposition).

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

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.[18]

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

[18] For those steeped in the mental model literature, in standard mental model theory, it is held that we are able to get from the default interpretation of the conditional to an intermediate interpretation including the alternative $\neg p \sqcup \neg q$ but not the alternative $\neg p \sqcup q$ (Barrouillet et al., 2000; Barrouillet and Lecas, 1998; but see O'Brien and Manfrinati, 2010). We can capture this in the present framework with a supplementary notion of inquire that can be defined without introducing any further anomalies. For any mental model $\{p\}$, $\langle \Gamma, B, i \rangle [\{p\}]^{\mathrm{AInq}} = \langle \Gamma, B, i \rangle [\{p\} \cup \{0\}]^{\mathrm{C}} [\ ]^{\mathrm{F}}$ The intuition here is that asymmetric inquire can be applied at will just like the notion of inquire we defined above. $\{0\}$ is the non-committal mental model that corresponds to *verum* in classical logic. *Verum* is always true, so we may always tautologously add *P or verum*, for any P. This device allows us to define the step-by-step expansion of mental model alternatives of the conditional envisaged by Johnson-Laid: $\|\textit{If } p \textit{ then } q\| = \{p \sqcup q, \neg p\}$ Step 1: Asymmetric Inquire on $\{\neg q\}$. $[\{\neg q\}]^{\mathrm{AInq}} = \{p \sqcup q, \neg p, \neg p \sqcup \neg q\}$ Step 2: Inquire on q. $[\{q\}]^{\mathrm{Inq}} = \{p \sqcup q, \neg p \sqcup \neg q, \neg p \sqcup q\}$. If one shares the intuition that asking 'what about P?' is not quite the same thing as asking 'what about not P?' then asymmetric inquire is independently motivated.

Example 21. (*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.

MP

P1	If the card is long then the number is even.	$\{l \sqcup e, \neg l\}$
P2	The card is long.	$\{l\}$
Conc.	The number is even.	$\{e\}$

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

$$[\{e\}]^{Up} = \langle\{l \sqcup e\},\{\{l\},\{e\}\},i\rangle$$

$$[e]^{MR} = \langle\{e\},\{\{l\},\{e\}\},i\rangle$$

MT

P1	If the card is long then the number is even.	$\{l \sqcup e, \neg l\}$
P2	The number is not even.	$\{\neg e\}$
Conc.	The card is not long.	$\{\neg l\}$

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

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

$$[\,]^{F} = \langle\{\neg l \sqcup \neg e\},\{\{\neg e\},\{\neg l\}\},i\rangle$$

$$[\neg l]^{MR} = \langle\{\neg l\},\{\{\neg e\},\{\neg l\}\},i\rangle$$

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). 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. *Reference suppressed* argues that a conservative extension of the theory presented in this paper yields the right predictions for both the classical Wason task and for variants that involve what seem like deontic rules (Griggs and Cox, 1982), but we have no room to consider this here.

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).

(11) 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:

(12) Bob danced. Therefore, if Bob danced then Hilary Clinton 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:

(13) 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 here, 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 seemingly

unique 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 22.

P1 John and Mary will come to the party, or Bill and Sue will. $\{j \sqcup m, b \sqcup s\}$

Conc. 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 ex:illusory-new.

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:

(14) If Jack danced, the music must have been good.

(15) John asked if Jack danced.

The question entertained in the second example 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 (15), where there is no *then*-clause, the relevant argument in the conditional-as-supposition analysis is the null nucleus {0}.

$$||\textit{if } \varphi||^{\mathfrak{D}} = \Gamma, \text{ such that } \mathfrak{D}[||\varphi||^{\mathfrak{D}}]^{\text{Sup}}[\{0\}]^{\text{Up}}[\ ]^{\text{Dep}}$$

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).

$$||\textit{if } \varphi||^{\mathfrak{D}} = \Gamma, \text{ such that } \mathfrak{D}[||\varphi||^{\mathfrak{D}}]^{\text{Sup}}[\ ]^{\text{Dep}} = \{||\varphi||^{\mathfrak{D}}, \textsc{Neg}(||\varphi||^{\mathfrak{D}})\}$$ ⊣

Thus:

$$||\text{if Jack danced}||^{\mathfrak{D}} = \{d, \neg d\}$$

Remarkably, it falls out of our account of the conditional that *if* without a *then*-clause contributes a question.

Moving on, 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.

6. Getting classical reasoning on the erotetic theory

6.1. Preliminaries

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

Definition 22 (Derivations in ETR). A derivation $\mathcal{D}$ is a non-empty sequence of operations on mental models, such that $\langle\{0\}, \emptyset, i\rangle\mathcal{D} = \langle\Gamma, B, i\rangle$, for i not a suppositional index. We call Γ the *conclusion* of $\mathcal{D}$, and the smallest set containing every Δ such that Δ^{Up} occurs somewhere in $\mathcal{D}$ the set of *hypotheses* of $\mathcal{D}$. ⊣

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

Definition 23 (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 $\Sigma \left|\overline{\overline{\text{ETR}}}\right. \Gamma$, iff there is a derivation $\mathcal{D}$ 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 $\mathcal{D} = \langle \{p \sqcup q, \neg p\}^{\text{Up}}, \{p\}^{\text{Up}}, q^{\text{MR}} \rangle$. Notice that it is also true that $\{\{p \sqcup q, \neg p\}, \{q\}\} \vdash \{p\}$, for consider $\mathcal{D} = \langle \{p \sqcup q, \neg p\}^{\text{Up}}, \{q\}^{\text{Up}}, p^{\text{MR}} \rangle$. Definition 23 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 22 that is sound for classical propositional semantics. We do this in the next section.

Before moving on to soundness, we 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 24 (Translation from mental models to propositional formulas). Let a molecular structure $\mathfrak{M}$ be given. We first define @, a function from mental molecules to propositional formulas as follows, for $a \in Atoms(\mathfrak{M})$ 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 $\Sigma^{*} = \bigcup_{i \leq n}\{\Gamma_i^{*}\}$. ⊣

Example 23. $\{p, q, r\}^{*} = p \vee q \vee r$, $\{p \sqcup \neg q, r \sqcup s\}^{*} = (p \wedge \neg q) \vee (r \wedge s)$, $\{\{p, q\}, \{r\}\}^{*} = \{p \vee q, r\}$

The precise results we show here make crucial use of this translation. Specifically, we 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, we 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.[19] 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 we need to cover classical reasoning completely.

We now give detailed proof-sketches of soundness and completeness.

6.2. 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.[20]

Definition 25 (Classical derivations in ETR). A non-empty sequence of operations on mental models $\mathcal{D}^C$ is a classical ETR derivation just in case (1) it is a derivation in the sense of Definition 22 and (2) every occurrence of Δ^{Up} and Δ^{Sup} in $\mathcal{D}^C$ is immediately preceded by a sequence of p^{Inq}, for each atom p that occurs somewhere in Δ. We will write $\Sigma \left|\overline{\frac{}{\text{CETR}}}\right. \Gamma$ iff there is a classical derivation $\mathcal{D}^C$ 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 1 on the next page.

Mental models in the erotetic theory of reasoning are in a sense typically underspecified. This was discussed in detail in Section 3.3: our mental models represent only the exact verifiers for complex propositions. Interestingly, this property of mental models doesn't *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*

[19] We simplify matters slightly in this discussion. Strictly speaking, every DNF formula *up to certain equivalences* (most notably idempotence and some cases of absorption) 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 we define in this paper 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 $\{p\}$.

[20] It is important to remark that, while every derivation that falls under Definition 25 will be classically sound, not all classically sound derivations will be classical derivations according to Definition 25. This is because we wanted to give a definition of classical derivations that guaranteed soundness but that was simple enough, for the purposes 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 our purposes, this is of little importance: we want to show that *there is* a subset of derivations guaranteed to be classically sound and that implement Part II of the erotetic principle.

to those questions that gives rise to non-classical reasoning patterns that are valid in ETR.[21] 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; 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 (we disregard the effects of the background set B and the index i for the purposes of 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, call it ClassUp, for classical update, 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 paper, 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. We 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 the effect of updating with $\{q\}$, the second premise, be? Recall that Q-update will eliminate every element of the context that has an empty intersection with $\{q\}$.

[21] ETR captures certain other non-classical properties of naïve 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 our 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. Concretely, 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.

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. Accordingly, this is the content of one of our most important metalogical results:

Lemma 1. *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 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 $E = \{\gamma \in \Gamma' : (\forall \delta \in \Delta)\ \gamma \sqcap \delta = 0\}$, that is E is the set of all molecules of Γ' that have empty intersections with all molecules of Δ.

Pick an arbitrary $\delta \in \Delta$. We consider first the case $\delta \neq 0$. Pick an arbitrary $\gamma \in \Gamma'$ and notice that γ and δ have an empty intersection just in case for all $p \sqsubseteq \delta$, $p \not\sqsubseteq \gamma$. Since $\delta \neq 0$, we can choose one such $p \sqsubseteq \delta$. But because $\gamma \in \Gamma'$ and Γ' is the result of inquiring on each atom that occurs somewhere in Δ, it must be that $\neg p \sqsubseteq \gamma$. Then $\gamma \sqcup \delta$ contains both p and $\neg p$, and is therefore a contradiction. Consequently (but keep in mind that we are for now only considering $\delta \neq 0$), $\gamma \in E$ 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.

In case $\delta = 0$, then necessarily $\gamma \sqcap \delta = 0$ for any γ. If $\Delta = \{\delta\}$, then $\Gamma'' = \emptyset$ and the claim follows immediately (for recall that, when Q-update fails, Update amounts to C-update). Otherwise, there will be other, non-0 molecules in Δ, in which case the fact that there is a $0 \in \Delta$ makes no difference: a molecule in Δ that excludes all molecules in Γ' effectively shifts the burden of selecting the γs to be excluded to other non-0 molecules in Δ. Reason as above. □

We also need to make sure that the background set B introduces no complications to soundness. Lemma 2 will suffice for our purposes.

Lemma 2. *For all classical derivations $\mathcal{D}^{C}$ with $\langle \{0\}, \emptyset, i\rangle \mathcal{D}^{C} = \langle \Gamma, B, i\rangle$, for all $\{b\} \in B$, b is an atom and $\Gamma \vdash_{\mathrm{CETR}} \{b\}$.*

Proof sketch. By induction on classical derivations. The condition is trivially satisfied for the base case. The only less-than-obvious inductive step is Update, which, according to Lemma 1, reduces to C-update plus Filter in the case of classical derivations. Now, $\langle \Gamma, B, i\rangle[\Delta]^{C} = \langle \Gamma', B', i\rangle$, where $\Gamma' = \Gamma \times \{\delta \in \Delta : (\neg \exists \beta \in B)\ (\forall b' \in \{\delta\} \times \beta)\ \textsc{Test}(b')\}$ and $B' = B^{*}(\Gamma', \Delta)$. Take an arbitrary $\{b\} \in B'$. According to the definition of $B^{*}(\alpha, \beta)$, there are three possible situations. 1. Γ' is categorical. Then b is an atom and $b \sqsubseteq \gamma$, for the $\gamma \in \Gamma'$. Obviously, b can be gotten from Γ' via MR on b. 2. Γ' is inquisitive and Δ is categorical. b is an atom and $b \sqsubseteq \delta$, for the $\delta \in \Delta$. Now, if Γ' is inconsistent the result follows by

(the procedure analogous to) *reductio ad absurdum*: b^{Sup}, Γ'^{Up}, Dep. Assume Γ' is consistent. Then $\Gamma' = \Gamma \times \{\delta\}$. This means we can use MR to get δ from Γ', and then MR again to get b from δ. 3. None of the above. Then $B' = B$ and the result follows from the induction hypothesis and the fact that $\Gamma \mid_{\overline{\text{CETR}}} \Delta \implies \Gamma \times \Gamma' \mid_{\overline{\text{CETR}}} \Delta$.

We reason in a similar way for the other operations that manipulate the background (MR, F, Inq, Sup). □

With these two intermediate results, we can now give a sketch of soundness via the $*$ translation. Because this proof involves an induction with many long (and rather tedious) steps, we present the proof sketch in a more informal fashion than other proof sketches, so as to improve readability.

Theorem 3 (Soundness via translation). $\Sigma \mid_{\overline{\text{CETR}}} \Gamma \implies \Sigma^* \models_{\text{CPL}} \Gamma^*$.

Proof sketch. We 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 atoms. The effects of Molecular Reduction amount to weakening of disjuncts, which is classically valid. For Δ^{Up}, we must use Lemma 1 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. By Lemma 2, B contains only (though not necessarily all) models that could be gotten from Γ. This means that, if a $\delta \in \Delta$ contradicts some model in B, then 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, we 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. These last steps are also clearly classically valid.

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 $\Sigma^* \models_{\text{CPL}} S^* \to \Gamma^*$. But $S^* \to \Gamma^*$ is classically equivalent to $\Gamma^* \vee (\textsc{Neg}(S))^*$. 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. □

6.3. 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, we outline in this section the proof of classical completeness for the classical erotetic theory of reasoning. The proof strategy we use here will be familiar to readers acquainted with standard completeness proofs for classical propositional logic.

It might be worth reminding the reader what the proof strategy is. We will show that, if Γ cannot be derived from some set of models Σ, then we can find a classical propositional model (a valuation on atoms) that makes the translation of Σ true but the translation of Γ false. The lemmas sketched below do this for us: Lemma 5 tells us that Σ is consistent with the negation of Γ, and Lemma 6 constructs a valuation v out of (a maximally consistent extension of) Σ with the negation of Γ. This valuation v makes all elements of Σ true and it makes the negation of Γ true. Thus v makes Γ false, and is therefore the desired classical propositional model.

Since we use the traditional proof-technique of maximally consistent sets, we start by recasting consistency for the classical erotetic theory in the most natural way.

Definition 26 (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 atoms (technically speaking, "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: one that is analogous to DeMorgan's laws, pushing a high negation into the level of atoms, and another to produce the mental model analog of a disjunctive normal form. As a consequence, $\textsc{Neg}(\textsc{Neg}(\Gamma))$ 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. $\Sigma \not\vdash_{\text{CETR}} \Gamma \implies \Sigma \cup \{\textsc{Neg}(\Gamma)\}$ *is consistent.*

Proof sketch. The equivalent claim $\Sigma \cup \{\textsc{Neg}(\Gamma)\}$ is inconsistent $\Longrightarrow \Sigma \vdash_{\text{CETR}} \Gamma$ 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, we 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). $\Sigma^* \models_{\text{CPL}} \Gamma^* \Longrightarrow \Sigma \vdash_{\text{CETR}} \Gamma$.

Proof sketch. We prove the contrapositive $\Sigma \not\vdash_{\text{CETR}} \Gamma \Longrightarrow \Sigma^* \not\models_{\text{CPL}} \Gamma^*$. Assume $\Sigma \not\vdash_{\text{CETR}} \Gamma$. By Lemma 5 $\Sigma \cup \{\textsc{Neg}(\Gamma)\}$ is consistent. By Lemma 6 there must be a classical valuation such that $[\![\Delta^*]\!] = 1$ for each $\Delta \in \Sigma$ and $[\![(\textsc{Neg}(\Gamma))^*]\!] = 1$, which entails that $[\![\Gamma^*]\!] = 0$. This is equivalent to $\Sigma^* \not\models_{\text{CPL}} \Gamma^*$. □

7. 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 naïve reasoning patterns described in the literature based on a dynamic update procedure that implements the principle in (15a). 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 solves 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 problem was to account for how our natural reasoning capacity can make it possible to learn how to reason correctly by classical standards – how it is possible to acquire the reasoning foundations for science and philosophy.

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 an exciting 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 independently motivated. Thus, reliance on a theory of reasoning such as ours could help assuage the classical semanticists' worry that since semantic analyses relying on operations over discourse representations use a grammatically unconstrained level of representation, they give up on explanatory power (Koralus, 2012).

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 paper, make us rational.

References

Anderson, C. A. (1982). Inoculation and counterexplanation: Debiasing techniques in the perseverance of social theories. *Social Cognition*, 1(2):126–139.

Arkes, H. R., Faust, D., Guilmette, T. J., and Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, 73(2):305–307.

Armstrong, D. M. (2004). *Truth and Truthmakers*. Cambridge: Cambridge University Press.

Baron, J. (1993). Deduction as an example of thinking. *Behavioral and Brain Sciences*, 16(02):336–337.

Barrouillet, P., Gauffroy, C., and Lecas, J. (2008). Mental models and the suppositional account of conditionals. *Psychological Review*, 115(3):760–771.

Barrouillet, P., Grosset, N., and Lecas, J. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition*, 75(3):237–266.

Barrouillet, P. and Lecas, J. (1998). How can mental models theory account for content effects in conditional reasoning? A developmental perspective. *Cognition*, 67(3):209–253.

Bauer, M. I. and Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, 4(6):372–378.

Braine, M. D. and O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review*, 98(2):182–203.

Braine, M. D. S., Reiser, B. J., and Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In Bower, G. H., editor, *The Psychology of Learning and Motivation*, chapter 18, pages 317–371. New York: Academic Press.

Braine, M. D. S. and Rumain, B. (1983). Logical reasoning. In Flavell, J. H. and Markman, E. M., editors, *Handbook of Child Psychology: vol 3, Cognitive Development*, pages 263–339. New York: Wiley.

Byrne, R. M. J. (2005). *The rational imagination: how people create alternatives to reality*. Cambridge, MA: MIT Press.

Ciardelli, I. A. (2009). Inquisitive semantics and intermediate logics. Master's thesis, University of Amsterdam.

Evans, J. S. B. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological bulletin*, 128(6):978.

Evans, J. S. B., Newstead, S. E., and Byrne, R. M. (1993). *Human reasoning: The psychology of deduction*. Psychology Press.

Fine, K. (2012). A difficulty for the possible world analysis of counterfactuals. *Synthese*.

Fox, J. F. (1987). Truthmaker. *Australasian Journal of Philosophy*, 65(2):188–207.

García-Madruga, J. A., Moreno, S., Carriedo, N., Gutiérrez, F., and Johnson-Laird, P. (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.

Gillies, A. S. (2009). On truth-conditions for if (but not quite only if). *Philosophical Review*, 118(3):325–349.

Girotto, V., Mazzocco, A., and Tasso, A. (1997). The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*.

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

Groenendijk, J. (2008). Inquisitive Semantics: Two possibilities for disjunction. ILLC Prepublications PP-2008-26, ILLC.

Groenendijk, J. and Roelofsen, F. (2009). Inquisitive semantics and pragmatics. In *Presented at the Workshop on Language, Communication, and Rational Agency, Stanford, May 2009*.

Groenendijk, J. and Roelofsen, F. (2010). Radical inquisitive semantics. In *Preliminary version, presented at the Colloquium of the Institute for Cognitive Science, University of Osnabrueck*.

Haiman, J. (1978). Conditionals are topics. *Language*, 54(3):564–589.

Hamblin, C. L. (1958). Questions. *Australasian Journal of Philosophy*, 36(3):159–168.

Harman, G. (1979). If and modus ponens. *Theory and Decision*, 11(1):41–53.

Harman, G. (1986). *Change in view: Principles of reasoning*. MIT press Cambridge, MA.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases*. PhD thesis, University of Massachusetts Amherst.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4):719–731.

Hodges, W. (1993). The logical content of theories of deduction. *Behavioral and Brain Sciences*, 16(02):353–354.

Johnson-Laird, P., Legrenzi, P., and Girotto, V. (2004). How we detect logical inconsistencies. *Current Directions in Psychological Science*, 13(2):41–45.

Johnson-Laird, P. N. (1970). The perception and memory of sentences. *New Horizons in Linguistic. Harmondsworth, Middlesex England: Penguin Book*, pages 261–270.

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

Johnson-Laird, P. N. (2008). Mental models and deductive reasoning. In Rips, L. and Adler, J., editors, *Reasoning: studies in human inference and its foundations*, pages 206–222. Cambridge: Cambridge University Press.

Johnson-Laird, P. N. and Byrne, R. M. (1991). *Deduction*. Erlbaum Hillsdale, NJ.

Johnson-Laird, P. N. and Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71(3):191–229.

Johnson-Laird, P. N. and Stevenson, R. (1970). Memory for syntax. *Nature*, 227(412).

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, J., Janssen, T., and Stokhof, M., editors, *Formal methods in the study of language*, pages 277–322. Amsterdam: Matematisch Centrum.

Karttunen, L. (1976). Discourse referents. In McCawley, J., editor, *Syntax and Semantics: Notes from the Linguistic Underground*, volume 7, pages 363–386. New York: Academic Press.

Khemlani, S., Orenes, I., and Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24(5):541–559.

Koralus, P. (2012). The open instruction theory of attitude reports and the pragmatics of answers. *Philosopher's Imprint*, 12(14).

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):107–118.

Kratzer, A. and Shimoyama, J. (2002). Indeterminate pronouns: the view from Japanese. In *Third Tokyo Conference on Psycholinguistics*.

Mackie, J. L. (1973). *Truth, probability and paradox: Studies in philosophical logic*. Oxford University Press.

Mascarenhas, S. (2009). Inquisitive semantics and logic. Msc thesis, University of Amsterdam.

Mascarenhas, S. (2013). An interpretation-based account of illusory inferences from disjunction. NYU, Ms under review.

Morris, B. J. and Hasson, U. (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.

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

Oaksford, M. and Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of experimental psychology. Learning, memory, and cognition*, 18(4):835–854.

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, 53(3):238–283.

O'Brien, D. and Manfrinati, A. (2010). The mental logic theory of conditional propositions. In Oaksford, M. and Chater, N., editors, *Cognition and Conditionals*, pages 39–54. Oxford: Oxford University Press.

Rips, L. (1994). *The Psychology of Proof*. Cambridge, MA: MIT Press.

Schroyens, W. J., Schaeken, W., and d'Ydewalle, G. (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.

Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, 5:269–286.

Starr, W. (forthcoming). What *if*? *Philosopher's Imprint*.

Stenning, K. and van Lambalgen, M. (2008a). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Stenning, K. and van Lambalgen, M. (2008b). Interpretation, representation, and deductive reasoning. In Adler, J. and Rips, L., editors, *Reasoning: studies in human inference and its foundations*, pages 223–249. Cambridge: Cambridge University Press.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315.

van Fraassen, B. C. (1969). Facts and tautological entailments. *The Journal of Philosophy*, 66(15):477–487.

von Fintel, K. (1999). The presupposition of subjunctive conditionals. In Sauerland, U. and Percus, O., editors, *The Interpretive Tract. MIT Working Papers in Linguistics*, volume 25, pages 29–44. Cambridge, MA: MITWPL.

Walsh, C. and Johnson-Laird, P. N. (2004). Coreference and reasoning. *Memory and Cognition*, 32:96–106.

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